Psychological Bulletin
2004, Vol. 130, No. 4, 631–663

0033-2909/04/$12.00 DOI: 10.1037/0033-2909.130.4.631

The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials

Drew Westen
Emory University

Catherine M. Novotny
Veterans Affairs Medical Center, San Francisco, California

Heather Thompson-Brenner
Boston University

This article provides a critical review of the assumptions and findings of studies used to establish psychotherapies as empirically supported. The attempt to identify empirically supported therapies (ESTs) imposes particular assumptions on the use of randomized controlled trial (RCT) methodology that appear to be valid for some disorders and treatments (notably exposure-based treatments of specific anxiety symptoms) but substantially violated for others. Meta-analytic studies support a more nuanced view of treatment efficacy than implied by a dichotomous judgment of supported versus unsupported. The authors recommend changes in reporting practices to maximize the clinical utility of RCTs, describe alternative methodologies that may be useful when the assumptions underlying EST methodology are violated, and suggest a shift from validating treatment packages to testing intervention strategies and theories of change that clinicians can integrate into *empirically informed therapies*.

> When the results of scientific studies are applied to new and important questions that may directly or indirectly affect clinical training, clinical treatment, and financial decisions about how to treat, it is useful for us to return to our roots in empirical science and to carefully consider again the nature of our scientific methods and what they do and do not provide in the way of possible conclusions relevant to those questions. (Borkovec & Castonguay, 1998, p. 136)

Robert Abelson (1995) has argued that the function of statistics is not to display "the facts" but to tell a coherent story—to make a principled argument. In recent years, a story has been told in the clinical psychology literature, in graduate programs in clinical psychology, in psychiatry residency programs, and even in the popular media that might be called "The Tale of the Empirically Supported Therapies (ESTs)." The story goes something like this.

> Once upon a time, in the Dark Ages, psychotherapists practiced however they liked, without any scientific data guiding them. Then a group of courageous warriors, whom we shall call the Knights of the Contingency Table, embarked upon a campaign of careful scientific testing of therapies under controlled conditions.
>
> Along the way, the Knights had to overcome many obstacles. Among the most formidable were the wealthy Drug Lords who dwelled in Mercky moats filled with Lilly pads. Equally treacherous were the fire-breathing clinician-dragons, who roared, without any basis in data, that their ways of practicing psychotherapy were better.
>
> After many years of tireless efforts, the Knights came upon a set of empirically supported therapies that made people better. They began to develop practice guidelines so that patients would receive the best possible treatments for their specific problems. And in the end, Science would prevail, and there would be calm (or at least less negative affect) in the land.

In this article we tell the story a slightly different way, with a few extra twists and turns to the plot. Ours is a sympathetic but critical retelling, which goes something like this.

> Once upon a time, psychotherapists practiced without adequate empirical guidance, assuming that the therapies of their own persuasion were the best. Many of their practices were probably helpful to many of their patients, but knowing which were helpful and which were inert or iatrogenic was a matter of opinion and anecdote.
>
> Then a group of clinical scientists developed a set of procedures that became the gold standard for assessing the validity of psychotherapies. Their goal was a valorous one that required tremendous courage in the face of the vast resources of the Drug Lords and the nonempirical bent of mind of many clinician-dragons, who tended to breathe some admixture of hot air, fire, and wisdom. In their quest, the Knights identified interventions for a number of disorders that showed substantial promise. The treatments upon which they bestowed Empirical Support helped many people feel better—some considerably so, and some completely.

Drew Westen, Department of Psychology and Department of Psychiatry and Behavioral Sciences, Emory University; Catherine M. Novotny, Veterans Affairs Medical Center, San Francisco, California; Heather Thompson-Brenner, Center for Anxiety and Related Disorders and Department of Psychology, Boston University.

Catherine M. Novotny is now at the Department of Mental Health Services, Opportunities for Technology Information Careers, Antioch, California.

Preparation of this article was supported in part by National Institute of Mental Health Grants MH62377 and MH62378 and by a Glass Foundation Grant to Drew Westen. We thank Hal Arkowitz, David Barlow, Rebekah Bradley, Glen Gabbard, Robert Rosenthal, Laura Westen, and Sherwood Waldron for their very useful comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Drew Westen, Department of Psychology and Department of Psychiatry and Behavioral Sciences, Emory University, 532 North Kilgo Circle, Atlanta, GA 30322. E-mail: dwesten@emory.edu

> In the excitement, however, some important details seemed to get overlooked. Many of the assumptions underlying the methods used to test psychotherapies were themselves empirically untested, disconfirmed, or appropriate only for a range of treatments and disorders. And although many patients improved, most did not recover, or they initially recovered but then relapsed or sought additional treatment within the next 2 years. Equally troubling, the Scientific Method (Excalibur) seemed to pledge its allegiance to whomsoever had the time and funding to wield it: Most of the time, psychotherapy outcome studies supported the preferred position of the gallant knight who happened to conduct them (Sir Grantsalot).
>
> Nevertheless, clinical lore and anecdotal alchemy provided no alternative to experimental rigor, and as word of the Knights' crusade became legendary, their tales set the agenda for clinical work, training, and research throughout the land. Many graduate programs began teaching new professionals only those treatments that had the imprimatur of Empirical Validation, clinicians seeking licensure had to memorize the tales told by the Knights and pledge allegiance to them on the national licensing examination, and insurance companies used the results of controlled clinical trials to curtail the treatment of patients who did not improve in 6 to 16 sessions, invoking the name of Empirical Validation.

This is a very different version of the story, but one that is, we hope to show, at least as faithful to the facts. The moral of the first, more familiar version of the story is clear: Only science can distinguish good interventions from bad ones. Our retelling adds a second, complementary moral: Unqualified statements and dichotomous judgments about validity or invalidity in complex arenas are unlikely to be scientifically or clinically useful, and as a field we should attend more closely to the conditions under which certain empirical methods are useful in testing certain interventions for certain disorders.

Let us be clear at the outset what we are not arguing. We are not advocating against evidence-based practice, looking to provide refuge for clinicians who want to practice as they have for years irrespective of empirical data. A major, and well-justified, impetus for the attempt to develop a list of ESTs was the literal Babble in the field of psychotherapy, with little way of distinguishing (or helping the public distinguish) useful therapeutic interventions from useless or destructive ones. And although we argue for a more nuanced story about efficacy and treatment of choice than sometimes seen in the scientific literature, we are not claiming our own narrative to be bias free (see Westen & Morrison, 2001). None of us is capable of being completely dispassionate about topics we are drawn to study, truth be told, by passionate beliefs. We have endeavored to present a balanced argument and have been aided in that endeavor by a number of colleagues, including several whose inclination might have been to tell a somewhat different tale. Fortunately, where our own critical faculties failed us, others' usually did not. What we *are* suggesting, however, is that the time has come for a thoroughgoing assessment of the empirical status of not only the data but also the methods used to assign the appellations empirically supported or unsupported.

We now tell our story in the more conventional language of science, beginning with an examination of the empirical basis of the assumptions that underlie the methods used to establish empirical support for psychotherapies. We then reexamine the data supporting the efficacy of a number of the treatments currently believed to be empirically supported.[1] We conclude by offering suggestions for reporting hypotheses, methods, and findings from controlled clinical trials and for broadening the methods used to test the clinical utility of psychosocial interventions for particular disorders. Throughout, we argue from data because ultimately the future of psychotherapy lies not in competing assertions about whose patients get more help but in replicable data. The question is how to collect and interpret those data so that, as a field, we maximize our chances of drawing accurate inferences.

Retelling the Story: The Assumptions Underlying ESTs

The idea of creating a list of empirically supported psychosocial treatments was a compelling one, spurred in part by concerns about other widely disseminated practice guidelines that gave priority to pharmacotherapy over psychotherapy in the absence of evidence supporting such priority (Barlow, 1996; Beutler, 1998, 2000; Nathan, 1998). The mandate to use, and train professionals exclusively in the use of, empirically validated therapies (now often called *empirically supported therapies*, or ESTs; Kendall, 1998) gained powerful momentum in 1995 with the publication of the first of several task force reports by the American Psychological Association (Task Force on Psychological Intervention Guidelines, 1995). This report, and others that followed and refined it, distinguished ESTs from the less structured, longer term treatments conducted by most practicing clinicians. Since that time, many advocates of ESTs have argued that clinicians should be trained primarily in these methods and that other forms of treatment are "less essential and outdated" (Calhoun, Moras, Pilkonis, & Rehm, 1998, p. 151; see also Chambless & Hollon, 1998; Persons & Silberschatz, 1998).

ESTs, and the research methods used to validate them, share a number of characteristics (see Chambless & Ollendick, 2000; Goldfried & Wolfe, 1998; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Nathan, Stuart, & Dolan, 2000). Treatments are typically designed for a single Axis I disorder, and patients are screened to maximize homogeneity of diagnosis and minimize co-occurring conditions that could increase variability of treatment response. Treatments are manualized and of brief and fixed duration to minimize within-group variability. Outcome assessment focuses primarily (though not necessarily exclusively) on the symptom that is the focus of the study. In many respects, these characteristics make obvious scientific sense, aimed at maximizing the internal validity of the study—the "cleanness" of the design. A valid experiment is one in which the experimenter randomly assigns patients, manipulates a small set of variables, controls potentially confounding variables, standardizes procedures as much as possible, and hence is able to draw relatively unambiguous conclusions about cause and effect.

What we believe has not been adequately appreciated, however, is the extent to which the use of RCT methodologies to validate ESTs requires a set of additional assumptions that are themselves neither well validated nor broadly applicable to most disorders and treatments: that psychopathology is highly malleable, that most patients can be treated for a single problem or disorder, that

[1] A number of researchers have pointed to important caveats in the enterprise of establishing a list of ESTs using randomized controlled trial (RCT) methodology, whose work we draw on here (Borkovec & Castonguay, 1998; Goldfried & Wolfe, 1996, 1998; Ingram & Ritter, 2000; Kazdin, 1997; Seligman, 1995).

psychiatric disorders can be treated independently of personality factors unlikely to change in brief treatments, and that experimental methods provide a gold standard for identifying useful psychotherapeutic packages. Psychotherapy researchers can put RCT methodology to different uses, some of which (hereafter referred to as *EST methodology* or *the methodology of ESTs*) entail all of these assumptions. Other uses of RCT methodology, such as those focused on testing basic theoretical postulates about change processes (Borkovec & Castonguay, 1998; Kazdin, 1997), mediators and moderators of outcome, or specific interventions (rather than entire treatments), entail only some of these assumptions some of the time. Here we focus on the assumptions of EST methodology rather than RCT methods more broadly and describe each of these assumptions and the extant data bearing on them. As the discussion below makes clear, we are not arguing that these assumptions are never valid. Rather, we are arguing that they are not *generally* valid—that is, that they apply to some instances but not others—and that researchers and consumers of research need to be more cognizant about the conditions under which their violation renders conclusions valid only with substantial qualification.

Psychological Processes Are Highly Malleable

The assumption of malleability is implicit in the treatment lengths used in virtually all ESTs, which typically range from about 6 to 16 sessions. As Goldfried (2000) has observed, historically the exclusive focus on brief treatments emerged less from any systematic data on the length of treatment required to treat most disorders effectively than from pragmatic considerations, such as the fact that if psychotherapy researchers were to compare their psychotherapies with medications, they needed to avoid the confound of time elapsed and hence tended to design treatments of roughly the length of a medication crossover design.[2] Equally important in determining the length of clinical trials was a simple, unavoidable fact of experimental method as applied to psychotherapy, the wide-ranging impact of which has not, we believe, drawn sufficient attention: The longer the therapy, the more variability within experimental conditions; the more variability, the less one can draw causal conclusions. As we argue, the preference for brief treatments is a natural consequence of efforts to standardize treatments to bring them under experimental control. Even 16 to 20 carefully controlled hour-long sessions pose substantial threats to internal validity. Indeed, we are aware of no other experiments in the history of psychology in which a manipulation intended to constitute a single experimental condition approached that length.

Given the centrality of the malleability assumption, one would expect that it rests on a strong evidentiary foundation; and for some disorders, brief, focal treatments do produce powerful results (see Barlow, 2002; Roth & Fonagy, 1996). However, a substantial body of data shows that, with or without treatment, relapse rates for all but a handful of disorders (primarily anxiety disorders with very specific therapeutic foci) are high. For example, data on the natural course of depression suggest that the risk of repeated episodes following an initial index episode exceeds 85% over 10 to 15 years (Mueller et al., 1999), and on average, individuals with major depressive disorder will experience four major depressive episodes of approximately 20 weeks duration each as well as a plethora of other depressive symptoms during periods of remission from major depressive episodes (Judd, 1997).

The malleability assumption is also inconsistent with data from naturalistic studies of psychotherapy, which consistently find a dose–response relationship, such that longer treatments, particularly those of 1 to 2 years and beyond, are more effective than briefer treatments (Howard, Kopta, Krause, & Orlinsky, 1986; Kopta, Howard, Lowry, & Beutler, 1994; Seligman, 1995). Of particular relevance is the finding from naturalistic samples that substantial symptom relief often occurs within 5 to 16 sessions, particularly for patients without substantial personality pathology; however, enduring "rehabilitation" requires substantially longer treatment, depending on the patient's degree and type of characterological impairment (Howard, Lueger, Maling, & Martinovich, 1993; Kopta et al., 1994). For example, Kopta et al. (1994) found that patients with characterological problems required an average of 2 years of treatment before 50% showed clinically significant change.

Although one might raise many legitimate methodological concerns about naturalistic designs, perhaps the most compelling data on the malleability assumption come from controlled trials themselves. As we discuss below, meta-analytic data on ESTs for a range of disorders using outcome intervals longer than 6 months suggest that most psychopathological vulnerabilities studied are in fact highly resistant to change, that many are rooted in personality and temperament, and that the modal patient treated with brief treatments for most disorders (other than those involving specific associations between a stimulus or representation and a highly specific cognitive, affective, or behavioral response) relapses or seeks additional treatment within 12 to 24 months.

Suggestive findings also come from research using implicit and other indirect measures (e.g., Gemar, Segal, Sagrati, & Kennedy, 2001; Hedlund & Rude, 1995), such as emotional Stroop tasks (in which participants are presented, for example, with neutral and depressive words in randomly varying colors and have to ignore the content of the word to report the color as quickly as possible; see Williams, Mathews, & MacLeod, 1996) or audio presentations of homophones associated with anxiety or depression (e.g., weak–week; see Wenzlaff & Eisenberg, 2001). Patients whose depression has remitted often show continued attentional biases toward depressive words, indexed by longer response latencies in Stroop tests and greater likelihood of choosing the depressive spelling of homophones. Research using implicit measures often finds continued biases toward depressive words and thematic content among people who are no longer depressed, suggesting that changes in state may or may not be accompanied by changes in diatheses for those states encoded in implicit networks and raising questions about the durability of change. A. T. Beck (1976) de-

[2] Other considerations have influenced the near-exclusive focus on brief treatments as well, such as considerations of cost, funding, and feasibility of research. Another is that most psychotherapy research is behavioral or cognitive–behavioral. Theorists from Skinner through Bandura have argued that human behavior is under substantial environmental control and that one system of responses can readily be changed without worrying about broader systems in which they may be embedded (see, e.g., Bandura, 1977; Skinner, 1953). Although this assumption was most strenuously advanced in the early days of behavior therapy, and many cognitive–behavioral therapy (CBT) researchers no longer explicitly endorse it, this assumption is now implicit in the design of virtually all clinical trials of psychotherapy.

scribed similar studies decades ago on the dream content of patients with remitted depression in his classic book on cognitive therapy for emotional disorders. These findings make sense in light of contemporary research in cognitive neuroscience (and social psychology) on implicit associational networks, which reflect longstanding regularities in the individual's experience, can be resistant to change, and likely provide a diathesis for many psychological disorders (Westen, 1998b, 1999, 2000). Although this is a frontier area of research, suggestive findings are beginning to emerge on prediction of future behavior, outcome, or relapse from indirect measures such as these (e.g., Segal, Gemar, & Williams, 1999; Wiers, van Woerden, Smulders, & De Jong, 2002).

Most Patients Have One Primary Problem or Can Be Treated as if They Do

The assumption that patients can be treated as if they have one primary, discrete problem, syndrome, or disorder—and the correlative assumption that if they have more than one disorder, the syndromes can be treated sequentially using different manuals (e.g., Wilson, 1998)—again reflects an admixture of methodological constraints and theoretical meta-assumptions. Perhaps most important are two features of the pragmatics of research. First, including patients with substantial comorbidities would vastly increase the sample size necessary to detect treatment differences if comorbidity bears any systematic relation to outcome. Thus, the most prudent path is arguably to begin with relatively "pure" cases, to avoid confounds presented by co-occurring disorders. Second, the requirement that research proposals be tied to categories defined by the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; 4th ed.; *DSM–IV*; American Psychiatric Association, 1994) to be considered for funding has virtually guaranteed a focus on single disorders (or at most dual diagnosis, such as posttraumatic stress disorder [PTSD] and substance abuse).

If we examine more carefully the empirical basis of this assumption, however, we find that, as a general rule, it fares no better than the malleability assumption. We focus here on three issues: the empirical and pragmatic limits imposed by reliance on *DSM–IV* diagnoses, the problem of comorbidity, and the way the different functions of assessing comorbidity in controlled trials and clinical practice may place limits on generalizability.

The Pragmatics of DSM–IV Diagnosis

Linking treatment research to *DSM*-defined categories has many benefits, the most important of which are the ability to generalize across different settings and the link between understanding psychopathology and identifying processes that might alter it. We note here, however, three costs.

First, *DSM* diagnoses are themselves created by committee consensus on the basis of the available evidence rather than by strictly empirical methods (such as factor analysis or latent class analysis), and in many cases they are under serious empirical challenge. For example, whether major depression is a distinct disorder or whether it simply represents the more severe end of a depressive continuum is unknown; nor is it known the extent to which the high comorbidity of major depressive disorder and generalized anxiety disorder (GAD) is an artifact of the way the two disorders are defined (overlapping criterion sets) or of a common diathesis for negative affect (see Brown, Chorpita, & Barlow, 1998; Westen, Heim, Morrison, Patterson, & Campbell, 2002). To the extent that some of the *DSM–IV* categories are themselves not empirically well supported, hitching our therapeutic wagons to these disorders may commit us to a range of empirically unsupported assumptions about psychopathology.

Second, the implicit assumption that patients typically present with symptoms of a specific Axis I diagnosis and can identify at the start of treatment precisely which one it is (with, perhaps, the aid of a telephone screen and a structured interview) is not generally valid.[3] For historical rather than rational or scientific reasons, treatment research has proceeded independently of any kind of systematic needs assessment of the reasons the average patient presents for psychotherapy in clinical practice. Instead, *DSM* (typically Axis I) categories have largely guided the psychotherapy research agenda in the past 20 years (Goldfried, 2000). Whether most patients seek treatment complaining primarily of Axis I disorders, either clinical or subclinical; whether most patients present primarily with interpersonal concerns (or with depression or anxiety in the context of interpersonal concerns, such as problematic relationship patterns, difficulties at work, etc.); or whether the average patient presents with a diffuse picture that requires more extensive case formulation than counting up diagnostic criteria is unknown (see Persons, 1991; Westen, 1998a). However, the best available data from both naturalistic and community (catchment) studies suggest that between one third and one half of patients who seek mental health treatment cannot be diagnosed using the *DSM* because their problems do not fit or cross thresholds for any existing category (see Howard et al., 1996; Messer, 2001). As Goldfried (2000) has observed, the requirement by funding agencies that researchers focus treatment research on *DSM*-defined psychiatric conditions has virtually eliminated research on problems that once dominated psychotherapy research, such as public speaking anxiety, interpersonal problems, or problems often associated with anxiety and depression both between and during episodes such as problematic self-esteem regulation.

A third problem in linking treatment research to Axis I categories is a pragmatic one. As several commentators have pointed out (e.g., Beutler, Moleiro, & Talebi, 2002; Weinberger, 2000), the sheer number of disorders in the *DSM–IV* renders the notion of clinicians learning disorder-specific manuals for more than a handful of disorders unrealistic. Given that 40% to 60% of patients do not respond to a first-line EST for most disorders (e.g., major depression or bulimia nervosa), clinicians would need to learn at least two or three manuals for each disorder. If researchers then start developing manuals for other disorders—including "atypical," "not otherwise specified," and subthreshold diagnoses—the number of manuals required for competent practice would be multiplied even further. This is a good example of a problem that is not inherent in the use of RCTs (e.g., for testing specific

[3] This assumption is, we suspect, rarely challenged in the treatment literature because of sampling techniques commonly used in psychotherapy research that render the problem opaque: Researchers typically establish specialty clinics for particular disorders and draw patients who self-identify as suffering primarily from those disorders. This is an area in which clinical observation may provide an important corrective to observation in the laboratory.

interventions, such as exposure, or theories of change) but that is inherent in the effort to identify *treatment packages* appropriate for a particular patient population and in the shift from manuals as tools for standardizing treatment in the laboratory to tools for standardizing treatment in clinical practice, a point to which we return.

The Problem of Comorbidity

Aside from the problem of linking treatment manuals to *DSM*-defined disorders is the question of whether, in fact, patients in clinical practice typically present with one primary disorder. The literature on comorbidity in both clinical and community samples suggests that single-disorder presentations are the exception rather than the rule. (We use the term *comorbidity* here only to imply co-occurrence, given the multiple potential meanings of the term; see Lilienfeld, Waldman, & Israel, 1994). Studies consistently find that most Axis I conditions are comorbid with other Axis I or Axis II disorders in the range of 50% to 90% (e.g., Kessler et al., 1996; Kessler, Stang, Wittchen, Stein, & Walters, 1999; Newman, Moffitt, Caspi, & Silva, 1998; Oldham et al., 1995; Shea, Widiger, & Klein, 1992; Zimmerman, McDermut, & Mattia, 2000).

The data on comorbidity are troublesome in light of the fact that the methodology underlying the identification of ESTs implicitly commits to a model of comorbidity that most psychotherapy (and psychopathology) researchers would explicitly disavow, namely that comorbidity is random or additive (i.e., that some people just happen to have multiple disorders, rather than that their symptoms might be interrelated). It may well be, as many advocates of ESTs have argued (e.g., Wilson, 1998), that the best way to approach a polysymptomatic picture is to use sequential manuals, one for depression, one for PTSD, one for GAD, and so forth. However, sequential symptom targeting may not be an optimal treatment strategy under conditions in which (a) seemingly distinct Axis I symptoms reflect common underlying causes, such as anxiety and depression that both stem from rejection sensitivity or a tendency to experience negative affect; (b) Axis I symptoms arise in the context of enduring personality patterns that create psychosocial vulnerabilities to future episodes; or (c) the presence of multiple symptoms can have emergent properties not reducible to the characteristics of each symptom independently.

As we argue below, the available data suggest that each of these conditions is frequently met. For example, depressed patients with a lifetime history of panic–agoraphobia spectrum symptoms not only show less response to interpersonal psychotherapy (IPT) in controlled clinical trials but also take substantially longer to respond to a sequential treatment strategy including selective serotonin reuptake inhibitors if they fail to respond to psychotherapy (Frank et al., 2000). This is not to say that such findings are universal; RCTs for some treatments and disorders have found just the opposite, that comorbidity has little impact on treatment outcome or that treatment of the target disorder leads to reduction of comorbid symptomatology (e.g., Borkovec, Abel, & Newman, 1995; Brown & Barlow, 1992). The point is simply that one cannot routinely assume that psychopathology is additive or can be treated as such.

The Function of Comorbidity Assessment and Generalizability to Everyday Clinical Practice

What is perhaps less obvious than the problem of comorbidity for treatments designed for single disorders is that the function of assessing for co-occurring conditions differs in research and practice in a way that can affect the generalizability of ESTs. Researchers typically begin by soliciting patients with a particular disorder, either through direct advertising or by informing clinicians in a treatment setting (usually a university clinic or medical center) about the kinds of patients suitable for the study. Respondents undergo a brief initial screen (often by telephone) to determine whether they are potentially appropriate for the treatment protocol, followed by a structured interview or set of interviews to make a final determination about their appropriateness for the study and to obtain pretreatment diagnostic data. Following this assessment, those admitted to the study arrive at the research therapist's office, and the treatment begins. Clinicians in studies assessing the efficacy of ESTs usually do not conduct their own evaluation and proceed on the assumption that the diagnosis is accurate and primary.

The point to note here is the function of assessing comorbid conditions in the laboratory, which is generally to eliminate patients who do not meet study criteria. The treating clinician may not even know whether the patient received a secondary diagnosis, which is typically immaterial to the treatment. Indeed, the clinician usually is kept blind to secondary diagnoses if one goal of the study is to assess their potential role as moderators of outcome.

In clinical practice, the situation is very different. Unless the patient has specifically sought out a specialist who works with a particular population, clinicians typically do not assume that one symptom or syndrome is primary. Rather than starting with one symptom or syndrome in mind, clinicians are likely to inquire broadly about the patient's symptoms, history, and so forth. Even for the unknown percentage of patients in clinical practice who identify a primary concern, the aim of inquiring about co-occurring conditions is not to decide whether to refer them elsewhere but to understand them better. This generally entails developing a tentative case formulation that cuts across symptoms and is likely to be substantially more varied than the standardized formulations about maladaptive schemas, interpersonal role transitions, and so forth that are essential in research to minimize within-group variation in interventions (see Persons & Tompkins, 1997; Westen, 1998a). We are not arguing here about the validity of clinicians' formulations, an issue addressed elsewhere (see Westen & Shedler, 1999a; Westen & Weinberger, 2003). Rather, we are simply noting the extent to which the requisites of experimental control in EST methodology limit the extent of variation permitted in case formulation, if variation in formulation is potentially related to variation in treatment delivered.

In clinical practice, symptoms initially identified as primary may not remain the focus of treatment over time, even if the clinician is appropriately responding to the patient's concerns. For example, many young people struggling with sexual orientation suffer from depression, anxiety, or suicidality (Harstein, 1996), and these psychiatric symptoms may be their primary complaint. In these cases, weeks or months of treatment may pass before the patient is able to recognize or acknowledge the source of distress. To what extent issues of this sort are responsible for some or most

symptomatology in everyday practice is unknown, but the methodology of ESTs commits to the assumption of their irrelevance, for two reasons. First, testing treatments brief enough to maintain experimental control and prescribing the number of sessions in advance to maximize comparability of treatments within and across conditions places a premium on rapid identification of treatment targets. Second, manualization presupposes that the same techniques (e.g., challenging dysfunctional cognitions, addressing problems in current relationships) should work for the same Axis I symptom or syndrome regardless of etiology, the circumstances that elicited it, the patient's personality, and so forth. This is one of many possible assumptions about the relationship between interventions and symptoms, but it is an untested one, and it should not be built into the structure of hypothesis testing for all forms of treatment for all disorders. It seems unlikely on the face of it, for example, that the same techniques useful for helping a depressed patient with situationally induced feelings of inadequacy (e.g., after a job loss) will always be optimal for treating someone with *chronic* feelings of inadequacy, let alone someone with the same symptom (depression) who is struggling with unacknowledged homosexuality, adult sequelae of childhood sexual abuse, aging in the context of a narcissistic personality style, or gene expression in the context of a family history of major depression.

Psychological Symptoms Can Be Understood and Treated in Isolation From Personality Dispositions

The assumption that psychological symptoms can be understood and treated in isolation from the personality of the person who bears them is essential to the methodology of ESTs, in large measure because of the brief, focal nature of treatment required to maximize experimental control and in part because of the focus on syndromes rather than processes or diatheses. Although treatments such as CBT and IPT target dysfunctional schemas or interpersonal patterns with roots in personality, neither treatment was intended to change enduring personality processes, and we know of no theory of personality or data suggesting that enduring personality processes or traits can typically be changed in 6 to 16 hour-long sessions. The only treatment considered an EST for personality disorders, Linehan's (1993) dialectical behavior therapy (DBT) for borderline personality disorder (BPD), takes roughly a year to complete what is essentially the first of several stages (M. M. Linehan, personal communication, May 2002). Research testing the efficacy of this first phase of DBT has found substantial behavioral change in parasuicidal behaviors (e.g., cutting) by 12 months along with a number of other clinically important outcomes (e.g., reduction in the number of days of hospitalization). However, personality variables such as feelings of emptiness showed little decline with even a year of treatment, and the enduring effects of DBT over years are unknown (Scheel, 2000).

The assumption that Axis I conditions can be treated as if they were independent of enduring personality dispositions has two complications, one empirical and one methodological, which we address in turn. The first is that, empirically, most Axis I syndromes are not independent of personality, and personality often moderates treatment response. The second is that, pragmatically, including patients who share a diagnosis such as depression but vary considerably in personality would require using sample sizes that are substantially larger than either customary or tenable for establishing ESTs.

Independence of Symptoms and Personality Processes

Accumulating evidence suggests that the first part of this assumption, that Axis I symptoms or syndromes can be understood apart from personality processes, is inaccurate for most disorders. Studies using factor analysis, latent class analysis, and structural equation modeling suggest that Axis I anxiety and mood disorders are systematically related to variables long considered personality variables, notably high negative and low positive affect (Brown et al., 1998; Krueger, 2002; Mineka, Watson, & Clark, 1998; Watson & Clark, 1992; Watson et al., 1994; Zinbarg & Barlow, 1996). Other research has found that different kinds of personality diatheses, such as vulnerability to loss versus vulnerability to failure, predispose different individuals to become depressed under different circumstances (e.g., Blatt & Zuroff, 1992; Hammen, Ellicott, Gitlin, & Jamison, 1989; Kwon & Whisman, 1998). The prevalence of comorbid Axis I conditions in patients treated for disorders such as depression, GAD, PTSD, and bulimia may actually provide an index of the prevalence of underlying personality diatheses. Studies using both adult (Newman et al., 1998) and adolescent (Lewinsohn, Rohde, Seeley, & Klein, 1997) samples suggest that the presence of multiple Axis I conditions is essentially a proxy for the presence of an Axis II condition, with the more Axis I symptoms present, the greater the likelihood of Axis II pathology.

Furthermore, a growing body of data suggests that the same Axis I symptom or syndrome may have different functions or implications in the presence of certain kinds of personality disturbance. Research on adolescents and adults with BPD has found differences on dozens of variables between patients diagnosed with major depressive disorder with and without BPD. A case in point is the way these patients experience, express, and attempt to regulate their distress. Borderline depression is not only quantitatively but qualitatively distinct from non-borderline depression, with markedly different correlates (Westen et al., 1992; Westen, Muderrisoglu, Fowler, Shedler, & Koren, 1997; Wixom, Ludolph, & Westen, 1993). For example, for people with both major depressive disorder and BPD, severity of depression is strongly correlated with a latent variable that includes abandonment fears, diffuse negative affectivity, an inability to maintain a soothing and constant image of significant others, and feelings of self-loathing and evilness. For people who have major depressive disorder without BPD, the same qualities are negatively correlated with severity of depression.

As noted above, data from many disorders and treatments (but not all; see, e.g., Hardy et al., 1995; Kyuken, Kurzer, DeRubeis, Beck, & Brown, 2001) suggest that patients treated for Axis I conditions often fare less well if they also have certain personality disorders, particularly BPD (e.g., Johnson, Tobin, & Dennis, 1991; Steiger & Stotland, 1996). Although this is typically described in terms of comorbidity as a moderator variable, the concept of comorbidity may be misleading because it implies that personality variables are an add-on to a symptom picture that is essentially distinct from them. This may be analogous to studying aspirin as a treatment for fever and viewing "comorbid" meningitis, influ-

enza, or appendicitis as moderating the relation between treatment (aspirin) and outcome (fever reduction). From a treatment perspective, the high correlations between trait anxiety and depression, and the substantial comorbidity between major depression and virtually every anxiety disorder, suggest that researchers might do well to develop treatments for negative affectivity and emotional dysregulation rather than focusing exclusively on *DSM*-defined syndromes.

The Paradox of Pure Samples

The prevalence of personality diatheses for psychopathology presents a methodological paradox. If researchers include patients with substantial personality pathology in clinical trials, they run the risk of ambiguous conclusions if these variables moderate outcome, unless sample sizes are sufficiently large to permit covariation or moderator analyses. If instead they exclude such patients (which, as we later note, is the norm, either explicitly or de facto through use of exclusion criteria such as suicidal ideation or substance abuse), one cannot assume generalizability to a target population that is rarely symptomatically pure.

The reader may object that starting with relatively pure cases is just the beginning of the enterprise: The appropriate way to develop and test a treatment is to begin with relatively circumscribed efficacy trials and then to move to community settings, where researchers can test experimental conditions that have already demonstrated efficacy in the laboratory. This sequential progression from pure to impure cases is probably an appropriate strategy for testing some therapies for some disorders (e.g., simple phobia or panic disorder, which may present as relatively discrete symptom constellations even within a polysymptomatic picture), but with two important caveats.

First, this approach commits de facto to many of the assumptions adumbrated here, most importantly the assumption that the polysymptomatic conditions seen in the community have no emergent properties that might call for different types of interventions. Interventions to address such emergent properties will, as a simple result of methodological preconditions, never be identified if investigators routinely start with less complex cases and focus studies in the community on interventions previously validated in RCTs. For example, a primary focus on eating symptoms may well be appropriate for some or many patients with bulimia nervosa; however, for others, such as those who are more impulsive, eating symptoms may need to be addressed within the context of broader problems with impulse and affect regulation, of which bingeing and purging may be one clinically salient example (Westen & Harnden-Fischer, 2001). The exclusion criteria frequently used in controlled clinical trials for bulimia nervosa, including substance abuse and suicidality (which exclude patients with substantial emotional dysregulation) and abnormally low weight (which excludes patients with anorexic symptoms) may be systematically constraining the phenomena seen in the laboratory and the interventions consequently chosen for examination (for empirical data, see Thompson-Brenner, Glass, & Westen, 2003).

The second caveat is that as researchers, educators, administrators, and clinicians, we need to exercise considerable circumspection in attempting to draw conclusions for training or public policy while we await data that could provide us with a fuller understanding of the conditions under which treatments developed in the laboratory are likely to be transportable to everyday clinical practice. It is one thing to say that cognitive therapy and IPT are the best treatments tested thus far in the laboratory for patients with major depression who pass rigorous screening procedures and that we do not know yet how these or other treatments will fare in naturalistic settings with more polysymptomatic patients. It is another to say that existing laboratory data already have demonstrated that we should stop teaching, and third-party payers should stop reimbursing, longer term, often more theoretically integrative treatments widely practiced for these disorders in the community. One can argue one or the other, but not both. As we suggest later, for some disorders and some treatments, existing laboratory data do appear to have strong implications for training and practice. For others, including several treatments widely viewed as ESTs, the empirical data support greater restraint in drawing conclusions until considerably more is known about the parameters within which these treatments are likely to operate effectively.

Controlled Clinical Trials Provide the Gold Standard for Assessing Therapeutic Efficacy

Perhaps the most central assumption underlying the enterprise of establishing ESTs is that RCT methodology provides the gold standard for assessing the efficacy of psychotherapeutic interventions. In this section we address a series of subassumptions or corollary assumptions central to assessing the validity of this assumption. These assumptions regard the functions of manualization, the pragmatics of dismantling, the independence of scientific conclusions from the processes used to select treatments to test, and the compatibility of the requisites of good science and good practice.

The Functions of Manualization

A key component of the assumption that experimental methods provide a gold standard for establishing ESTs is the corollary assumption that the elements of efficacious treatment can be spelled out in manualized form and that the interventions specified in the manual are the ones that are causally related to outcome. This corollary assumption is central to the rationale for the experimental study of psychotherapy because the aim of manualization is standardization of the intervention across participants and the control of potential confounding variables (see Wilson, 1998). Here we examine the logic of this assumption and the empirical data bearing on it.

The logic of manualization. There can be no question that some form of manualization, whether in the form of specific prescriptions or in the form of more general "practice guidelines" for therapists in RCTs, is essential in psychotherapy research, for multiple reasons. Manualization is essential to minimize variability within experimental conditions, to insure standardization across sites, and to allow consumers of research to know what is being tested. One cannot test experimental manipulations one cannot operationalize. We argue, however, that EST methodology imposes constraints on the ways manualization can be implemented that limit its flexibility and utility in generating scientifically and clinically useful data.

From the standpoint of experimental methodology, the best manual is one that can standardize the "dose," the timing of the

dose, and the specific ingredients delivered in each dose. This is the only way to minimize within-group variation and hence to be certain that all patients in a given treatment condition are really receiving the same treatment. The ideal manual from an experimental point of view would thus specify not only the number of sessions but precisely what is to happen in each session or at least within a narrow band of sessions. The more a manual deviates from this ideal, the less one can draw causal conclusions about precisely what caused experimental effects.

This simple methodological desideratum has broad implications, the most important of which is as follows: The extent to which a treatment requires a competent clinical decisionmaker who must decide how and where to intervene on the basis of principles (even principles carefully delineated in a manual) is the extent to which that treatment will not be able to come under experimental control in the laboratory. This places a premium on development of treatment packages that minimize clinical judgment because such treatments are the only ones that allow researchers to draw firm causal conclusions. If clinicians are then to use these treatments in everyday practice, the most empirically defensible way to do so is to adhere closely to the manual. This simple logical entailment of scientific method as applied to ESTs has led to a significant shift in training goals in many clinical psychology programs, away from training clinicians who can intervene with patients on the basis of their knowledge of relatively broad, empirically supported principles of change (e.g., efforts at response prevention must include attention to covert forms of avoidance that prevent extinction or habituation) toward training clinicians who can competently follow one manual for depression, another for BPD, another for social phobia, and so forth.

Historically, manuals did not arise as prescriptions for clinical practice. Manualization was simply a method for operationalizing what investigators were trying to study. The goal of manual development was to obviate the need for the kinds of secondary correlational analyses that are becoming increasingly common in psychotherapy research as researchers address the limits of experimental control in complex treatments (e.g., predicting outcome from therapist competence or adherence). Secondary analyses of this sort shift the nature of the question from a causal one (does this treatment produce better results than another treatment or a control condition?) to a correlational one (are these particular intervention strategies associated with positive outcome?). The more researchers must ask the second question, the less valuable manualization becomes (and indeed, the more problematic it becomes, because it artificially restricts the range of interventions tested to those predicted to be useful a priori and hence limits what might be learned about mechanisms of change).

The reader may object that manualization is a broad construct, and one that is currently undergoing considerable discussion and revision (see, e.g., Carroll & Nuro, 2002). However, as argued above, the logic of EST methodology requires a very particular form and use of manualization, one that many of its advocates may explicitly reject. As RCT methodology has metamorphosed into EST methodology, a shift has occurred from a view of experimental manipulations as *exemplars* of specific constructs to a view of experimental conditions as *constitutive* of those constructs. Put another way, a reversal of means and ends is taking place whereby manuals are not just convenient ways of operationalizing treatments in the laboratory but are the defining features of the treatments themselves. In the former approach to experimentation, as in most psychological research, the investigator sees the experimental intervention as drawn from a sample of possible interventions instantiating a particular construct. Just as a researcher studying the impact of positive affect on problem solving can operationalize induction of positive affect by having participants eat a candy bar, think about pleasant memories, or receive positive feedback, a researcher studying the impact of exposure on specific social phobia can operationalize exposure in dozens of ways. The goal in these cases is to generalize about the impact of positive affect or exposure on the dependent variables of interest, not about the impact of receiving a candy bar or performing a particular set of role-plays in a group. In the latter approach, in contrast, the researcher views the intervention not as an example of how one might proceed but as how one actually should proceed. Viewed this way, deviation from the package is just as problematic in everyday practice as in the laboratory because it renders the intervention different from the one that has been tested.

The difference between these two approaches to manualization is subtle, but the implications are enormous. Consider the case of IPT for depression. The IPT manual was originally devised simply as an attempt to operationalize, for research purposes, the kinds of interventions dynamically informed psychopharmacologists of the late 1960s used with their patients, particularly as a complement to acute medication treatment (see Frank & Spanier, 1995). Within a short span of years, however, researchers were exhorting clinicians to practice IPT but not the kinds of treatments it was attempting to approximate because the latter, unlike the former, had never been empirically validated.

Along with this shift in means and ends has come a shift from the study of treatment principles to the validation of treatment packages and a corresponding shift in the function of manuals from a descriptive one (allowing researchers to describe their experimental manipulations precisely) to a prescriptive one (standardization of clinical activity in everyday practice, so that clinicians carry out interventions in the precise ways they have been tested). In a prior era, clinicians who kept abreast of the empirical literature might have tried an exposure-based technique with a patient who manifested some form of avoidance, regardless of whether the patient carried a particular diagnosis. Today, an empirically minded clinician faces a dichotomous choice when confronted with a patient who meets certain diagnostic criteria: either to implement an empirically supported treatment package as a whole or to disregard psychological science. The clinician cannot, in good empirical faith, pick and choose elements of one treatment package or another because it is the package as a whole, not its specific components or mechanisms, that has been validated. Any divergence from the manual represents an unfounded belief in the validity of one's clinical judgment, which the clinician has learned is likely, on average, to produce worse outcomes.

What has not, we believe, been adequately appreciated is the extent to which a particular view of clinicians is an unintended but inexorable consequence of EST methodology. Any exercise of clinical judgment represents a threat to internal validity in controlled trials because it reduces standardization of the experimental manipulation and hence renders causal inferences ambiguous. A good clinician in an efficacy study (and, by extension, in clinical practice, if practitioners are to implement treatment manuals in the

ways that have received empirical support) is one who adheres closely to the manual, does not get sidetracked by material the patient introduces that diverges from the agenda set forth in the manual, and does not succumb to the seductive siren of clinical experience. The more researchers succeed in the scientifically essential task of reducing the clinician to a research assistant who can "run subjects" in a relatively uniform (standardized) way, the more they are likely to view psychotherapy as the job of paraprofessionals who cannot—and should not—exercise clinical judgment in selecting interventions or interpreting the data of clinical observation.

The logic of experimental method in ESTs actually dictates not only the kind of therapist interventions that can be tested or permitted (those that can be rigorously manualized) but also the kind of patient activity. The scientific utility of treatment manuals is maximized in treatments in which the therapist sets the agenda for each session. Where patients have a substantial degree of control over the content or structure of treatment hours, therapists by definition have less control. Where therapists have less control, standardization is diminished and within-group variance attributable to sources other than standardized technique is correspondingly increased. The paradox of manualization for disorders such as depression and GAD is that the patient's active involvement in the treatment is likely to be essential to good outcome but destructive of experimental control. Modeled after dosing in medication trials (an analogy explicit in dose–response curves in psychotherapy research; see Stiles & Shapiro, 1989), manualization commits researchers to an assumption that is only appropriate for a limited range of treatments, namely that therapy is something done to a patient—a process in which the therapist applies interventions—rather than a transactional process in which patient and therapist collaborate. As we note below, within the range of cognitive–behavioral treatments, those that require genuine collaboration and creative problem solving on the part of the patient, such as A. T. Beck's (1976) cognitive therapy for depression (which explicitly aims at a "collaborative empiricism" between therapist and patient), have proven most recalcitrant to experimental control and require the most secondary correlational analyses to understand what is curative.

Empirical data on manualization. We have argued thus far that the logic of manualization is problematic for many disorders and treatments. So too are the empirical data bearing on assumption that the interventions specified in treatment manuals are causally linked to change. For many brief treatments for many disorders, the lion's share of the effect emerges before the patient has been administered the putatively mutative components of the treatment. For example, most of the treatment effects demonstrated in studies of cognitive therapy for depression occur by the fifth session, with treatment effects leveling off asymptotically after that (Ilardi & Craighead, 1994). Although researchers have challenged these findings (Tang & DeRubeis, 1999), studies using CBT to treat bulimia nervosa similarly have found that patients who do not reduce purging by 70% by the sixth session (prior to most of the interventions aimed at cognitive restructuring) are unlikely to respond to treatment (Agras et al., 2000; see also Wilson, 1999), and recent research with a different treatment, supportive–expressive therapy, has similarly found that sudden gains tend to occur around the fifth session (Asay, Lambert, Gregersen, & Goates, 2002). Similar findings have also emerged repeatedly in naturalistic samples of psychotherapy for patients with a range of problems, who tend to experience a "remoralization" process that restores hope and reduces symptomatology after a handful of sessions (Howard et al., 1993).

Furthermore, therapist adherence to manuals has proven only variably associated with outcome—sometimes positively correlated, sometimes negatively, and sometimes not at all (e.g., Castonguay, Goldfried, Wiser, Raue, & Hayes, 1996; Feeley, DeRubeis, & Gelfand, 1999; Henry, Strupp, Butler, Schacht, & Binder, 1993; Jones & Pulos, 1993)—and correlational analyses have sometimes identified important but unexpected links between process and outcome, such as the finding that focusing on parental issues may be associated with positive outcome in cognitive therapy for depression (Hayes, Castonguay, & Goldfried, 1996). In one study (Ablon & Jones, 1998), researchers used the Psychotherapy Process Q Set (Jones, 2000; Jones & Pulos, 1993) to measure process variables from psychotherapy transcripts of both cognitive and psychodynamic short-term therapies for depression. Not only did therapists of both persuasions use techniques from the other approach (a finding similar to that reported by Castonguay et al., 1996), but in both forms of treatment, positive outcome was associated with the extent to which the treatment matched the empirical prototype of psychodynamic psychotherapy. In this study, the extent to which cognitive therapists used cognitive techniques was actually unrelated to outcome.

In a second study (Ablon & Jones, 1999, 2002), the investigators used the Psychotherapy Process Q Set to study the process of psychotherapy in the National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program (Elkin et al., 1989). They found that both treatments, as actually practiced, strongly resembled the empirical prototype of cognitive therapy, and neither resembled the psychodynamic prototype, even though IPT was derived from the work of theorists such as Sullivan (1953; see also Frank & Spanier, 1995) and is frequently described as a brief psychodynamic variant. Ablon and Jones (1999, 2002) suggested that despite careful efforts at manualization and adherence checks, the NIMH Collaborative Research Program may have compared two cognitive therapies. In this study, adherence to the cognitive therapy prototype was most predictive of change, regardless of which treatment the clinician was attempting to practice.

Another study, using an instrument designed specifically to distinguish CBT and IPT, did find small but significant mean differences between the CBT and IPT conditions on factors designed to distinguish them (Hill, O'Grady, & Elkin, 1992). However, both treatments were best characterized by items designed to assess two nonspecific aspects of treatment characteristic of the control condition, labeled *explicit directiveness* and *facilitative directiveness*.

To what extent similar findings would emerge for other disorders is unknown. We suspect that for specific anxiety disorders, such as simple phobia, specific social phobia, and obsessive–compulsive disorder (OCD), different treatments would be more readily distinguishable. The point, however, is that, as a general assumption, the assumption that the interventions specified in treatment manuals are causally linked to change is not well supported and needs to be demonstrated empirically for a given set of treatments rather than assumed.

Dismantling and the Scientific Testing of Treatment Packages

Another corollary to the assumption that experimental methods provide a gold standard for establishing the validity of therapeutic interventions is that the elements of efficacious treatment are dissociable and hence subject to dismantling. Again, as with the other assumptions and corollary assumptions described here, this one is likely applicable to varying degrees to different treatments and disorders. Dismantling is most readily applied to brief treatments with highly specific procedures, where therapists can adhere closely to a manual and either include or exclude a particular set of interventions, such as cognitive restructuring in exposure-based treatments for OCD or PTSD.

The dismantling assumption is appropriate for RCT methodology (and is indeed one of the advantages of that methodology), but it is invalid as a general rule for EST methodology. The reason lies again in what is being tested, namely treatment packages rather than specific interventions or classes of intervention. Consider, for example, the manual for CBT for bulimia nervosa (Fairburn, Marcus, & Wilson, 1993), which has received considerable empirical support. The manual prescribes that clinicians begin with psychoeducational and behavioral interventions, then move to cognitive interventions, and conclude with interventions aimed at maintenance of change over time. But would treatment as prescribed by this manual as currently configured be superior to the same treatment delivered without the behavioral interventions, or with the order of interventions inverted (cognitive first, behavioral second), or with an initial 5 sessions devoted to alliance building, or with an additional module aimed at addressing interpersonal problems or affect regulation, or simply with the exact same treatment extended to 60 sessions? No one has ever tested, or will ever likely test, any of these variations, even though each of them could be equally justified by theory and might well be more efficacious. The process of selecting the particular package of interventions the investigators selected is, in the philosopher of science Karl Popper's (1959) terms, a *pre*scientific process (i.e., prior to hypothesis testing), and one that has set the agenda for the subsequent scientific process of testing this manual against other treatments and control conditions. Or to use the language of Paul Meehl (1954), it is a prime example of *clinical prediction* (nonquantitative, synthetic judgments about what might work).

The reality is that researchers generally solidify treatment packages (manuals) so early and on the basis of so little hard data on alternative strategies, even within the same general approach, that clinicians have to accept *on faith* that the treatment as packaged is superior to the myriad variants one could devise or improvise with a given patient. It is difficult enough to conduct one or two methodologically rigorous clinical trials with a single manual. To expect researchers to test one or more of the infinite variants of it that could potentially have better efficacy is simply untenable. As we suggest below (see also Beutler, 2000), investigators may do better to focus RCT methodology on the testing of interventions, intervention strategies, and processes of change rather than putatively complete treatments and to strive for guidelines that foster the practice of empirically informed rather than empirically validated psychotherapies.

Science and Prescience: Selection of Treatments to Test as a Source of Bias

Another significant caveat to the assumption that experimental methods provide a gold standard for testing treatments is the problem of determining which treatments to test. One can only separate lead from gold by testing the properties of both. If, as a field, we choose to study only certain kinds of treatments, we cannot draw conclusions about treatment of choice except within the (small) universe of treatments that have received empirical attention. Because of its requirement of brevity and experimenter control, the methodology of ESTs has precluded the testing of treatments widely used in the community, leading to the conclusion that such treatments are empirically unsupported. This conclusion, however, is logically entailed by the method, not determined empirically. Treatments that cannot be tested using a particular set of methods by definition cannot be supported using those methods. Given the powerful allegiance effects documented in psychotherapy research, in which the treatment favored by the investigator tends to produce the superior outcome (Luborsky et al., 1999),[4] perhaps the best predictors of whether a treatment finds its way to the empirically supported list are whether anyone has been motivated (and funded) to test it and whether it is readily testable in a relatively brief format.

Lest the reader object that this is an unfair characterization, consider a recent monograph commissioned by the American Psychological Society (APS) on the treatment of depression (Hollon, Thase, & Markowitz, 2002). As the authors noted, numerous studies have shown that CBT and IPT (and a number of lesser known brands) produce initial outcomes comparable with those obtained with medications. Over the course of 3 years, however, patients who receive these 16-session psychotherapies relapse at unacceptably high rates relative to patients in medication conditions if the latter are maintained on medication during the follow-up period (Hollon et al., 2002). These findings have prompted researchers to test maintenance psychotherapy, which essentially extends brief manualized treatments into long-term treatments.

The results have been promising. As the authors suggested, although monthly IPT maintenance treatment over 3 years does not fare as well as continuous provision of medication, studies testing it have compared low-dose IPT with high-dose imipramine (Hollon et al., 2002), and IPT might do considerably better if provided continuously for 3 years on a more frequent basis. At the end of the monograph, the authors reiterated that CBT and IPT are the psychotherapies of choice for depression but suggested that the wave of the future may be long-term maintenance CBT and IPT:

> Despite real progress over the past 50 years, many depressed patients still do not respond fully to treatment. Only about half of all patients

[4] Luborsky et al. (1999) found that by measuring allegiance in multiple ways, they could account for over 69% of the variance in outcome across a large set of studies by allegiance alone. If one converts their multiple correlation (*R*) of .85 to a binomial effect size (Rosenthal, 1991), the implication is that 92.5% of the time, they could predict which treatment will be most successful based on investigator allegiance alone. Although this may be a liberal estimate, even an estimate one third of this magnitude would have tremendous consequences for the enterprise of testing psychotherapies.

> respond to any given intervention, and only about a third eventually meet the criteria for remission. . . . Moreover, most patients will not stay well once they get better unless they receive ongoing treatment. (Hollon et al., 2002, p. 70)

The authors likened depression to chronic disorders such as diabetes, suggested that depression "may require nearly continuous treatment in order to ensure that symptoms do not return," and concluded with the familiar lamentation that "too few patients have access to empirically supported treatments" (Hollon et al., 2002, p. 70).

If one steps back for a moment, however, the argument appears circular. Thirty years ago, a group of researchers, convinced that the therapies practiced by most clinicians were needlessly long and unfocused, quite reasonably set about to use experimental methods to test more focal treatments aimed at changing explicit thoughts and feelings and current interpersonal circumstances contributing to depression. After an initial 20 years or so of enthusiasm, counterevidence began to amass, first and most importantly from the NIMH Collaborative Research Program (Elkin et al., 1989). The NIMH Collaborative Research Program had an enormous sample size relative to prior studies, and it eliminated two confounds that had rendered interpretation of prior findings difficult: common factors (a confound eliminated by incorporating a rigorous "medical management" placebo control group) and allegiance effects (eliminated by employing investigators at all three sites with allegiance to each of the treatments under investigation). Despite a promising initial response, by 18 months posttreatment, the outcome of brief psychotherapy was indistinguishable from a well-constructed placebo. Subsequent studies (see Hollon et al., 2002) found that 16 weeks of IPT or CBT could not compare in efficacy with a continuous course of medication.

Placed in their broader context, these studies appear to provide a definitive disconfirmation of the central hypothesis that motivated this line of research, namely that depression is amenable to brief psychotherapies, specifically those focusing on explicit cognitive processes or current interpersonal patterns. Yet the authors of the monograph came to a very different conclusion. On the basis of data showing that extending short-term interventions by several months substantially improves outcome, they concluded that only long-term versions *of these short-term treatments* are empirically supportable (Hollon et al., 2002). This conclusion makes sense of the available data, but the available data were predicated on a set of methodological assumptions that presume the disconfirmed hypothesis, that depression is malleable in the face of brief interventions. These methods precluded from the start the testing of the kind of long-term psychotherapies the researchers had set out to show 30 years ago were unnecessarily lengthy, and these methods continue today to preclude the testing of integrative treatments that might address current states and diatheses, explicit and implicit processes, current and enduring interpersonal problems, and so forth (see Westen, 2000).[5] This is not to say that such treatments would turn out to be more effective. That is an unknown. But it will remain unknown as long as treatments are required to fit the requisites of methods rather than vice versa.[6]

Can hypothesis testing be isolated from hypothesis generation? What we are suggesting here is that the influence of prescientific processes can lead to scientifically invalid conclusions despite the safeguards of scientific method imposed at the level of hypothesis testing. Consider again the example of psychotherapy for depression and what might have happened if the NIMH Collaborative Research Program had compared CBT not with IPT but with psychodynamic psychotherapy, which at that time was the most widely practiced psychotherapy in the community. Given the ultimate convergence of the findings from the NIMH Collaborative Research Program with the results of decades of psychotherapy research indicating that brief psychotherapies for depression tend to show similar results as long as they are tested by investigators invested in them (Luborsky et al., 1999; Wampold et al., 1997), what would probably be taught today is that CBT and psychodynamic psychotherapy are the psychotherapeutic treatments of choice for depression.

This example highlights the extent to which the conclusions reached in the EST literature depend on a highly problematic tenet of Popper's (1959) philosophy of science that as a field we have implicitly embraced: that the essence of science lies in hypothesis testing (the context of scientific justification) and that where one finds one's hypotheses (the context of discovery) is one's own business. There can be no more powerful way to create a gulf between clinical practice and research than to compare laboratory-derived interventions with everything but what clinicians practice in the community. The paradoxical effect of doing so is that it places empirically minded clinicians in the position of having to guess, without data, how their own ways of intervening might fare relative to laboratory-based treatments.

The reader may object that a host of studies have compared established therapies with "treatment as usual" (TAU). Unfortunately, TAU comparison groups virtually all consist of low-

---

[5] The reader may object, with some justification, that clinical experience should not dictate the treatments that are tested. We suspect, however, that the failure to test treatments widely used in clinical practice is imprudent, given that clinicians, like other organisms subject to operant conditioning, are likely to learn something useful, if only incidentally, when they peck at a target long enough. They may also develop all kinds of superstitious behavior (as well as false beliefs, illusory correlations, and all the other biases and heuristics that inflict information processors, including clinical information processors), but one should not assume that expertise in clinical work or any other domain leads only to such biases and errors.

[6] One could, in fact, tell a very important story from the data summarized by the authors of the APS monograph (Hollon et al., 2002): that helping patients in an acute depressive episode problem solve, recognize ways they may be inadvertently maintaining their depression, and get moving again behaviorally and interpersonally can, in the context of a supportive relationship, be extremely useful in reducing the severity and duration of depressive episodes (and that some patients can remember and mobilize these resources the next time they become severely depressed). Ellen Frank, who has been one of the most productive contributors to the IPT literature, reached a similar conclusion in one of the most balanced presentations of the results of RCTs of CBT and IPT for depression we have seen (Frank & Spanier, 1995, p. 356). Such a conclusion is, we believe, justified by the available data and is in fact a variation on the theme of the story the authors told. But it requires a substantial shift in aims, from using RCTs to validate treatment packages for depression to using RCTs to assess intervention strategies that may prove useful to clinicians at particular junctures with patients for whom depressive symptoms are clinically significant.

budget, low-frequency treatments with minimally trained paraprofessionals struggling to cope with enormous caseloads (see, e.g., Scheel, 2000). The use of this kind of TAU comparison is not likely to change the mind of many clinicians and in fact should not do so if they understand scientific method, because such conditions do not control for several obvious confounds that render causal inference impossible (e.g., treatment frequency, caseload size, commitment of clinicians to the treatment, and level of training and supervision; see Borkovec & Castonguay, 1998). As suggested below, as researchers, we should exercise more caution in using terms such as *treatment as usual*, *traditional therapy*, or treatment as practiced *in the community* (e.g., Weiss, Catron, & Harris, 2000) if what we really mean is treatment as practiced by masters-level clinicians in community mental health centers (CMHCs) with low-income patients, where notoriously difficult treatment populations intersect with notoriously limited care.

Empirically unvalidated and empirically invalidated. The failure to apply scientific methods to the selection of treatments to subject to empirical scrutiny has contributed to a widespread confusion in the literature, sometimes explicit and sometimes implicit, between empirically untested and empirically disconfirmed, or empirically *un*validated and empirically *in*validated, psychotherapies (Roth & Fonagy, 1996; Weinberger, 2000; Westen & Morrison, 2001). Consider, for example, the following statement from the chapter on CBT for bulimia nervosa in the *Handbook of Treatment for Eating Disorders*:

> Many patients will be somewhat symptomatic at the end of the 19-session manual-based treatment. In our clinical experience, patients in the United States, with its ready availability of different forms of psychological therapy and a tradition of largely open-ended treatment, will often wish to seek additional therapy at the end of the 19 sessions of CBT. We reiterate the caveat issued by Fairburn, Marcus and Wilson . . . about the inadvisability of a rush into further therapy. Patients should be encouraged to follow through on their maintenance plans and to "be their own therapists" as CBT has emphasized. If after a period of some months their problems have not improved, or possibly deteriorated, they can then seek additional treatment. (Wilson, Fairburn, & Agras, 1997, p. 85)

What is clear from this quotation is that the authors do not take an agnostic attitude toward empirically untested treatments practiced in the community. They clearly view the absence of evidence for efficacy of treatments practiced in the community as evidence for absence of efficacy and hence feel confident informing non- or partial-responders (who constitute more than half of patients who undergo CBT or any other brief treatment for bulimia nervosa) that other treatments are unlikely to help them.[7] The authors of the APS monograph on treatment of depression similarly equated untested treatments with inadequate treatments when they concluded that "the empirically supported psychotherapies are still not widely practiced. *As a consequence* [italics added], many patients do not have access to adequate treatment" (Hollon et al., 2002, p. 39).

Incompatibilities Between the Requisites of Experimental Design and Practice

A final problem with the assumption that experimental methods provide a gold standard for separating the clinical wheat from the chaff is the extent to which the requisites of experimental research aimed at identifying ESTs can diverge from the requisites of good treatment, leading to a state of affairs in which the methodological tail wags the clinical dog. Consider again the case of IPT as an empirically supported treatment for bulimia. (We hope readers do not interpret our occasional oversampling of research on bulimia nervosa, which has produced some of the most impressive findings in the treatment literature, as indicative of anything other than our familiarity with it.) When Fairburn, Kirk, O'Connor, and Cooper (1986) conducted their first RCT for bulimia, their explicit goal was to test a cognitive–behavioral treatment previously piloted in an uncontrolled study against a nondirective, nonspecific comparison treatment with some putative credibility (Fairburn, 1997). Thus, they designed a short-term focal comparison treatment, intended as a psychodynamic treatment, in which the therapist first assessed "underlying difficulties" that precipitated the bulimia and then focused on these issues for the remainder of the treatment (Fairburn et al., 1986, p. 632). In their next trial, Fairburn et al. (1991) substituted IPT for the original short-term focal psychotherapy because "it was similar to it in style and focus, but had the advantages of being better known and having a treatment manual available" (Fairburn, 1997, p. 280). In the first four sessions, the role of the IPT therapist was to analyze the interpersonal context in which the eating disorder occurred. Thereafter, "no attention was paid to the patients' eating habits or attitudes to shape and weight" (Fairburn et al., 1991, pp. 464–465). The reason for this injunction was to avoid any overlap with CBT, because the aim of the study was to test the effects of the specific interventions prescribed in the CBT manual.

The results of this second study were unexpected: CBT initially showed the predicted superiority to IPT, but patients in the IPT condition caught up in outcome over the months following termination (Fairburn, 1997; Fairburn et al., 1991, 1993). As a result of the apparent success of IPT in this trial (recently replicated by Agras et al., 2000), Klerman and Weissman (1993) published the IPT manual for treatment of bulimia nervosa. The practice of IPT for bulimia nervosa as summarized by Fairburn (1997) faithfully mirrors the manual designed for experimental use, including "little emphasis on the patient's eating problem as such, except during the assessment stage" (p. 281). The therapist explains to the patient this paradoxical injunction against discussing the symptoms that brought her in for treatment as follows: "This is because focusing on the eating disorder would tend to distract the patient and therapist from dealing with the interpersonal difficulties" (Fairburn, 1997, p. 281).

In fact, the developers of IPT for bulimia did not proscribe discussion of food, body image, eating behavior, or eating attitudes because they or their colleagues had noticed that doing so seemed to be effective. Nor did they do so because they had reason to believe, theoretically or empirically, that talking about eating

[7] This example is not unusual. The national licensing examination in psychology now includes a series of questions about the "correct" treatment for disorders such as depression. Indeed, in an oral examination for licensure, one colleague who indicated that his theoretical orientation was other than CBT was asked why he practiced "an outmoded form of treatment."

behavior should be counterproductive or distracting. Indeed, their own prior controlled trials of CBT had demonstrated just the opposite. The reason the IPT manual proscribes any focus on the symptoms is that doing so made for a clean experiment, in which the effects of the two experimental conditions could be readily distinguished. And when, by accident, IPT turned out to be helpful to many patients, suddenly an experimental manipulation never intended as anything but a credible-enough control found its way into review articles as an EST for bulimia, despite the lack of any empirical evidence for one of its key components (or noncomponents), the counterintuitive injunction against discussing one of the main things the patient came in to talk about.

This example is not an anomaly. It reflects a confusion of two uses of RCT methodology, one reflecting the goal of discovering what kinds of interventions work, and the other reflecting the goal of distinguishing valid from invalid treatment packages. The latter goal becomes particularly problematic in light of the common factors problem (the repeated finding that common factors account for much of the variance in RCTs; see Lambert & Bergin, 1994; Luborsky, Barton, & Luborsky, 1975; Wampold et al., 1997; Weinberger, 1995). A researcher testing a novel intervention in an RCT needs to control for common factors (either by eliminating them from the experimental treatment or using a rigorous control condition that includes them) to test its incremental efficacy, but this does not mean clinicians should do so. Controlling for common factors is essential for causal inference in RCTs but would be counterproductive in clinical practice, given their powerful effects on outcome. As RCTs metamorphosed into tests of the utility of treatment packages taken as a whole, however, researchers had to maximize the purity of their treatments to distinguish them from other treatments or credible controls, leading them to minimize common factors in manuals intended for use by practicing clinicians.

The case of IPT for bulimia (and the fact that CBT for bulimia places limited emphasis on interpersonal problems, reflecting the same effort to minimize treatment overlap with IPT; Wilson et al., 1997, p. 87) is an example of what might be called the *uncommonly differentiated factors paradox* (Westen, 2002): To maximize detection of clinically and statistically significant between-groups effects for ESTs, researchers need to design treatments that are maximally differentiable. Doing so, however, renders them vulnerable to developing treatments that lack precisely the factors that produce much of the effect of brief psychotherapies for many disorders. To put it another way, the demands of experimental investigation in the real world, where researchers cannot easily collect samples of several hundred patients that might help them assess the incremental effects of specific over common factors, often conflict with the demands of clinical practice in the real world. Just as experimenters cannot afford the loss of statistical power that invariably follows from implementation of *impure* treatments, clinicians cannot afford the loss of therapeutic power that follows from implementation of *pure* treatments, particularly where common factors play a role in outcome or where more than one treatment has shown incremental efficacy beyond common factors. If clinicians tend to prefer eclectic or integrative treatments for disorders such as bulimia or depression over treatments that fail to address aspects of their patients' pathology that are obvious to the naked eye but proscribed by one manual or another to maximize their distinctiveness in experiments, they are probably exercising both common sense and good clinical judgment.

Summary: The Assumptive Framework of ESTs

The question of what works for whom is an empirical question that can only be addressed using empirical methods. Yet the effort to identify ESTs has led to the parallel evolution of "practice guidelines" for the conduct of psychotherapy research whose assumptions need to be carefully examined. These assumptions—that psychopathology is highly malleable, that most patients can be treated for a single problem or disorder, that personality is irrelevant or secondary in the treatment of psychiatric disorders, and that a straightforward application of experimental methods as used in other areas of psychology and in research in psychopharmacology provides the primary if not the only way to identify therapeutically useful interventions strategies—appear to be applicable to some degree to some treatments for some disorders. However, when applied indiscriminately, they are likely to lead to substantial error because they are only applicable with substantial qualification and under particular conditions. A central task ahead, and a focus of the final section of this article, is to examine more systematically the conditions under which these assumptions are likely to be accurate or inaccurate, or violated in ways that do or do not produce systematic error.

Retelling the Story: A Reexamination of the Data Supporting ESTs

Thus far we have examined the assumptions underlying the methodology widely assumed to provide the best answers to the question of what works for whom. We now turn to a reconsideration of the empirical findings using this methodology.

Consider a study of cognitive therapy for depression, which illustrates a well-designed investigation of an exemplary EST. Thase et al. (1992) screened 130 patients with depression, of whom 76 were suitable for the treatment protocol, for an inclusion rate of 58%. Of the 76 patients included, 64 (81%) completed the treatment. Of these 64, 23 were described as fully recovered and 27 as partially recovered at the end of treatment, for a full recovery rate of roughly 36% and a partial recovery rate of slightly greater magnitude (42%). At 1-year follow-up, 16 of these 50 fully to moderately successful cases had fully relapsed, leaving 34 at least partially successful treatments at follow-up. When the definition of *relapse* was relaxed to include not only those who developed a subsequent major depressive episode but also those who developed a diagnosable mood disorder short of major depression or who required further treatment, the number who remained improved or recovered fell to 38% of those who entered treatment, or 29 of the 130 who originally sought treatment.

Whether this is the story of an empirically supported or an empirically disconfirmed therapy depends on where one puts the asterisk. In our laboratory we are in the process of completing a series of multidimensional meta-analyses of data from RCTs for a range of disorders, which provide a set of indices yielding information on both outcome and generalizability that we believe are essential for drawing scientifically and clinically meaningful con-

clusions from the literature (Westen & Morrison, 2001).[8] We first briefly describe those variables and then examine the findings with respect to five disorders: major depressive disorder, panic disorder, GAD, bulimia nervosa, and OCD. Next, we place these findings in the context of naturalistic studies recently completed that bear on the external validity of RCTs used to establish treatments as empirically supported. Finally, we consider recent research attempting to address concerns about the external validity of ESTs.

Multidimensional Meta-Analysis: Aggregating a Range of Indicators of Outcome

The most common way of assessing the value of a treatment is to compare mean outcome of treatment, usually (but not always or exclusively) focusing on the symptom or syndrome deemed primary, with pretreatment scores, outcome obtained in a placebo or control conditions, or outcome obtained using another treatment. This method leads to a significance test, which is useful but can be misleading because statistical significance is a joint function of effect size and sample size, so that varying sample size can produce substantial fluctuations in significance values for treatments with equal effects; and to a quantifiable effect size estimate (e.g., Cohen's *d*) that provides a relatively straightforward measure of central tendency that can be readily summarized meta-analytically.

Effect size estimates are essential in evaluating the efficacy of a psychotherapy; however, they have certain limits. Pre–post effect size estimates, though widely reported, are difficult to interpret because passage of time, regression to the mean, spontaneous remission in disorders with fluctuating course, tendency to present for treatment (and hence for research) when symptoms are particularly severe, and other variables not specific to a given treatment can lead to symptomatic change over time. Treatment-control effect size estimates, which do not share these limitations, provide a better estimate of the extent to which a treatment is useful for the average patient. However, they do not provide information on clinically meaningful variation in treatment response. A treatment that has an enormous effect in 20% of patients can appear superior to another treatment that has a smaller but clinically meaningful impact on 90% of patients. These caveats are not meant to "demean the mean," or to devalue probability statistics, only to suggest that mean differences and their corresponding significance values and effect size estimates provide only one measure of efficacy.

A second common index of outcome, readily available in most published reports but rarely aggregated meta-analytically, is percentage improved or recovered. This metric, which we believe is an essential meta-analytic complement to effect size estimates, has a number of variations that need to be distinguished. One variation depends on the numerator (i.e., the number of patients who improved): How does one define clinically significant improvement or recovery? One could, for example, require that patients be symptom free, that their scores on outcome measures fall one or two standard deviations below their original means or within one to two standard deviations of published norms of nonclinical samples, or that they fall below a predetermined cutoff (e.g., the cutoff for major depressive disorder or panic disorder). A considerable body of literature on clinical significance has emerged to attempt to address these issues but has not yet led to any consensus (see, e.g., Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991; Kendall et al., 1999).[9] From a meta-analytic standpoint, the best one can usually do in aggregating across studies is to adopt "local standards" for a given disorder (e.g., rely on the most widely used definitions of improvement in the literature for a particular disorder) and carefully distinguish between improvement and complete recovery. A treatment could, for example, lead to substantial declines in symptoms for most patients but leave all patients symptomatic. That may or may not be an indictment of an experimental treatment, depending on the severity of the disorder, the disability it imposes, and the availability of other treatments.

A second variation in estimating the percentage of patients improved or recovered involves the denominator (the number improved in relation to whom, i.e., success rates divided by what number?). Percentage improved can be calculated relative to the number of patients who complete treatment or the number who entered treatment (intent-to-treat sample). If dropout rates are as high as even 20%—which they usually are—these metrics can yield very different estimates of improvement or recovery.

A third metric that can be readily obtained from most published reports but is rarely noted in reviews is the average level of symptomatology after treatment. A treatment may be highly efficacious in reducing symptoms in the average patient or even in most patients but still leave the vast majority of patients symptomatic. Thus, another way to describe outcome is to look at mean scores on widely used outcome measures or face-valid measures, such as number of panic episodes per week, to assess the absolute value of symptomatology at the end of treatment.

The question of when to measure outcome is as important as how. A key distinction in this regard is between initial response and sustained efficacy. Most nonpsychotic psychiatric conditions show an initial response to a very wide range of psychosocial interventions. Fifteen percent of patients improve significantly after making the initial call to a therapist's office, before attending the first session (see Kopta et al., 1994), and as mentioned above, much of the change seen in many brief therapies occurs within the first few sessions. Whether changes that occur by the fifth of sixth session are durable, and whether they bear any relation to long-term efficacy, is a crucial question.

Thus, a fourth set of indices assess outcome at long-term follow-up intervals. An important distinction at follow-up is between percentage improved or recovered at follow-up, and the percentage that *remained* improved or recovered at follow-up. Many psychiatric disorders are characterized by a course of multiple periods of remission and relapse or symptom exacerbation over many years; hence, knowing whether a patient is better 1, 2, or 5 years later is not the same as knowing that he or she got better as a result of treatment and remained better. Major depression, for example, is an episodic disorder, with an average duration of roughly 20 weeks if left untreated (Judd, 1997). Thus, patients who did not respond to therapy are likely, a year later, to appear

[8] See also McDermut, Miller, and Brown (2001) for an example of the creative application of meta-analytic techniques to a range of metrics important for drawing inferences about efficacy.

[9] For an example of the use of clinically significant change indices in meta-analytic investigations, see McDermut et al. (2001).

improved, recovered, or no longer meeting the diagnostic threshold for major depression, but this says nothing about efficacy, particularly since patients in control conditions are rarely followed for comparison. Data on the percentage of patients who seek additional treatment in the 1 or 2 years following a controlled clinical trial can also be useful in painting a clear portrait of what works for whom. Although treatment seeking can be evidence that patients found a treatment helpful (see Kendall et al., 1999), healthy people typically do not seek further treatment; thus, treatment seeking can provide useful information on incomplete outcomes.

A final set of indices bear on generalizability. One simple metric is the percentage of potential participants excluded at each step of screening (usually once after a phone screen and then again after a structured interview). A second way to provide research consumers with data on the kinds of patients to whom the results of a study can be assumed to generalize is to count the number of exclusion criteria and compile a list of prototypical exclusion criteria across studies. Aside from using this index as a potential moderator variable, researchers can also apply these prototypical exclusion criteria to naturalistic samples to assess the extent to which patients included in RCTs resemble patients with the same disorder treated in clinical practice (and whether comorbid conditions that lead to exclusion of patients from controlled trials are associated in everyday practice with variables such as treatment length and outcome).

Meta-analysis, like any procedure, has its advantages and limits (see Eysenck, 1995; Feinstein, 1995; Rosenthal, 1991; Rosenthal & DiMatteo, 2000), and we do not believe that our approach is without limitations. For example, because we were interested in reexamining conclusions drawn from the published literature, we did not attempt to address the "file drawer" problem by tracking down unpublished studies that might have had null findings, and hence our results are likely to be biased slightly toward positive outcomes. Similarly, too little is written about investigator bias in meta-analysis and the importance of maintaining investigator blindness in making determinations that can substantially affect the findings (Westen & Morrison, 2001). On the other hand, as a field we have known since Meehl's (1954) classic work about the advantages of actuarial over informal, synthetic (in his terms, *clinical*) judgments, and this applies as much to literature reviews as to diagnostic judgments. The best one can do is to present a range of statistics that summarize the data as comprehensively as possible and let readers study the tables and draw their own conclusions.

Efficacy of ESTs for Common Psychological Disorders: A Meta-Analytic Reassessment

In our laboratory, we have thus far completed multidimensional meta-analyses of controlled clinical trials of psychotherapy for five disorders (Eddy, Dutra, & Westen, 2004; Thompson-Brenner et al., 2003; Westen & Morrison, 2001) and are in the process of completing similar analyses for four others. We begin with the take-home message: Empirical support is a matter of degree, which varies considerably across disorders. A dichotomous judgment of empirically supported versus not supported (implicit in the enterprise of constructing a list of ESTs) provides a very crude assessment of the state of the art.

Efficacy of Treatments for Depression, Panic, and GAD

In a first set of studies, Westen and Morrison (2001) examined all studies of ESTs for depression, panic, and GAD published in the major high-quality journals that publish controlled outcome studies of psychotherapy during the 1990s. With the partial exception of treatments for depression, effect sizes for these treatments (in standard-deviation units) were generally impressive, similar to the findings of meta-analyses of psychotherapy published since the pioneering study by Smith and Glass (1977). At termination of treatment, the median effect sizes for depression, panic, and GAD relative to placebo or control conditions were .30, .80, and .90, respectively. Data on percentage of patients improved painted a more variable picture than effect size estimates, depending on the number used as the denominator. Of those who completed treatment, success rates (defined variably across studies, but including patients who improved as well as those who recovered) ranged from 63% for panic to 52% for GAD. Of those who entered treatment (intent-to-treat analysis), improvement rates ranged from 37% for depression to 54% for panic.

Although the average patient improved substantially in active treatment conditions, the average patient also remained symptomatic (Westen & Morrison, 2001). For example, depressed patients completed the average EST with a Beck Depression Inventory (A. T. Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) score above 10, which is above the cutoff for clinically significant pathology using Jacobson and Truax's (1991) criteria for clinical significance (Bouchard et al., 1996). The average panic patient continued to panic about once every 10 days and had slightly over four out of the seven symptoms required for a *DSM–IV* panic disorder diagnosis, enough to qualify for limited-symptom attacks. This is not to diminish the very powerful effects of many of these treatments, especially for panic, given that the average patient began with frequencies of attacks that substantially affected their possibility for life satisfaction. It is simply to suggest that empirical support of validation comes in shades of gray.

For all three disorders, long-term follow-up data were almost nonexistent, and where they did exist, they tended to support only treatments for panic. Across all disorders, by 2 years posttreatment, roughly half of patients in active treatment conditions had sought further treatment. Of those treated for depression, only one third had improved and remained improved over 2 years. For panic, the success rates were higher. Roughly half of patients who entered or completed treatment improved and remained improved. Even for treatments for panic, however, the investigators found that many of the patients who were symptom free at 2 years were not symptom free at 1 year, and vice versa, suggesting a variable course of waxing and waning symptoms for many whose outcomes were generally positive (Brown & Barlow, 1995). For GAD, the authors could locate no data on efficacy at 2 years or beyond.

One question that is difficult to answer because of ethical limitations of keeping patients treatment free for long periods is how treated versus untreated patients fare at extended follow-up. The only study reporting follow-up data at 18 months or longer for both treatment and control conditions was the NIMH Treatment of Depression Collaborative Research Program, which included a relatively active control condition (see Shea, Elkin, et al., 1992). In this study, 78% to 88% of those who entered treatment completely relapsed or sought further treatment by 18 months. Shea, Elkin, et

al. (1992) found no significant differences on any outcome measure among any of the active treatments (cognitive therapy, IPT, and imipramine) and controls at follow-up.

Finally, with respect to generalizability, exclusion rates in Westen and Morrison's (2001) meta-analysis ranged from 65% for GAD to 68% for depression. Thus, the average study excluded two thirds of patients who presented for treatment. Researchers studying all three disorders appropriately excluded patients with psychotic, bipolar, and organic mental disorders; medical conditions that might affect interpretability of results; and those in imminent danger of suicide. However, additional exclusion criteria that were common across studies render generalizability for many of these treatments difficult to deduce. The prototypical study of depression excluded patients if they had suicidal ideation or comorbid substance use disorders, both of which are common symptoms in patients with depression. For panic, prototypical exclusion criteria were moderate to severe agoraphobic avoidance, any concurrent Axis I or Axis II disorder in need of immediate treatment, major depression deemed primary, and recent previous therapy. Prototypical GAD exclusion criteria included major depression, substance use disorders, and suicidality. The fact that such a high percentage of patients had to be excluded across all three disorders suggests that comorbidities of the types excluded may be the rule rather than the exception. Exclusion criteria for all three disorders also tended to eliminate many of the more troubled, comorbid, difficult-to-treat patients, such as patients with borderline features, who are likely to be suicidal and to have substance use disorders.

Efficacy of Treatments for Bulimia Nervosa and OCD

This first set of studies led, we believe, to some important incremental knowledge about the strengths and limitations of treatments currently described as empirically supported, using the best available published studies as data sources. Nevertheless, the meta-analyses of these three disorders had two primary limitations. First, to maximize the quality of the sample and to make the task manageable, Westen and Morrison (2001) included only studies published in major journals with relatively rigorous methodological standards (except for GAD, for which a broader computer search was conducted because of the dearth of studies) and focused on data published in the 1990s to capitalize on methodological advances since the 1980s. Second, because preliminary analyses showed only minor differences in outcome across types of treatment for most of the disorders when enough studies were available to meta-analyze, and because allegiance effects tend to yield higher effect sizes for investigators' preferred treatments (Luborsky et al., 1999), Westen and Morrison did not report findings for specific treatments (e.g., cognitive, cognitive–behavioral, and strictly behavioral treatments for depression). Thus, in subsequent studies, our research team has broadened criteria to include all published studies meeting methodological criteria (experimental methods and randomization of patients) published since publication of the third edition of the *DSM* in 1980 (American Psychiatric Association, 1980). Here we briefly describe the results of the first two such studies, of RCTs for bulimia nervosa and OCD.

Treatments for bulimia nervosa. For bulimia nervosa (Thompson-Brenner et al., 2003), mean effect sizes of treatments compared with controls were substantial (0.88 and 1.01 for binge eating and purging, respectively). However, most patients continued to be symptomatic at the end of treatment. Of those who completed treatment, 40% recovered; of those who entered treatment, 33% recovered. The average patient continued to binge 1.7 times per week and purge 2.3 times per week at the end of treatment. Although this still comes close to the diagnostic threshold for bulimia nervosa in the *DSM–IV*, it nevertheless represents a very substantial improvement from baseline. Findings at 1-year follow-up, though hard to come by, were similar to posttreatment data, with the average patient across treatments showing substantial improvement over pretreatment baseline but also substantial residual symptomatology. However, only one third of patients across treatments or in individual CBT (which tended to fare slightly better than other treatments, particularly group CBT) showed sustained recovery at 1 year (i.e., recovered at termination and remained recovered at 1 year).

With respect to exclusion rates and criteria, the average study excluded 40% of the patients screened. Approximately half the studies excluded patients for either low or high weight (excluding patients with both anorexic symptoms and obesity) and suicide risk, and an additional one third excluded patients for substance abuse or dependence (31%). A large number of studies also excluded patients who had "major psychiatric illness," "serious comorbidity," or similar nonspecific exclusion criteria.

Treatments for OCD. For OCD (Eddy et al., 2004), as reported in previous meta-analyses (e.g., Cox, Swinson, Morrison, & Paul, 1993; Kobak, Greist, Jefferson, Katzelnick, & Henk, 1998; van Blakom, van Oppen, Vermeulen, van Dyck, & Harne, 1994), effect sizes were very high, averaging 1.50 to 1.89 depending on the outcome measure, and were uniformly high across treatment conditions (behavioral, cognitive, and cognitive–behavioral). Approximately two thirds of patients who completed treatment improved (defined variously as 30% to 50% reduction in symptoms), and one third recovered. Among the intent-to-treat sample, about one half improved and one fourth recovered. Posttreatment scores on standardized instruments suggested that the average patient experienced substantial improvement but also remained symptomatic.

Eddy et al. (2004) intended to meta-analyze follow-up data as in the previous studies, but only two studies included follow-up at or beyond 1 year, and both of these reported data using the last observation carried forward, which does not allow readers to distinguish between data collected at 12 weeks and 12 months for patients who were inaccessible for follow-up. With respect to generalizability, few studies reported on the percentage of patients screened out, but among those that did, the average study excluded 62% of patients. The most common exclusion criteria were substance abuse and a variety of co-occurring disorders that varied across studies.

A counterpoint: Randomized trials of psychopharmacology for bulimia and OCD. One useful way of contextualizing these findings is to apply the same metrics to pharmacological interventions for the same disorders, which our laboratory has done thus far for both bulimia and OCD. Psychopharmacology for bulimia appears to be useful in many cases as an adjunctive treatment, but outcomes obtained using medication alone do not compare with the results of psychotherapies such as CBT (Nakash-Eisikovits et al., 2002). The average effect sizes for bulimia nervosa were 0.64 for binge episodes and 0.59 for purge episodes, slightly over half the effect sizes for psychotherapy. Although many patients improved, few recovered (slightly over 20%). On average, at termi-

nation patients binged 4.3 times a week and purged 6.2 times weekly, which represents roughly twice the posttreatment means for CBT.

The data on psychopharmacological treatments for OCD are much more encouraging, with outcomes that rival behavioral and cognitive–behavioral interventions for the same disorder (Eddy et al., 2004). Treatment-control effect sizes in this study, as in prior meta-analyses of the same literature, were large (e.g., for clomipramine, which outperformed the other medications, $d = 1.35$). Almost two thirds of patients who completed and half of those who entered a medication trial improved. As in the psychotherapy studies, however, recovery is a rare event, and high exclusion rates and the absence of follow-up data at clinically meaningful intervals rendered clinically meaningful conclusions more difficult to come by in this and virtually every other psychopharmacological literature we have examined.

In many respects, meta-analysis of the results of medication trials, like psychotherapy trials, underscores the problems inherent in making dichotomous determinations of empirical support or nonsupport when the data call for more nuanced appraisals. Medication for bulimia is useful, but only in certain ways for certain patients. The data on medication for OCD are much stronger in terms of effect size, but medication rarely leads to cure for OCD. Whether to call pharmacological treatments for one of these disorders empirically supported and the other empirically unsupported is unclear because they are each efficacious and inefficacious in their own ways, if to differing degrees.

Comparing ESTs and Naturalistic Studies of Psychotherapy

Naturalistic studies of treatment in the community provide another useful context for assessing the findings of RCTs for ESTs (see, e.g., Asay et al., 2002; Kopta et al., 1994; Seligman, 1995). Our research team recently followed up the meta-analyses described above with naturalistic studies of several disorders designed to shed light on external validity. Naturalistic studies as implemented thus far (including our own) tend to have a number of methodological shortcomings, such as nonrandom assignment of patients and lack of experimental control. However, they provide a window to phenomena not readily observed in the laboratory and can be particularly useful both for hypothesis generation and for providing a context within which to interpret data from RCTs, particularly data bearing on external validity.

Morrison, Bradley, and Westen (2003) began with a simple naturalistic study involving 242 clinicians randomly selected from the registers of the American Psychiatric and American Psychological Associations as participants. Approximately 20% of clinicians contacted (1/3 psychiatrists and 2/3 psychologists) returned completed materials, for which they received no compensation. (Despite their differential response rates, psychologists and psychiatrists provided highly similar data, suggesting that response rates do not likely account for the bulk of the findings.) The clinicians tended to have multiple institutional affiliations: 31% worked in hospitals at least part time, 20% worked in clinics, 82% worked in private practice, and 11% worked in forensic settings. Respondents were a highly experienced group, with 18 years of posttraining experience being the median.

As a follow-up to our research team's first set of multidimensional meta-analyses, Morrison et al. (2003) asked responding clinicians to describe their last completed psychotherapy with three patients: one who presented with clinically significant depressive symptoms, one who presented with clinically significant panic symptoms, and one who presented with clinically significant anxiety symptoms other than panic. (The decision was made to widen the diagnostic net to include patients with clinically significant depression, panic, and anxiety because the emphasis was on generalizability to the clinical population.) Clinicians provided information about length of treatment, Axis I comorbidity, and Axis II comorbidity. They also completed a checklist of other clinically significant personality variables found in previous research to be frequent targets of therapeutic attention, such as problems with intimacy, relatedness, or commitment in close relationships; difficulty with assertiveness or expression of anger or aggression; authority problems; problems with separation, abandonment, or rejection; and so forth (Westen & Arkowitz-Westen, 1998).

Two findings are of particular relevance from the present point of view. First, median treatment length ranged from 52 sessions for panic to 75 sessions for depression. When Morrison et al. (2003) stratified clinicians by theoretical orientation (psychodynamic, cognitive–behavioral, and eclectic, which were the primary orientations in the sample), the briefest treatments, not surprisingly, were cognitive–behavioral. Even these treatments, however, were almost twice as long on the average as manualized CBTs for the same disorders. It may be the case, of course, that these therapies were long relative to ESTs because clinicians were inefficient or influenced by monetary incentives to retain patients longer than necessary. However, the consistent finding in RCTs for these disorders that the average patient remains symptomatic at the end of a trial of brief psychotherapy and seeks further treatment suggests that clinicians were likely responding to the fact that patients continued to manifest clinically significant symptoms.

Second, comorbidity was the norm rather than the exception. As in previous community and clinical samples, roughly half of patients for each disorder had at least one comorbid Axis I condition, and slightly less than half had an Axis II disorder. The data on depression are illustrative: Half of patients had at least one comorbid Axis I condition, half had at least one comorbid Axis II disorder, and virtually no clinician reported treating any patient exclusively for depression when completing the personality problems checklist. For example, 67% of clinicians described their patients diagnosed with depression as suffering from clinically significant problems with intimacy, relatedness, or commitment in relationships that the patient and clinician agreed was causing substantial distress or dysfunction, and 77% of clinicians reported clinically treating the patient for clinically significant problems with assertiveness or expression of anger. These percentages were invariant across therapeutic orientations, appearing with virtually identical frequencies regardless of clinicians' theoretical preconceptions, and were systematically related to treatment length. Across disorders and theoretical orientations, average treatment length doubled when patients had any form of Axis I comorbidity or Axis II comorbidity, and the presence of clinically significant personality problems also predicted treatment length. For example, presence of externalizing pathology (an aggregate variable from

the personality problem checklist) was strongly associated with treatment length ($r = .40$).

This first set of studies (Morrison et al., 2003), though providing useful data on common comorbidity and on current practices in the community, had several limitations. It focused only on successful treatments, so we could not assess the relation between comorbid conditions and outcome; it provided no data on the interventions used by clinicians, so that we could not rule out the possibility that clinicians were simply using inefficient strategies; it was retrospective, leaving open the possibility of reporting bias; and the treating clinician was the only source of data.

Thompson-Brenner and Westen (2004a, 2004b) addressed the first two of these problems in a subsequent study in which they asked a random national sample of clinicians to describe their most recently terminated patient with clinically significant bulimic symptoms, including treatment failures. The demographics of the clinicians were similar to the first study. Patients in the study averaged 28 years of age, and were, like the population from which they were drawn (women with eating disorders), primarily middle class and Caucasian.

Although most clinicians described their patients as improved over the course of treatment, only 53% of patients completely recovered (a percentage that was similar across all theoretical orientations). Once again, clinicians of all theoretical orientations reported treating patients for much longer than the 16 to 20 sessions prescribed in the most widely tested and disseminated manuals. Although CBT treatments were of shorter duration than eclectic/integrative and psychodynamic treatments, the average CBT treatment lasted 69 sessions on average, substantially longer than the 19 prescribed in the manual.

Comorbidity was also the rule rather than the exception, and both Axis I and Axis II comorbidity were negatively associated with treatment outcome. Over 90% of the sample met criteria for at least one comorbid Axis I diagnosis other than an eating disorder. Axis II comorbidity was also high: One third of the sample met criteria for at least one Cluster B (dramatic, erratic) diagnosis, and the same proportion met criteria for at least one Cluster C (anxious) diagnosis. Several comorbid Axis I disorders (notably major depressive disorder, PTSD, and substance use disorders) and Axis II disorders (borderline, dependent, and avoidant) commonly seen in patients with eating disorders were positively correlated with treatment length and negatively correlated with outcome, with small to medium effect sizes. When Thompson-Brenner and Westen (2004a) applied four common exclusion criteria from RCTs to the naturalistic sample (substance use disorder, weight 15% or more over ideal, weight 15% or more below ideal, and bipolar disorder), they found that approximately 40% of the naturalistic sample would have been excluded (the same percentage excluded in the average RCT). Two thirds of the patients with BPD would have been excluded on the basis of these criteria, and the 40% of patients who would have been excluded (whether or not they had BPD) showed worse treatment outcome across a number of indices.

Finally, Thompson-Brenner and Westen (2004a, 2004b) measured intervention strategies by asking clinicians to complete an interventions questionnaire adapted from Blagys, Ackerman, Bonge, and Hilsenroth (2003). Factor analysis of the interventions questionnaire yielded three factors: Psychodynamic, Cognitive–Behavioral, and Adjunctive interventions (e.g., pharmacotherapy, hospitalization). Across the entire sample, greater use of CBT interventions was associated with more rapid remission of eating symptoms, whereas greater use of Psychodynamic interventions was associated with larger changes in global outcome, such as Global Assessment of Functioning (American Psychiatric Association, 1994) scores. Clinicians of all theoretical backgrounds reported using more Psychodynamic interventions when treating patients with comorbid pathology, which is perhaps not surprising given that these interventions are more oriented toward personality. Psychodynamic clinicians reported using more CBT interventions (such as structuring the therapy hours) with emotionally constricted patients. In contrast, CBT clinicians reported using more Psychodynamic interventions (e.g., exploring patterns in relationships, exploring sexuality, and exploring unconscious processes) when treating emotionally dysregulated patients (i.e., those with borderline features, substance use disorders, etc.). These data suggest that clinicians of all theoretical orientations attend to personality and comorbid symptomatology and adjust their intervention strategies accordingly. An important point of note is that clinicians did not appear reluctant to describe unsuccessful cases or to self-report the use of interventions explicitly associated with their nonpreferred theoretical orientation, suggesting that the data cannot simply be reduced to clinician bias.

We do not consider the data from these naturalistic studies by any means definitive. The exclusive reliance on clinicians as respondents, the retrospective design, and the use of a brief therapy process measure completed by the clinician without independent verification by external observers impose severe constraints on what we can conclude. We also do not know whether the patients in these naturalistic studies fared better or worse than patients in the ESTs examined in Thompson-Brenner et al.'s (2003) meta-analysis, except by clinicians' own report (slightly greater than 50% recovery from bulimia nervosa at termination). That question can only be answered by comparing outcome in naturalistic and manualized treatments for similar patients and including shared outcome measures. The data are consistent, however, with the dose–response relationship found in virtually all naturalistic studies, which shows that patients tend to show greater improvement with more extensive treatment, particularly when they have characterological problems (Howard et al., 1986; Kopta et al., 1994; Seligman, 1995). Perhaps most important, the data suggest limitations in manualized treatments designed to address specific Axis I syndromes that do not address enduring personality dispositions relevant to these syndromes to which clinicians of all theoretical orientations attend and which are not readily explained in terms of sampling or response bias.

Studies Testing the Transportability of ESTs

A critic might object that the data presented thus far do not address the amassing literature on the transportability of ESTs to more naturalistic settings. Reading the contemporary literature, one is indeed impressed with how rapidly and successfully researchers have responded to the clarion call for research addressing critics' concerns about the generalizability of treatments tested in the laboratory (e.g., Persons & Silberschatz, 1998). Within a short span of years, a number of effectiveness and benchmarking studies have found manualized treatments to be highly transport-

able, with little if any decrement in effect size or response rates (see, e.g., Chambless & Ollendick, 2000). This emerging consensus is somewhat surprising, given that researchers have presumably been imposing relatively stringent exclusion criteria in RCTs for 20 years for a reason. We have no doubt that many manualized treatments (and, more broadly, many interventions tested in RCTs) will ultimately show substantial transportability. However, at this juncture, we suspect that the best way to advance knowledge of what works for whom would be to begin testing in a systematic way the conditions under which particular treatments or interventions are likely to be useful in everyday practice, rather than to try to make dichotomous judgments about transportability.

Consider some of the studies now widely cited as evidence for transportability. One showed an impressively low relapse rate at 1-year follow-up for patients treated with CBT for panic in a community mental health setting (Stuart, Treat, & Wade, 2000). Another examined patients excluded from RCTs conducted at the same site (a superb comparison sample to address the question) to assess the transportability of exposure-based treatment for OCD (Franklin, Abramowitz, Levitt, Kozak, & Foa, 2000). This study produced outcomes (average pre–post effect sizes above 3.00) that exceeded the mode in the investigators' own table of benchmark RCTs. Although we suspect that these are indeed two of the most portable of all the ESTs—and the data may well reflect the robustness of these treatments—both studies suffered from a substantial limitation that has not been noted in any review of which we are aware, namely nonblind assessment.[10]

Another recent study, designed to test the hypothesis that CBT for depression in children and adolescents is superior to treatment in the community, used a creative benchmarking design to compare treatment response among child and adolescent patients with depression treated at what appear to have been six inner-city CMHCs with the average response of patients of similar age and severity of depression treated in RCTs (Weersing & Weisz, 2002). Patients in the RCTs showed much more rapid improvement, although outcome converged by 1-year follow-up. The authors concluded that the treatment trajectories of CMHC-treated youth "more closely resembled those of control condition youth than youth treated with CBT" in RCTs, and drew implications about the transportability and benefits of manualized treatments relative to "the effectiveness of community psychotherapy for depressed youth" (Weersing & Weisz, 2002, p. 299). They noted that the "CMHC services were predominantly psychodynamic, whereas therapists in clinical trials provided a pure dose of CBT" (Weersing & Weisz, 2002, p. 299) and suggested that these treatment differences likely accounted for much of the difference in outcome.

Several features of Weersing and Weisz's (2002) study, however, suggest caution in drawing even preliminary conclusions about transportability of manualized therapies or about their superiority to therapies practiced in the community. Although the authors reported no data on socioeconomic status of the CMHC sample, their description of the sample suggests that they compared a low-socioeconomic status CMHC sample with a set of benchmark studies of primarily Caucasian, presumably less socioeconomically disadvantaged, patients. The authors noted that mood disorder diagnoses did not differ substantially between the CMHC and benchmark samples, nor did severity of depression, suggesting equivalence of the samples. However, our calculations from data provided in tabular form in their article indicate very different rates of comorbid conditions known to influence outcome in children and adolescents. Frequency of conduct disorder and oppositional defiant disorder averaged only 11% in benchmark studies that either reported or excluded these diagnoses but averaged 61% in the CMHC sample. Rates of anxiety disorders were 25% and 58%, respectively. Whereas some investigators studying generalizability from RCTs to clinical practice have maximized comparability of samples by applying the same inclusion and exclusion criteria to the community sample (e.g., Humphreys & Weisner, 2000; Mitchell, Maki, Adson, Ruskin, & Crow, 1997; Thompson-Brenner & Westen, 2004a, 2004b), Weersing and Weisz (2002) excluded patients only if they "were unable to complete study measures as a result of psychosis or developmental disability" (p. 301). Although the intent appears to have been to maximize external validity (which would have been appropriate if the goal were to compare CBT in the laboratory with CBT in a CMHC sample), we are unaware of any RCT for depression in either adults or children with comparable inclusion criteria. Thus, any obtained results could reflect differences in the samples, differences in the treatments delivered, or both.

Sampling issues aside, of particular note was Weersing and Weisz's (2002) conclusion that patients treated in RCTs not only did better than those treated in the community but that the CMHC patients actually looked more like patients in the control conditions than in experimental conditions in controlled trials. The authors rested this conclusion on an extrapolation from the slope of change in RCT control conditions from pretreatment to 3 months, arguing that by 12 months, continuation of this slow but steady reduction of symptoms would have yielded data indistinguishable from the treated CMHC patients (i.e., return to normalcy). Such a procedure, however, would dissolve any treatment effect ever documented at 1 year for any EST for child or adult depression of which we are aware. The authors explained the convergence in outcome at 1 year between benchmark and CMHC patients by suggesting that improvement in the CMHC group likely reflected the natural course of the illness (i.e., gradual waning of a depressive episode), which it may well have. However, this explanation is notably different from the conclusion typically drawn from the same finding when obtained in RCTs, for which converging outcomes at 1 year for IPT and CBT for bulimia, for example, have been interpreted as demonstrating a delayed treatment effect for IPT.

[10] In the panic study, follow-up assessment was conducted by graduate research assistants who knew that all patients being followed up had been in the active treatment condition. In the OCD study, the sole OCD outcome measure reported was a semistructured interview, with the interviewers presumably aware both that all patients had been treated with the same therapy and that the purpose of the study was to demonstrate ecological validity of this treatment. In both studies, secondary measures such as the Beck Depression Inventory, which are less likely to be contaminated by experimenter expectancy effects, provided promising corroborating data but unfortunately did not directly address the target symptom. Surprisingly, however, the rival explanation of experimenter bias, which rendered both studies perhaps more comparable to open label trials than controlled clinical trials, was not discussed as a potential limitation in either research report, and both have been cited frequently as evidence for the transportability of manualized treatments to clinical practice.

Our point here is not to criticize particular studies or investigators but simply to note the danger of confirmatory biases when a community of scientists feels some urgency to respond to published critiques with creative demonstration projects that enter into the empirical lore as disconfirmations of the critiques. At this point, we believe the most scientifically appropriate way to resolve the question of generalizability of ESTs is a moratorium on exclusion criteria in RCTs other than those a reasonable clinician might apply in everyday practice (e.g., reliably documented organic brain disease) or that are medically or ethically necessary and the use of correlational analyses to identify moderators of outcome worth studying in future investigations. Either exclusion criteria are necessary, and generalizability from RCTs is correspondingly limited, or exclusion criteria are unnecessary, and studies using them have a serious sampling flaw and should not be published in first-tier journals or cited without substantial qualification. The near-universal exclusion of patients with suicidality from clinical trials is no exception: If, as a field, we truly believe that the state of our science justifies the kinds of distinctions between empirically supported and unsupported treatments for disorders such as depression that are leading to massive shifts in training, practice, and third-party payment, it is neither scientifically nor ethically justifiable to relegate suicidal patients to unvalidated treatments, at least in studies that could randomly assign patients with suicidal ideation to two or more active interventions.

Summary: How Valid Are Empirically Validated Psychotherapies?

The data from RCTs of treatments widely described as empirically supported for depression, panic, GAD, bulimia nervosa, and OCD over the past decade suggest that these treatments do indeed lead to substantial initial reductions in painful mood states and pathological behaviors that substantially affect life satisfaction and adaptive functioning. The treatments we have studied meta-analytically have proven as or more effective than pharmacotherapies that have had the advantage of billions of dollars of industry support for testing and marketing. At the same time, the existing data support a more nuanced and, we believe, empirically balanced view of treatment efficacy than implied by widespread use of terms such as *empirically supported*, *empirically validated*, or *treatment of choice*.

As a discipline we have clearly identified a set of techniques that can be helpful to many patients suffering from many disorders. However, the effects of brief manualized psychotherapies are highly variable across disorders, with some findings justifying claims of empirical support and others reminding us that in science, empirical support is usually not a dichotomous variable. With the exception of CBT for panic, the majority of patients receiving treatments for all the disorders we reviewed did not recover. They remained symptomatic even if they showed substantial reductions in their symptoms or fell below diagnostic thresholds for caseness; they sought further treatment; or they relapsed at some point within 1 to 2 years after receiving ESTs conducted by clinicians who were expert in delivery of the treatment, well supervised, and highly committed to the success of their treatment of choice.

The extent to which even the more qualified conclusions offered here can be generalized to the population of patients treated in the community is largely unknown because of high, and highly variable, exclusion rates and criteria that render the findings from different studies difficult to aggregate and apply to the treatment-seeking population. In the modal efficacy study, somewhere between 40% and 70% of patients who present for treatment with symptoms of the disorder are excluded (not including the unknown percentage of patients who are screened by referring clinicians who know the study's exclusion criteria), and patients treated in everyday clinical practice who resemble those excluded tend to take longer to treat and to have poorer outcomes.[11] The data from naturalistic studies suggest that, in fact, most patients are polysymptomatic, and the more co-occurring conditions with which the patient suffers, the longer and more wide-ranging the treatment appears to be—a conclusion similarly reached in a qualitative review by Roth and Fonagy (1996). The correlation between treatment length and comorbidity is one of the few generalizations that appears to apply across treatments and disorders. The polysymptomatic nature of patient pathology in clinical practice suggests that a primary or exclusive clinical focus on a single Axis I syndrome does not appear to be appropriate in the majority of cases, particularly if polysymptomatic cases have any emergent properties that cannot be reduced to variance accounted for by each symptom or syndrome in isolation.

A striking gap in the literature is the relative absence of follow-up studies that span the length of time during which relapse is known to be common in untreated patients for the disorders in question. More encouraging findings are on the horizon for some treatments, such as data suggesting that CBT can help prevent relapse and recurrence of panic following discontinuation of alprazolam several years after treatment (Bruce, Spiegel, & Hegel, 1999) or that cognitive therapy for depression may help prevent relapse in a subset of patients (see Hollon et al., 2002). However, the limited data on long-term outcome of ESTs suggest that initial response may or may not bear a relationship to efficacy at 2- to 5-year follow-up (e.g., Shea, Elkin, et al., 1992; Snyder, Wills, & Grady, 1991). Treatments that differ in initial response may yield highly similar efficacy estimates at 2 years, and treatments that appear similar in initial response may have very different outcomes at 5 years.

Rewriting the Story: Implications for Evidence-Based Practice

Although the story we have told thus far has been critical of many of the uses and interpretations of controlled clinical trials, we are not arguing against the utility of RCTs or experimental methods more generally in psychotherapy research. This final section has two goals. The first is to suggest standards for reporting aspects of design and findings in RCTs to maximize their clinical utility and applicability. The second is to examine ways research-

[11] These findings do not appear to be limited to the disorders we have studied thus far. Humphreys and Weisner (2000) recently applied the prototypical exclusion criteria taken from alcohol treatment studies to two large community samples to assess the external validity of efficacy trials for alcoholism. The resulting samples were highly unrepresentative, heavily composed of Caucasian, stable, higher functioning patients with less substantial comorbidity—the kinds of patients who, empirically, are likely to respond more favorably to treatment.

ers might design rigorous studies of interventions that do not violate the assumptions outlined in the first part of this article.

Maximizing the Efficacy of Clinical Trials

In reviewing hundreds of studies over the course of several meta-analyses, we have observed a number of common problems in reporting that limit the impact and applicability of many controlled trials to everyday clinical practice. Here we describe some of these problems and potential solutions. Most of the suggestions we offer may appear obvious, and many have been detailed elsewhere (e.g., Chambless & Hollon, 1998) but are not yet normative in practice.

Describing the Hypotheses and Experimental Conditions

A central issue that has received little attention in the literature involves the framing and reporting of hypotheses. Because the interpretation of most statistical tests depends on whether the analysis is planned or post hoc, it is essential that researchers clearly label their hypotheses as primary, secondary, a priori, post hoc, one-tailed, or two-tailed. Nothing is wrong with unexpected findings, but readers need to know which hypotheses were predicted. Of particular relevance in this regard is the clear labeling of comparison and control groups. Researchers need to specify clearly and a priori whether a condition is intended as an active, credible treatment; a dismantled component of a treatment that may or may not prove useful; a credible control that is likely to be superior to no treatment because of common factors; a weak control that largely controls for number of sessions and some very nonspecific factors; or a simple wait-list or similar group that controls only for the passage of time.

The frequent use of wait-list and TAU controls in RCTs for many disorders can lead to substantial problems of data interpretation (Borkovec, 1994; Borkovec & Castonguay, 1998). Given that virtually any 10- to 20-session intervention will produce an initial response in most patients if carried out by a clinician who expects it to be effective (Luborsky, McLellan, Diguer, Woody, & Seligman, 1997), the only scientifically valid conclusion one can draw from observed differences between experimental and wait-list control conditions is that doing something is better than doing nothing, yet it is remarkable how many literatures rely primarily on such comparisons and how many investigators draw much more specific conclusions (see Kazdin, 1997).

As noted earlier, TAU conditions are often interpreted as demonstrating the superiority of experimental treatments to everyday clinical practice, but this is only a valid conclusion if control therapists are well paid, motivated, and see the patients on a regular basis. This is seldom the case. The literature on DBT, for example, has relied almost exclusively on TAU comparisons (Scheel, 2000), but DBT therapists (who are available around the clock, provide several hours a weeks of group and individual therapy, undergo weekly supervision, etc.) and community mental health practitioners differ on so many variables that it is difficult to draw causal conclusions from such comparisons. Our own suspicion is that DBT is a highly efficacious treatment for many aspects of borderline pathology, but the frequent use of TAU conditions renders confidence in this conclusion weaker than it likely need be.

Researchers should also exercise caution in labeling control treatments not constructed to maximize their efficacy (what Wampold et al., 1997, described as *non-bona fide* treatments, or what might be called *intent-to-fail* conditions) with brand names that are readily confused with genuine treatments and create sleeper effects in the literature. For example, as described earlier, to test the efficacy of CBT for bulimia, Garner et al. (1993) developed a treatment they called *supportive–expressive therapy*, an abbreviated treatment described as nondirective and psychodynamically inspired, in which clinicians were forbidden to discuss the target symptoms with the patient and were instead instructed to reflect them back to the patient. Such a practice is not in fact characteristic of psychodynamic therapies for eating disorders (e.g., Bruch, 1973) and is analogous to a researcher creating a cognitive therapy comparison condition in which the therapist is instructed to say, "That's irrational," every time a patient tries to discuss the symptom. Unfortunately, this study is frequently cited in review articles as demonstrating that CBT is superior to psychodynamic psychotherapy for bulimia nervosa (e.g., Compas, Haaga, Keefe, & Leitenbert, 1998).

A similar example can be seen in the graph on the cover of the APS monograph on the treatment of depression (Hollon et al., 2002), which was adapted from a practice guidelines document published almost a decade earlier (a document whose purported biases had in fact contributed to the movement among psychologists to construct a list of ESTs). The graph showed response rates above 50% for IPT, CBT, and medication, compared with response rates hovering slightly above 30% for placebo and psychodynamic therapy. Given the absence of any manual for psychodynamic psychotherapy for depression (at least ca. 1993, the publication date of the practice guidelines), it is unclear what treatments one would include in such a graph. As best we could ascertain from the authors' thorough (and very balanced) review in the text of the monograph, the treatments summarized in the graph for psychodynamic therapy were largely intent-to-fail controls. By the methodological standards used elsewhere in the monograph and in the EST literature more generally, the most appropriate conclusion to draw (and the one the authors in fact drew in the text) is that there are no credible data either way on psychodynamic therapy for depression and certainly none for the longer term therapies most commonly denoted by that term. Unfortunately, the message all but the most careful readers are likely to take away from the monograph is that psychodynamic therapy for depression is empirically invalidated, not unvalidated.

Reporting, Justifying, and Interpreting the Data in the Context of Inclusion and Exclusion Criteria

Another important issue pertains to the reporting, justification, and interpretation of findings in light of inclusion and exclusion criteria. As we argued earlier, as a field we would do well to stop using exclusion criteria other than those that are medically necessary or similar to those a clinician might be expected to apply in everyday practice (e.g., brain damage) if our goal is to guide practice. If researchers impose criteria other than those that are obviously necessary, they should routinely provide the following information. In the Method section, they need to describe and justify precisely what the criteria were and whether they made these decisions prior to the study, prior to examining the data, after

noticing that certain kinds of patients had anomalous results, and so forth. In the Results section, they should describe how many patients were excluded at each step (e.g., both at initial phone screen and upon interview) and for what reasons. In the Discussion section, they should describe the precise population to which they expect the results to generalize. The abstract should also be qualified if the study excluded patients a reader might readily assume were included, such as depressed patients with suicidal ideation. As noted earlier, only a minority of research reports describes the number of patients excluded at each level of screening and how and when decisions were made, and fewer still carefully limit their conclusions to what is usually a subpopulation. By the time the results find their way into reviews or meta-analyses, qualifications about patients excluded tend to disappear.

Describing the Clinicians

Another aspect of reporting that requires greater attention regards the clinicians who conducted the treatments. Investigators need to describe clearly the number and characteristics of the treating clinicians in each condition and to describe and justify any choices that could bear on interpretation of the results. Given the small number of clinicians who can participate in any RCT, there are no right answers to the question of how to select therapists. For example, having different therapists in two treatment conditions creates the possibility of uncontrolled therapist effects, whereas using the same therapists in two conditions raises questions about allegiance effects. Crits-Christoph and Mintz (1991) reviewed 140 clinical trials of psychosocial treatments and found that 26 used only one therapist, one pair of therapists working together, or one therapist per treatment, which rendered treatment effects and therapist effects confounded. Seventy-seven studies made no mention of therapist effects; and 32 conducted one-way analyses of variance to rule out therapist effects but set the significance level at the conventional .05, which left the analyses seriously underpowered to detect even substantial effects.

In a large percentage of the studies we reviewed, it was difficult or impossible to ascertain precisely how many therapists conducted the treatments, who these therapists were, and whether they had commitments to one approach or another. Frequently (especially in smaller studies) the first author, who was expert in and committed to the treatment approach under consideration, appeared to be the therapist in the active treatment condition, but the report provided no data on whether the clinician or clinicians in other conditions were similarly expert and motivated. As Luborsky and others have noted (e.g., Luborsky et al., 1999), differing levels of expertise and commitment provide one of the likely mediators of allegiance effects. Protecting against this threat to validity would probably require multiallegiance research teams to become the norm in treatment research.

Assessing Psychopathology and Outcome

The reliability and validity of assessment is obviously crucial in treatment studies and extends to the way investigators assess the primary diagnosis required for inclusion in the study, the diagnosis of comorbid conditions, and outcome. With respect to reliability and validity of diagnosis for inclusion, clinical trials for many anxiety disorders are exemplary (e.g., J. G. Beck, Stanley, Baldwin, Deagle, & Averill, 1994; Borkovec & Costello, 1993; Bouchard et al., 1996; Shear, Pilkonis, Cloitre, & Leon, 1994). Less optimal reporting, however, is common in clinical trials for many other disorders. For example, of 26 studies reviewed for a meta-analysis of psychotherapies for bulimia, many reported data on the reliability and validity of the assessment instrument in general (e.g., as used by its developers), but none that used a structured interview reported data on interrater reliability in the study being reported (Thompson-Brenner et al., 2003).

Even when researchers establish the primary diagnosis carefully, they generally pay less attention to the diagnosis or reporting of comorbid conditions or exclusion criteria (see Strupp, Horowitz, & Lambert, 1997). This is understandable given that diagnosis of other conditions is generally considered secondary to the point of the study. Because exclusion of patients on the basis of comorbid diagnoses or other criteria is crucial to generalizability, however, assessment of exclusion criteria can be just as important as reliable assessment of the target syndrome. For example, many studies require that patients have a primary diagnosis of the index disorder to be included in the clinical trial. However, researchers virtually never describe how, how reliably, and when in the screening process they made that determination of primacy. This is again a potential avenue for unnoticed allegiance effects and threats to generalizability that cannot be detected from reading published reports, as interviewers or screeners may determine that the problems of more difficult patients are primary and hence lead to their exclusion. Because comorbid conditions can have implications for treatment outcome, even studies that do not exclude patients for common forms of comorbidity need to provide reliability data on diagnosis of comorbid conditions that might bear on clinical utility outside of the laboratory.

With respect to the measurement of outcome, the primary focus of most outcome studies is outcome vis-à-vis the syndrome under investigation, and appropriately so. At the same time, outcome studies should always supplement assessment of primary symptom measures in four ways, some of which are becoming more common (e.g., measures of high end-state functioning or global adaptation) and some of which are rarely used. First, because most patients have multiple problems that affect their adaptation and life satisfaction, studies should routinely include measures of other Axis I conditions, adaptive functioning, and quality of life. Second, given the strong links between Axis I conditions and personality, efficacy studies should routinely include measures of relevant personality variables, particularly where data are available suggesting that these variables may be diatheses that render the patient vulnerable to future episodes. Third, given the growing evidence distinguishing implicit from explicit processes (and linking implicit processes to underlying vulnerabilities; e.g., Segal et al., 1999), studies should routinely include measures designed to assess implicit networks or implicit attentional biases (e.g., emotional Stroop tasks) that may indicate the likely durability of changes. Finally, studies that include extended follow-up should implement reliable and systematic ways of assessing posttermination treatment seeking. Given the demand characteristics inherent in participating in a controlled trial (e.g., wanting to please the therapist by reporting feeling better at the end), behavioral indices are crucial to supplement self-reports, and one of the most useful

such indices is whether the patient seeks further treatment (and whether he or she seeks it from the investigators or from someone else).

In the studies we reviewed, the reliability, validity, and demand characteristics of outcome measures for even the symptoms considered primary was highly variable. Many studies of anxiety and eating disorders, for example, use symptom diaries as primary outcome measures (Arntz & van den Hout, 1996; Ost & Westling, 1995; Shear et al., 1994). The benefits of diary-type recording are clear: Patients can observe and record panic or eating behavior as it occurs, rather than relying on memory during an interview that may occur weeks or months after the fact. At the same time, diaries impose a number of problems that are rarely controlled or addressed. Because diaries are often used as treatment tools in cognitive–behavioral treatments, the accuracy and calibration of diary reports are likely to differ across conditions if one condition repeatedly uses diaries throughout treatment and the other does not. Furthermore, in many treatments, patients and clinicians use diaries to chart progress, leading to demand characteristics as the treatment draws to an end if the diary is used as a primary outcome measure. How diaries are collected is another relevant issue. In several studies we reviewed, the final outcome diary appears to have been collected directly by the therapist, which could obviously influence patients' symptom reports.

Diaries are not the only measures subject to potentially differential calibration or demand characteristics across conditions. In several studies we reviewed, psychoeducational components of the treatment could well have biased the way respondents used self-report measures. For example, when clinicians provide education on the precise meaning of *panic attack* or *binge* in one experimental condition, patients' assessment of the frequency of these events is likely to change irrespective of behavioral change, particularly if clinicians are motivated to see improvement and subtly bias the severity required to "count" as an episode. Bouchard et al. (1996) addressed this potential problem in a creative way by explicitly educating all participants on the definition of a panic attack before they began keeping diaries. Such practices should be standard. At the very least, researchers need to report explicitly how and when patients in each condition are provided information that might impact the way they respond to outcome questionnaires for symptoms whose definition is at all ambiguous.

Tracking Relevant Ns

As suggested earlier, tracking precisely how many patients have come through each stage of a study is essential for assessing both efficacy and generalizability (see Table 1). In this respect, psychotherapy researchers should follow the lead of medical researchers, who have recently developed consolidated standards for the reporting of trials that provide guidelines for reporting sample sizes at each stage, including providing flow diagrams (Egger, Juni, & Bartlett, 1999). For the modal study, in which the investigators use both an initial telephone screen and a subsequent, more extensive screen via structured interview, research reports need to describe the percentage excluded at each step and the reasons for exclusion. Rarely do researchers provide data on exclusion at both of these two points (indeed, the majority of studies we have reviewed did not include either), which is essential for assessing generalizability. Many exemplary cases of reporting exist in the literature, however, that should serve as models for standard practice (e.g., Barlow, Gorman, Shear, & Woods, 2000; Fairburn et al., 1991; Ost & Westling, 1995).

Definitions of dropout and completion also deserve greater attention. A large number of published reports use idiosyncratic definitions of completion, and it is often unclear whether these definitions were set a priori. This is another potential locus for intrusion of allegiance effects, as researchers may, with no conscious intention of deception, select definitions of completion that tell the best story for one condition or another. Including relatively early dropouts as *completers*, for example, can bias results in either a positive or a negative direction, depending on whether the last observation carried forward is high or low. Consider a comparison of cognitive therapy, applied relaxation, and imipramine for depression by Clark et al. (1994). Although the intended length of the psychosocial treatments was 15 sessions, the investigators defined *dropouts* as those who completed no more than 3 sessions; thus, patients who attended only 4 out of 15 sessions would be considered completers. This could be a very conservative way of ana-

Table 1
The Complex Meanings of N: Indispensable Data for Interpreting the Results of Clinical Trials

Stage of patient participation	*N* necessary to report
Participants assessed for eligibility	Estimated *N* in the community or treatment center who were likely to have seen recruitment or referral advertisements or notices
	N who responded and received phone screen
	N who passed initial phone screen and received a subsequent interview
Participants excluded	*N* who failed to meet each specific inclusion criterion
	N who met inclusion criteria but met each specific exclusion criterion
Participants included	*N* randomized to each condition
	N who began each condition; reasons, if known, why any patients randomized did not enter treatment
	N who completed each condition; reasons, if known, why any patients did not complete each intervention
	N who completed each condition but did not participate in assessments; reasons, if known, why patients were not available for assessment
	N analyzed, for each analysis
	N excluded from each analysis and reasons, including exclusion of outliers

lyzing the data, retaining the last observation carried forward of patients who did not find the treatment useful or tolerable. Alternatively, it could inflate completion rates or include data from patients who made immediate improvements that had little to do with any specific treatment components. Decisions such as this are rarely discussed or justified in published reports. Many of the reports we reviewed across disorders not only failed to explain decisions about completer definitions but used different definitions in different analyses. The only reason we even noticed this problem was that we were meta-analyzing data that required us to record *N*s, and noticed different *N*s in different tables or discrepancies between the *N*s reported in the tables and the text.

Reporting and Interpreting Results

One of our primary conclusions in this article has been the importance of reporting a range of outcome statistics and indicators of generalizability that allow readers, reviewers, and meta-analysts to draw more accurate and nuanced conclusions. As a number of researchers have persuasively argued, primary research reports should always supplement tests of statistical significance with effect size estimates such as Cohen's *d* or Pearson's *r* (see, e.g., Rosenthal, 1991). Others have noted, however, that even these effect size estimates can sometimes fail to represent clinical significance (e.g., Jacobson & Truax, 1991). The investigation that demonstrated the value of aspirin in reducing heart attacks so clearly that the study had to be discontinued for ethical reasons produced an *r* of only .03—and an R^2 less than .01. Although seemingly minimal to null, this effect size translates to 15 people out of 1,000 who will live or die depending on whether they take aspirin (Rosenthal, 1995). The debate about how to measure clinical significance is an important one that will likely continue for some time. In the meantime, we recommend that all published RCTs report, at minimum, each of the metrics described earlier.

Several other data reporting and interpretation issues also require attention. One involves pretreatment group differences in symptom levels that occur despite randomization. Much of the time, researchers call attention to and report attempts at controlling for accidental pretreatment group differences (e.g., J. G. Beck et al., 1994; Ost & Westling, 1995). However, particularly in studies with small samples, researchers need to be careful to note and address potentially important pretreatment differences that may exceed one to two standard deviations but not cross conventional significance thresholds (e.g., Wolf & Crowther, 1992).

An additional set of reporting issues seems obvious but represents such a widespread problem in the literature for both psychotherapy and psychopharmacology that it requires vigilance by reviewers and editors, namely the clear reporting of sample and subsample sizes, treatment of outliers, and appropriate measures of central tendency. In reviewing studies across disorders, it was surprising how frequently we had difficulty ascertaining which subgroups of the sample were used in different analyses (e.g., intent-to-treat vs. completer samples). When researchers present intent-to-treat analyses, they should also state whether the analyses used the last observation carried forward, particularly in follow-up analyses. More generally, tables and analyses reported in the text should always include the *N* or degrees of freedom and indicate whether participants were excluded and for what reasons. If outliers are deleted, investigators need to state explicitly whether (and why or why not) they deleted the corresponding outliers at the other end of the distribution or the same number of participants from the other treatment conditions. Similarly, researchers should always report means and standard deviations where appropriate, and if they switch to medians in some analyses, they should justify the reasons for doing so. A surprisingly large number of reports did not include pretreatment and posttreatment means for the primary outcome measures. In some cases research reports included only figures or charts without the raw numbers, which makes meta-analysis of findings difficult. In other cases, researchers reported means without standard deviations, which are impossible to interpret, except if the investigators also provided all the relevant *F* values, *p* values, degrees of freedom, and so forth that could allow a motivated reader to calculate the size of the effect.

Finally, several issues concerning reporting of follow-up data deserve attention. As noted earlier, the dearth of follow-up data for ESTs over extended periods is a serious problem with the existing literature, as it is for the medication literature for the same disorders. Researchers routinely refer to treatments that show short-term effects in multiple studies as empirically supported, when they know only about initial response and effects at brief follow-up intervals (e.g., 3 to 9 months). Including a plan for maintaining contact with patients after treatment should be a prerequisite for funding.

When researchers conduct follow-ups at multiple intervals (e.g., at 3, 6, and 12 months posttreatment), they need to report their data with and without noncompleters and with and without the last observation carried forward. A typical practice is to follow up only the completers (or responders) and to carry forward the last follow-up observation (e.g., at 3 months) if the patient is no longer available to the researchers at later intervals such as 12 months (e.g., Foa et al., 1999). Unfortunately, doing so renders the 12-month data uninterpretable because these data are contaminated by 3-month data on what is often a sizeable proportion of patients. Researchers also need to be very cautious in the way they summarize follow-up findings in abstracts and reviews. If they follow up only completers or responders, this needs to be clearly stated in summarizing follow-up findings in the abstract.

Not only is the interval between termination and follow-up important, but so is the timeframe patients are asked to consider in follow-up assessments (Brown & Barlow, 1995). Most long-term follow-up studies ask patients to report whether they experienced the target symptoms over a very brief span of the recent past, usually 1 to 4 weeks. This is likely to be an unreliable sample of behavior over the course of a 12- to 24-month period and can lead to the mistaken impression that a treatment shows sustained efficacy. The problem is exacerbated by the lack of comparison to controls, who are usually not assessed at long-term follow-up for ethical reasons (e.g., they have received treatment in the interim).

To maximize both reliability of reporting (e.g., because patients may not accurately recall precisely how many times they purged per week over the past 6 months) and comprehensiveness, the most sensible course may be for researchers to assess both the immediate past (e.g., symptoms in the past month) as well as the entire period since termination or the last follow-up assessment. Dimensional symptom assessment and assessment of related diagnoses (e.g., not-otherwise-specified diagnoses) should also be standard, given that many patients may no longer meet threshold for a

disorder but still show residual symptoms that are clinically meaningful and bear on genuine efficacy.

Selecting an Appropriate Design: RCTs and Their Alternatives

We have suggested that reporting changes could maximize the utility of controlled trials of psychotherapy. Better reporting, however, can only address some of the problems we have outlined. To the extent that the assumptions underlying efficacy trials are violated in ways that threaten the robustness of the conclusions, researchers need to turn to other methods, or at least to triangulate on conclusions using multiple methods. We focus here, first, on the conditions under which RCTs are likely to be useful and second, on strategies that may prove useful when RCTs are not.

When RCT Designs Are Useful

Throughout this article, we have suggested a distinction between RCT methodology and EST methodology, the latter denoting a particular use of the former. An important question regards the conditions under which researchers can use RCT designs without making the assumptions that derail many EST designs relying on RCT methodology. We note here two such conditions.

First, the assumptions of EST methodology are only minimally violated for particular kinds of symptoms and particular kinds of interventions. (In these cases EST methodology and RCT methodology are for all intents and purposes identical.) The symptoms or syndromes that least violate the assumptions of EST methodology involve a link between a specific stimulus or representation and a specific cognitive, affective, or behavioral response that is not densely interconnected with (or can be readily disrupted despite) other symptoms or personality characteristics. Prime examples are simple phobia, specific social phobia, panic symptoms, obsessive–compulsive symptoms, and PTSD following a single traumatic experience (particularly one that does not lead to disruption of foundational beliefs about the world, such as its safety, beneficence, etc.; see Janoff-Bulman, 1992). In each of these instances, patients should, theoretically, be able to obtain substantial relief from an intervention aimed at breaking a specific associative connection, regardless of whatever other problems they may have, presuming that these problems do not interfere with compliance or other aspects of the treatment. Empirically, exposure-based treatments of such disorders have in fact produced many of the most impressive results reported over decades of psychotherapy research (see Roth & Fonagy, 1996). Syndromes that involve *generalized* affect states, in contrast, violate virtually all of the assumptions of EST methodology: They are highly resistant to change, they are associated with high comorbidity, they are strongly associated with (if not constitutive of) enduring personality dispositions, and efficacy trials testing treatments for them have typically required secondary correlational analyses to make sense of the findings. Perhaps not coincidentally, empirically, brief manualized treatments for these disorders tend to fail to produce sustained recovery at clinically meaningful follow-up intervals for any but a fraction of the patients who receive them.[12]

With respect to treatments, those that readily lend themselves to parametric variation (and hence to genuine dismantling) and to the degree of within-condition standardization required for causal inference in RCTs are least likely to violate the assumptions of EST methodology. Such treatments are brief, highly prescriptive, and comprise only a small number of distinct types of intervention (e.g., exposure, or exposure plus cognitive restructuring). Any treatment that (a) requires principle-based rather than intervention-based manualization, (b) prescribes a large set of interventions from which clinicians must choose on the basis of the material the patient presents, or (c) allows the patient to structure the session will introduce too much within-condition variability to permit the optimal use of EST designs. A convergence of theory and data should now be apparent: The studies that have yielded the best results in the psychotherapy literature are those that have targeted syndromes that least violate the assumptions regarding comorbidity and personality inherent in EST methods, applying treatments that least violate the requisites of experimental design as applied in EST methodology.

Second, even where many of the assumptions of EST methods are violated, researchers can still make considerable use of RCT designs to assess specific intervention strategies or principles, general approaches to treatment, and moderators of outcome. Rather than assuming the burden of demonstrating that they have developed a complete package for treating depression that is superior to any other package for the heterogeneous population of patients who become depressed, investigators might address the more modest goal of testing whether a specific intervention strategy (e.g., challenging dysfunctional explicit beliefs) is associated with a clinically significant reduction in depressed mood, and if so, for how long. The goal of establishing a list of "approved" treatments has led to a primary focus on main effects (e.g., "cognitive therapy is an empirically supported treatment for depression"), with the assumption that once researchers have done the hard work of identifying main effects, they can turn their attention to studying moderators and interactions (i.e., the conditions under which the effect holds or does not hold). Although sensible from one point of view, this strategy poses serious dilemmas for clinicians, who need to know today, rather than 10 or 20 years from now, whether they should try a given treatment for depression with patients who are acutely suicidal, have BPD, have chronic low-level depression rather than major depression, and so forth.

In contrast, a focus on testing specific interventions allows researchers to move more quickly from main effects to clinically meaningful questions. For example, does the intervention produce effects that last for hours, days, weeks, or months? Does it work for patients who are mourning a specific loss? Does it work for individuals who have recently become unemployed and suffered a loss in self-esteem? Does it work for patients with various forms of comorbidity? The goal of a study of this sort is not to test a complete treatment package for a specific disorder that would need

[12] In retrospect, the reasons for this seem clear: Neither 2 decades of psychotherapy research nor any other data of which we are aware suggest that any treatment is likely to change lifelong patterns ingrained in neural networks governed by Hebbian learning principles in a brief span of hours, particularly when these patterns are highly generalized or serve affect-regulatory functions. The neural networks governing these disorders likely extend so far and wide that models and methods appropriate for targeting specific associative links are likely to be a poor fit.

to be modified and tested in modified form with every new subpopulation. Rather the goal is to isolate intervention strategies that clinicians can integrate into their practice when working with a patient for whom depression is a prominent symptom.

Perhaps one of the most important uses of RCTs would be to establish, empirically, the length of a given treatment required to produce a clinically meaningful response, a relatively enduring change, and so forth, vis-à-vis a particular set of outcome variables. For example, consider what might have happened if, in the 1980s, having found cognitive interventions to be useful in the first two or three studies, researchers had systematically compared outcome at 1, 2, and 5 years of 16, 52, 100, and 200 sessions of cognitive therapy for depression. This question, of course, reflects the benefits of hindsight, but it points to important implications for the ways researchers could maximize the use of clinical trials in the future.

The use of RCT designs we are advocating here is clearly more limited than the use propounded in the EST literature, although several prominent RCT methodologists have come to very similar conclusions (Borkovec & Castonguay, 1998; Kazdin, 1997). If the goal is to study specific interventions or mechanisms of change, RCT designs may be more limited still if the disorder is one for which the malleability assumption does not apply and if the goal is to address diatheses for disorders as well as current states or episodes. For example, RCT methods can be very useful in identifying interventions designed to curtail (or perhaps forestall) a depressive episode, because effects of the intervention on target symptoms are proximal (within weeks or months), and natural variation in patients (if exclusion criteria are minimized) should allow for testing of moderators of treatment response (e.g., presence of comorbid conditions). If, however, the goal is to test interventions useful for treating depression over the long run or treating psychological diatheses for depression (e.g., negative affectivity, emotional dysregulation, or deficits in self-esteem regulation), it is unlikely that any small set of relatively brief, readily operationalizable procedures will produce a large enough signal to be detected across several subsequent years of noise (except, perhaps, through secondary correlational analyses), particularly given the likelihood that patients will seek other forms of treatment in the interim. Matters are even more complicated for polysymptomatic presentations with substantial personality diatheses, which are unlikely to show enduring change in response to brief, targeted interventions. In such cases, psychotherapy researchers will need to supplement RCTs with alternative designs or to tailor experimental methods to real-world clinical problems in ways that go beyond adding an extra 6-session module to a 12- or 16-session treatment.

Alternatives to RCT Designs

The question, then, is how to supplement RCT designs to converge on findings that are both scientifically rigorous and clinically relevant (see also Beutler, 2000; Lambert, Hansen, & Finch, 2001; Seligman, 1995). We argue that doing so requires a rethinking of the relation between science and practice in clinical science and a reconsideration of ways researchers and clinicians may be able to collaborate to spin clinical yarn(s) into empirical gold.

A transactional approach to knowledge generation in clinical science. Designing treatments and methodologies to match the clinical complexities of polysymptomatic patients, personality diatheses, and symptoms that are resistant to change requires, we believe, rethinking some pervasive assumptions not only about methodology in psychotherapy research but also, more broadly, about the relation between science and practice. Perhaps most important is the need to reconsider the top-down, unidirectional model of science and practice that has become increasingly prevalent in recent years, which assumes that knowledge flows primarily from researchers to clinicians. This assumption is implicit in the EST movement, in frequently voiced concerns about the need for better dissemination or marketing of well-tested manuals that clinicians do not seem to use (e.g., Addis, Wade, & Hatgis, 1999; Wilson, 1998), in clinical scientist models that are rapidly replacing scientist–practitioner models in the most prestigious clinical psychology graduate programs (e.g., McFall, 1991), and in the prevailing view of effectiveness research as a second stage of research in which laboratory-proven treatments are implemented in the community. The implicit metaphor underlying this view is essentially a biomedical one (see Stiles & Shapiro, 1989), in which researchers develop new medications, which pharmaceutical companies then market to physicians, who are perceived as consumers.

In some cases this is an appropriate metaphor, as when developments in basic science lead to novel therapeutic procedures. Perhaps the best example can be seen in exposure-based treatments that emerged from research on classical conditioning. In other cases, however, as a field we might be better served by a more transactional philosophy of clinical science, in which the laboratory and the clinic are both seen as resources for hypothesis generation and testing, albeit with different strengths and limitations. As researcher–clinicians, we can sometimes capitalize on our knowledge of relevant basic and applied literatures as well as our clinical observation, and those who happen to be talented clinical observers as well as talented researchers (names like Barlow, Borkovec, and Linehan come to mind) are likely to generate novel approaches to treatment that few who live on only one side of the mountain would have been likely to develop. On the other hand, those of us who practice both research and psychotherapy (or just research) cannot, over time, match the sheer number of hours full-time clinicians spend with their patients that allows them greater opportunities for uncontrolled observation, innovation, and the kind of trial and error that constitutes the most scientific aspect of clinical practice.

The reality is that many, if not most, of the major clinical innovations in the history of our field have come from clinical practice. Cognitive therapy did not emerge in the laboratory. It emerged from the practice of a psychoanalyst, Aaron T. Beck (and from converging observations of another psychoanalyst, Albert Ellis), whose clinical data convinced him that the psychoanalytic methods of the time were too inefficient and whose *clinical empiricism*—that is, trial and error and careful assessment of the results—led him to develop a set of more structured, active procedures that he and others subsequently put to more formal experimental test. What has perhaps distinguished A. T. Beck over the course of his career has been his ability to integrate what he and others have seen in the consulting room with both basic and

applied research and his willingness, in a true scientific spirit, to change his theories and techniques when the available data—clinical and empirical—suggest the importance of doing so.

Using practice as a natural laboratory. We argued earlier that a failure to use systematic procedures to determine which treatments to test ultimately undermines any conclusions reached, however rigorously one tests interventions of interest. One way of selecting treatment strategies more systematically is to use clinical practice as a natural laboratory. Thus, as investigators, we might take advantage of the wide variation that exists in what clinicians do in practice and use correlational analyses to identify intervention strategies associated with positive outcome, initially and at multiple-years follow-up. Instead of requiring individual investigators to predict, on the basis of their best guesses and theoretical preferences, which treatments are most likely to work, this approach allows us to extend scientific method to the context of discovery. Once we have identified potentially useful interventions (and moderators) correlationally, we can then set our experimental sights on the interventions that appear most likely to pay off *as well as* on experimentally derived interventions that have received support in the laboratory.

Consider what such a design might look like in practice. An investigative team enlists the support of a large sample of clinicians who agree to participate and enlist participation of their next 3 patients with a given problem or disorder (e.g., clinically significant depression, bulimia nervosa, substance abuse, low self-esteem, narcissistic personality disorder, negative affectivity, emotional dysregulation), irrespective of comorbid conditions or whether one disorder or another appears primary. The researchers may choose to study a random sample of clinicians, or they may use a peer nomination procedure to select clinicians whose peers (within or across theoretical orientations) consider master clinicians.

At the beginning of treatment and at periodic intervals thereafter, the investigators obtain data on symptoms, personality, and adaptive functioning from clinicians, patients, and independent interviewers. To assess what is happening in the treatment, clinicians audiotape two consecutive sessions at regular intervals until the treatment is over, which the investigators code for process–intervention variables using an instrument such as the Psychotherapy Process Q Set (Jones & Pulos, 1993). For each of the next 5 years, the investigators examine the process–intervention profiles of patients in the upper and lower 25th percentile on important outcome measures to see which items, singly and in combination, are predictive of success or failure. They may develop a single prototype of what effective treatment looks like (by simply aggregating data across all treatments in the upper 25th percentile), or use techniques such as Q-factor analysis to identify prototypes of successful treatments (on the assumption that more than one strategy may be effective and that mean item ratings may conceal two or more effective ways of working with patients that are very different or even polar opposites). If the sample is large enough, the investigators may develop different prototypes for different kinds of patients (e.g., those with particular personality styles, levels of severity of depression) or different profiles for different points in the treatment (e.g., the use of more structured interventions early, when depressive or bulimic symptoms are most acute, and the use of more exploratory interventions later). Clinicians would treat patients as they normally do, with no limits on number of sessions or type of interventions.

To move to an experimental phase that permits the testing of causal hypotheses, the investigators then use a similar design, but this time they add an experimental condition in which a group of randomly selected clinicians from the same sample is supervised to match the prototype of a successful treatment derived from the correlational phase of the study. Thus, clinicians in this condition might receive regular feedback on audiotaped hours, including identification of items from the Psychotherapy Process Q Set on which they diverged by at least one standard deviation from the empirically generated prototype and consultation on how they might alter their technique to approximate the prototype. Aside from a no-supervision condition, an additional comparison group might receive regular feedback on the Q-sort profile of their treatment without comparison to any prototype, or supervision by a master clinician of their theoretical orientation. The investigators could then compare outcomes in these conditions to see whether clinicians supervised to match the empirical prototype were able to do so and whether doing so was associated with greater success. In so doing, researchers could use clinical practice not only to generate hypotheses but to test them using experimental methods. They could also take specific interventions associated with positive outcomes in the correlational phase of the research, particularly interventions with relatively proximal correlates, back to the laboratory to refine and test.

This approach represents an inversion of the relation between efficacy and effectiveness designs as currently understood. Rather than starting with pure samples and treatments in the laboratory and then taking only those with proven efficacy into the community, this strategy works in precisely the opposite direction: It begins with unselected samples in the community and then turns to experimental designs, in the community, in the laboratory, or both. This strategy essentially retains the advantages of relatively open-ended case studies in the process of scientific discovery but reduces their disadvantages by aggregating across cases from the start. Westen and colleagues have applied a similar approach to basic science research on psychopathology by quantifying clinicians' observations of their patients and then aggregating across cases to generate ways of classifying disorders empirically (e.g., Westen & Shedler, 1999a, 1999b; Westen, Shedler, Durrett, Glass, & Martens, 2003). For both basic and applied practice network research of this sort (see also, e.g., Asay et al., 2002; Borkovec, Echemendia, Ragusea, & Ruiz, 2001; West, Zarin, Peterson, & Pincus, 1998), the goal is not to survey clinical opinion (e.g., about what works or how to classify psychopathology) but to quantify data from clinical practice in such a way as to derive scientifically valid generalizations across cases. Using this approach does not assume the wisdom of any given clinician's clinical experience. Rather, it assumes that variability among clinicians will allow the investigator to identify and distill clinical wisdom empirically.

Summary: Maximizing the Efficacy of Psychotherapy Research

To summarize, we are not arguing against the use of controlled clinical trials to assess the efficacy of psychotherapies. We are arguing, rather, that as a field, we need to make better use of them, and to triangulate on conclusions with other methods for which

RCT designs are not optimal. For disorders and treatments for which the use of RCTs is appropriate, we need to apply standards of reporting with which most scientists would agree in theory but are not routinely implemented in practice. For disorders and treatments for which the assumptions of EST methodology are violated, we need not abandon scientific method. We can make better use of RCT designs if we focus on intervention strategies and change processes rather than treatment packages, and if we focus on target symptoms likely to change within relatively brief intervals. Where RCT designs provide only limited information, we should take advantage of one of the greatest resources at our disposal, clinical practice, which can and should serve as a natural laboratory for both generating and testing hypotheses. Correlational designs can extend the reach of scientific method into the context of discovery, identifying promising interventions that can then be tested experimentally, both in the community and the laboratory.

Conclusion: Toward Empirically Informed Psychotherapy

A reconsideration of both the assumptions and the findings of RCTs generally interpreted as evidence for the validity of a specific set of brief manualized treatments suggests the need for both a more nuanced view of outcome and a reexamination of the enterprise of compiling a list of empirically supported psychotherapies. Inherent in the methodology that has been shaping the agenda for clinical training, practice, licensing, and third-party reimbursement is a series of assumptions that are violated to a greater or lesser degree by different disorders and treatments. These assumptions include symptom malleability, incidental comorbidity, dissociation between symptoms and personality dispositions, and a one-size-fits-all model of hypothesis testing. For disorders characterized by readily identifiable, maladaptive links between specific stimuli or representations and specific responses, and treatments capable of a kind of manualization that allows genuine experimental control, these conditions are minimally violated. Not coincidentally, these are the disorders and treatments that have generated the clearest empirical support using RCT methodology: exposure-based treatments for specific anxiety symptoms (as well as many behavioral treatments of the 1960s and 1970s, which focused on specific problems such as speech anxiety and assertiveness).

For most disorders and treatments, however, the available data suggest that the need to rethink the notion of empirical support as a dichotomous variable, a notion on which practice guidelines comprising a list of validated treatments is implicitly predicated. The average RCT for most disorders currently described as empirically supported excludes between one third and two thirds of patients who present for treatment, and the kinds of patients excluded often appear both more representative and more treatment resistant in naturalistic studies. For most disorders, particularly those involving generalized symptoms such as major depression or GAD, brief, largely cognitive–behavioral treatments have demonstrated considerable efficacy in reducing immediate symptomatology. The average patient for most disorders does not, however, recover and stay recovered at clinically meaningful follow-up intervals.

Reporting practices in the psychotherapy literature also require substantial changes to maximize the benefits of RCT designs. Frequently consumers cannot obtain details essential for assessing the internal and external validity of even high-quality studies. These details, particularly relating to external validity, tend to be absent from qualitative and quantitative reviews as well, leading to conclusions that are often underqualified and overgeneralized.

Despite frequent claims in the literature about treatment of choice, few data are available comparing manualized treatments with treatment as usual for patients with the financial resources to obtain treatment from experienced professionals in private practice, who may or may not provide as good or better care. What *is* known is that treatments in the community tend to be substantially longer than treatments in the laboratory, regardless of the therapist's theoretical orientation, and that in naturalistic samples, more extensive treatments tend to achieve better results according to both patient and clinician reports. To what extent this is a methodological artifact is unknown. For the polysymptomatic patients who constitute the majority of patients in the community, researchers and clinicians need to collaborate to make better use of natural variations in clinical practice to identify interventions associated empirically with good outcomes and to subject correlationally identified intervention strategies to experimental investigation to assess their potential causal impact.

Rather than focusing on treatment packages constructed in the laboratory designed to be transported to clinical practice and assuming that any single design (RCTs) can answer all clinically meaningful questions, as a field we might do well to realign our goals, from trying to provide clinicians with step-by-step instructions for treating decontextualized symptoms or syndromes to offering them empirically tested interventions and empirically supported theories of change that they can integrate into *empirically informed treatments*. This realignment would require a very different conception of the nature of clinical work, and of the relation between science and practice, than is current in our discipline, where researchers and clinicians often view each other with suspicion and disrespect (see Goldfried & Wolfe, 1996). Perhaps most important, it would require the assumption of clinically competent decision makers (rather than paraprofessionals trained to stay faithful to a validated manual) who have the competence to read and understand relevant applied *and* basic research, as well as the competence to read people—competencies we suspect are not highly correlated. Learning how to create such clinicians will, we suspect, prove at least as challenging as designing treatments for them to conduct.

References

Abelson, R. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Ablon, J. S., & Jones, E. E. (1998). How expert clinicians' prototypes of an ideal treatment correlate with outcome in psychodynamic and cognitive-behavioral therapy. *Psychotherapy Research, 8,* 71–83.

Ablon, J. S., & Jones, E. E. (1999). Psychotherapy process in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology, 67,* 64–75.

Ablon, J. S., & Jones, E. E. (2002). Validity of controlled clinical trials of psychotherapy: Findings from the NIMH Treatment of Depression Collaborative Research Program. *American Journal of Psychiatry, 159,* 775–783.

Addis, M. E., Wade, W. A., & Hatgis, C. (1999). Barriers to dissemination

of evidence-based practices: Addressing practitioners' concerns about manual-based psychotherapies. *Clinical Psychology: Science and Practice, 6,* 430–441.

Agras, W. S., Crow, S. J., Halmi, K. A., Mitchell, J. E., Wilson, G. T., & Kraemer, H. C. (2000). Outcome predictors for the cognitive behavior treatment of bulimia nervosa: Data from a multisite study. *American Journal of Psychiatry, 157,* 1302–1308.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arntz, A., & van den Hout, M. (1996). Psychological treatments of panic disorder without agoraphobia: Cognitive therapy versus applied relaxation. *Behaviour Research and Therapy, 34,* 113–121.

Asay, T. P., Lambert, M. J., Gregersen, A. T., & Goates, M. K. (2002). Using patient-focused research in evaluating treatment outcome in private practice. *Journal of Clinical Psychology, 58,* 1213–1225.

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice-Hall.

Barlow, D. (1996). The effectiveness of psychotherapy: Science and policy. *Clinical Psychology: Science and Practice, 1,* 109–122.

Barlow, D. (2002). *Anxiety and its disorders* (2nd ed.). New York: Guilford Press.

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

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Beck, J. G., Stanley, M. A., Baldwin, L. E., Deagle, E. A., & Averill, P. M. (1994). Comparison of cognitive therapy and relaxation training for panic disorder. *Journal of Consulting and Clinical Psychology, 62,* 818–826.

Beutler, L. E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting and Clinical Psychology, 66,* 113–120.

Beutler, L. E. (2000). David and Goliath: When empirical and clinical standards of practice meet. *American Psychologist, 55,* 997–1007.

Beutler, L. E., Moleiro, C., & Talebi, H. (2002). How practitioners can systematically use empirical evidence in treatment selection. *Journal of Clinical Psychology, 58,* 1199–1212.

Blagys, M., Ackerman, S., Bonge, D., & Hilsenroth, M. (2003). *Measuring psychodynamic-interpersonal and cognitive–behavioral therapist activity: Development of the Comparative Psychotherapy Process Scale*. Unpublished manuscript.

Blatt, S., & Zuroff, D. (1992). Interpersonal relatedness and self-definition: Two prototypes for depression. *Clinical Psychology Review, 12,* 527–562.

Borkovec, T. D. (1994). Between-group therapy outcome design: Design and methodology. In L. S. Onken & J. D. Blaine (Eds.), *NIDA Research Monograph No. 137* (pp. 249–289). Rockville, MD: National Institute on Drug Abuse.

Borkovec, T. D., Abel, J. L., & Newman, H. (1995). Effects of psychotherapy on comorbid conditions in generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 63,* 479–483.

Borkovec, T. D., & Castonguay, L. G. (1998). What is the scientific meaning of empirically supported therapy? *Journal of Consulting and Clinical Psychology, 66,* 136–142.

Borkovec, T. D., & Costello, E. (1993). Efficacy of applied relaxation and cognitive–behavioral therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 61,* 611–619.

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruiz, M. (2001). The Pennsylvania practice research network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice, 8,* 155–167.

Bouchard, S., Gauther, J., Laberge, B., French, D., Pelletier, M., & Godbout, C. (1996). Exposure versus cognitive restructuring in the treatment of panic disorder with agoraphobia. *Behaviour Research and Therapy, 34,* 213–224.

Brown, T. A., & Barlow, D. H. (1992). Comorbidity among anxiety disorders: Implications for treatment and *DSM–IV*. *Journal of Consulting and Clinical Psychology, 60,* 835–844.

Brown, T. A., & Barlow, D. H. (1995). Long-term outcome in cognitive–behavioral treatment of panic disorder: Clinical predictors and alternative strategies for assessment. *Journal of Consulting and Clinical Psychology, 63,* 754–765.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the *DSM–IV* anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107,* 179–192.

Bruce, T. J., Spiegel, D. A., & Hegel, M. T. (1999). Cognitive–behavioral therapy helps prevent relapse and recurrence of panic disorder following alprazolam discontinuation: A long-term follow-up of the Peoria and Dartmouth studies. *Journal of Consulting and Clinical Psychology, 67,* 151–156.

Bruch, H. (1973). *Eating disorders: Obesity, anorexia nervosa, and the person within*. New York: Basic Books.

Calhoun, K. S., Moras, K., Pilkonis, P. A., & Rehm, L. (1998). Empirically supported treatments: Implications for training. *Journal of Consulting and Clinical Psychology, 66,* 151–162.

Carroll, K. M., & Nuro, K. F. (2002). One size cannot fit all: A stage model for psychotherapy manual development. *Clinical Psychology: Science and Practice, 9,* 396–406.

Castonguay, L. G., Goldfried, M. R., Wiser, S., Raue, P., & Hayes, A. M. (1996). Predicting the effect of cognitive therapy for depression: A study of unique and common factors. *Journal of Consulting and Clinical Psychology, 64,* 497–504.

Chambless, D., & Hollon, S. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Chambless, D., & Ollendick, T. (2000). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52,* 685–716.

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

Compas, B. E., Haaga, D. A. F., Keefe, F. J., & Leitenbert, H. (1998). Sampling of empirically supported psychosocial treatments from health psychology: Smoking, chronic pain, cancer, and bulimia nervosa. *Journal of Consulting and Clinical Psychology, 66,* 89–112.

Cox, B. J., Swinson, R. P., Morrison, B. L., & Paul, S. (1993). Clomipramine, fluoxetine, and behavior therapy in the treatment of obsessive-compulsive disorder: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 24*(2), 149–153.

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

Eddy, K. T., Dutra, L., & Westen, D. (2004). *A multidimensional meta-analysis of psychotherapy and pharmacotherapy for obsessive–compulsive disorder*. Unpublished manuscript, Emory University, Atlanta, GA.

Egger, M., Juni, P., & Bartlett, C. (1999). Value of flow diagrams in reports of randomized controlled trials. *Journal of the American Medical Association, 285,* 1996–1999.

Elkin, I., Shea, T., Watkins, J., Imber, S., Sotsky, S., Collins, J., et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry, 46,* 971–982.

Eysenck, H. J. (1995). Meta-analysis squared—Does it make sense? *American Psychologist, 50,* 110–111.

Fairburn, C. G. (1997). Interpersonal psychotherapy for bulimia nervosa. In D. M. Garner & P. E. Garfinkel (Eds.), *Handbook of treatment for eating disorders* (2nd ed., pp. 278–294). New York: Guilford Press.

Fairburn, C. G., Jones, R., Peveler, R., Carr, S. J., Solomon, R. A., O'Connor, M., et al. (1991). Three psychological treatments for bulimia nervosa: A comparative trial. *Archives of General Psychiatry, 48,* 463–469.

Fairburn, C. G., Kirk, J., O'Connor, M., & Cooper, P. J. (1986). A comparison of two psychological treatments for bulimia nervosa. *Behaviour Research and Therapy, 24,* 629–643.

Fairburn, C. G., Marcus, M. D., & Wilson, G. T. (1993). Cognitive behaviour therapy for binge eating and bulimia nervosa: A comprehensive treatment manual. In C. G. Fairburn & G. T. Wilson (Eds.), *Binge eating: Nature, assessment, and treatment* (pp. 361–404). New York: Guilford Press.

Feeley, M., DeRubeis, R. J., & Gelfand, L. (1999). The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 67,* 578–582.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48,* 71–79.

Foa, E. B., Dancu, C. V., Hembree, E. A., Jaycox, L. H., Meadows, E. A., & Street, G. P. (1999). A comparison of exposure therapy, stress inoculation training, and their combination for reducing posttraumatic stress disorder in female assault victims. *Journal of Consulting and Clinical Psychology, 67,* 194–200.

Frank, E., Shear, M. K., Rucci, P., Cyranowski, J., Endicott, J., Fagiolini, A., et al. (2000). Influence of panic-agoraphobic spectrum symptoms on treatment response in patients with recurrent major depression. *Archives of General Psychiatry, 157,* 1101–1107.

Frank, E., & Spanier, C. (1995). Interpersonal psychotherapy for depression: Overview, clinical efficacy, and future directions. *Clinical Psychology: Science and Practice, 2,* 349–369.

Franklin, M. E., Abramowitz, J. S., Levitt, J. T., Kozak, M. J., & Foa, E. B. (2000). Effectiveness of exposure and ritual prevention for obsessive–compulsive disorder: Randomized compared with nonrandomized samples. *Journal of Consulting and Clinical Psychology, 68,* 594–602.

Garner, D. M., Rockert, W., Davis, R., Garner, M. V., Olmstead, M. P., & Eagle, M. (1993). A comparison of cognitive–behavioral and supportive–expressive therapy for bulimia nervosa. *American Journal of Psychiatry, 150,* 37–46.

Gemar, M. C., Segal, Z. V., Sagrati, S., & Kennedy, S. J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. *Journal of Abnormal Psychology, 110,* 282–289.

Goldfried, M. R. (2000). Consensus in psychotherapy research and practice: Where have all the findings gone? *Psychotherapy Research, 10,* 1–16.

Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained relationship. *American Psychologist, 51,* 1007–1016.

Goldfried, M. R., & Wolfe, B. E. (1998). Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology, 66,* 143–150.

Hammen, C., Ellicott, A., Gitlin, M., & Jamison, K. R. (1989). Sociotropy/autonomy and vulnerability to specific life events in patients with unipolar depression and bipolar disorders. *Journal of Abnormal Psychology, 98,* 154–160.

Hardy, G. E., Barkham, M., Shapiro, D. A., Stiles, W. B., Rees, A., & Reynolds, S. (1995). Impact of Cluster C personality disorders on outcomes of contrasting brief psychotherapies for depression. *Journal of Consulting and Clinical Psychology, 63,* 997–1004.

Harstein, N. B. (1996). Suicide risk in lesbian, gay, and bisexual youth. In R. P. Cabaj, T. S. Stein, et al. (Eds.), *Textbook of homosexuality and mental health* (pp. 819–837). Washington, DC: American Psychiatric Press.

Hayes, A. M., Castonguay, L. G., & Goldfried, M. R. (1996). Effectiveness of targeting the vulnerability factors of depression in cognitive therapy. *Journal of Consulting and Clinical Psychology, 64,* 623–627.

Hedlund, S., & Rude, S. S. (1995). Evidence of latent depressive schemas in formerly depressed individuals. *Journal of Abnormal Psychology, 104,* 517–525.

Henry, W., Strupp, H., Butler, S., Schacht, T., & Binder, J. (1993). Effects of training in time-limited dynamic psychotherapy: Changes in therapist behavior. *Journal of Consulting and Clinical Psychology, 61,* 434–433.

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

Hollon, S. D., Thase, M. E., & Markowitz, J. C. (2002). Treatment and prevention of depression. *Psychological Science in the Public Interest, 3,* 39–77.

Howard, K. I., Cornille, T. A., Lyons, J. S., Vessey, J. T., Lueger, R. J., & Saunders, S. M. (1996). Patterns of mental health service utilization. *Archives of General Psychiatry, 53,* 696–703.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41,* 159–164.

Howard, K. I., Lueger, R., Maling, M., & Martinovich, Z. (1993). A phase model of psychotherapy: Causal mediation of outcome. *Journal of Consulting and Clinical Psychology, 54,* 106–110.

Humphreys, K., & Weisner, C. (2000). Use of exclusion criteria in selecting research subjects and its effects on the generalizability of alcohol treatment outcome studies. *American Journal of Psychiatry, 157,* 588–594.

Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive–behavior therapy for depression. *Clinical Psychology: Science and Practice, 1,* 138–156.

Ingram, R. E., & Ritter, J. (2000). Vulnerability to depression: Cognitive reactivity and parental bonding in high-risk individuals. *Journal of Abnormal Psychology, 109,* 588–596.

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

Jacobson, N. J., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

Janoff-Bulman, R. (1992). *Shattered assumptions: Towards a new psychology of trauma*. New York: Free Press.

Johnson, C., Tobin, D. L., & Dennis, A. (1991). Differences in treatment outcome between borderline and nonborderline bulimics at one-year follow-up. *International Journal of Eating Disorders, 9,* 617–627.

Jones, E. E. (2000). *Therapeutic action*. Northvale, NJ: Aronson.

Jones, E. E., & Pulos, S. M. (1993). Comparing the process in psychodynamic and cognitive–behavioral therapies. *Journal of Consulting and Clinical Psychology, 61,* 306–316.

Judd, L. (1997). The clinical course of unipolar major depressive disorders. *Archives of General Psychiatry, 54,* 989–991.

Kazdin, A. E. (1997). A model for developing effective treatments: Progression and interplay of theory, research, and practice. *Journal of Clinical Child Psychiatry, 26,* 114–129.

Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology, 66,* 3–6.

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

Kessler, R. C., Nelson, C. B., McGonagle, K. A., Liu, J., et al. (1996). Comorbidity of DSM-III—R major depressive disorder in the general population: Results from the US National Comorbidity Survey. *British Journal of Psychiatry, 168*(Suppl. 30), 17–30.

Kessler, R. C., Stang, P., Wittchen, H. U., Stein, M., & Walters, E. E. (1999). Lifetime comorbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychological Medicine, 29,* 555–567.

Klerman, G. L., & Weissman, M. M. (1993). *New applications of interpersonal psychotherapy*. Washington, DC: American Psychiatric Press.

Kobak, K. A., Greist, J. A., Jefferson, J. W., Katzelnick, D. J., & Henk, H. J. (1998). Behavioral versus pharmacological treatments of obsessive compulsive disorder: A meta-analysis. *Psychopharmacologia, 136,* 205–216.

Kopta, S., Howard, K., Lowry, J., & Beutler, L. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology, 62,* 1009–1016.

Krueger, R. F. (2002). The structure of common mental disorders. *Archives of General Psychiatry, 59,* 570–571.

Kwon, P., & Whisman, M. A. (1998). Sociotropy and autonomy as vulnerabilities to specific life events: Issues in life event categorization. *Cognitive Therapy and Research, 22,* 353–362.

Kyuken, W., Kurzer, N., DeRubeis, R. J., Beck, A. T., & Brown, G. K. (2001). Response to cognitive therapy in depression: The role of maladaptive beliefs and personality disorders. *Journal of Consulting and Clinical Psychology, 69,* 560–566.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143–189). New York: Wiley.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172.

Lewinsohn, P., Rohde, P., Seeley, J. R., & Klein, D. N. (1997). Axis II psychopathology as a function of Axis I disorders in childhood and adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry, 36,* 1752–1759.

Lilienfeld, S. O., Waldman, I., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice, 1,* 71–83.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: Guilford Press.

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

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6,* 95–106.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology: Science and Practice, 4,* 53–65.

McDermut, W., Miller, I. W., & Brown, R. A. (2001). The efficacy of group psychotherapy for depression: A meta-analysis and review of the empirical research. *Clinical Psychology: Science and Practice, 8,* 98–116.

McFall, R. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist, 44*(6), 75–88.

Meehl, P. E. (1954). *Clinical vs. statistical prediction*. Minneapolis: University of Minnesota Press.

Messer, S. (2001). Empirically supported treatments: What's a nonbehaviorist to do? In B. D. Slife, R. N. Williams, & D. Barlow (Eds.), *Critical issues in psychotherapy: Translating new ideas into practice* (pp. 3–19). Thousand Oaks, CA: Sage.

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49,* 377–412.

Mitchell, J. E., Maki, D. D., Adson, D. E., Ruskin, B. S., & Crow, S. J. (1997). The selectivity of inclusion and exclusion criteria in bulimia nervosa treatment studies. *International Journal of Eating Disorders, 22,* 243–252.

Morrison, C., Bradley, R. & Westen, D. (2003). The external validity of efficacy trials for depression and anxiety: A naturalistic study. *Psychology and Psychotherapy: Theory, Research and Practice, 76,* 109–132.

Mueller, T. I., Leon, A. C., Keller, M. B., Solomon, D. A., Endicott, J., Coryell, W., et al. (1999). Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *American Journal of Psychiatry, 156,* 1000–1006.

Nakash-Eisikovits, O., Dierberger, A., & Westen, D. (2002). A multidimensional meta-analysis of pharmacotherapy for bulimia nervosa: Summarizing the range of outcomes in controlled clinical trials. *Harvard Review of Psychiatry, 10,* 193–211.

Nathan, P. E. (1998). Practice guidelines: Not yet ideal. *American Psychologist, 53,* 290–299.

Nathan, P. E., Stuart, S. P., & Dolan, S. L. (2000). Research on psychotherapy efficacy and effectiveness: Between Scylla and Charybdis? *Psychological Bulletin, 126,* 964–981.

Newman, D. L., Moffitt, T., Caspi, A., & Silva, P. A. (1998). Comorbid mental disorders: Implications for treatment and sample selection. *Journal of Abnormal Psychology, 107,* 305–311.

Oldham, J. M., Skodol, A. E., Kellman, H. D., Hyler, S. E., Doidge, N., Rosnick, L., et al. (1995). Comorbidity of Axis I and Axis II disorders. *American Journal of Psychiatry, 152,* 571–578.

Ost, L., & Westling, B. E. (1995). Applied relaxation vs. cognitive behavior therapy in the treatment of panic disorder. *Behaviour Research and Therapy, 33,* 145–158.

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

Persons, J. B., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66,* 126–135.

Persons, J., & Tompkins, M. (1997). Cognitive-behavioral case formulation. In T. D. Ells (Ed.), *Handbook of psychotherapy case formulation* (pp. 314–339). Oakland, CA: Center for Cognitive Therapy.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Thousand Oaks, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

Rosenthal, R., & DiMatteo, M. R. (2000). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

Roth, A., & Fonagy, P. (1996). *What works for whom? A critical review of psychotherapy research*. New York: Guilford Press.

Scheel, K. R. (2000). The empirical basis of dialectical behavior therapy: Summary, critique, and implications. *Clinical Psychology: Science and Practice, 7,* 68–86.

Segal, Z. V., Gemar, M., & Williams, S. (1999). Differential cognitive response to a mood challenge following successful cognitive therapy pharmacotherapy for unipolar depression. *Journal of Abnormal Psychology, 108,* 3–10.

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

Shea, M., Elkin, I., Imber, S., Sotsky, S., Watkins, J., Collins, J., et al. (1992). Course of depressive symptoms over follow-up: Findings from the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry, 49,* 782–787.

Shea, M., Widiger, T., & Klein, M. (1992). Comorbidity of personality disorders and depression: Implications for treatment. *Journal of Clinical and Consulting Psychology, 60,* 857–868.

Shear, M. K., Pilkonis, P. A., Cloitre, M., & Leon, A. C. (1994). Cognitive behavioral treatment compared with nonprescriptive treatment of panic disorder. *Archives of General Psychiatry, 51,* 395–402.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

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

Snyder, D. K., Wills, R. M., & Grady, F. A. (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: A 4-year follow-up study. *Journal of Consulting and Clinical Psychology, 59,* 138–141.

Steiger, H., & Stotland, S. (1996). Prospective study of outcome in bulimics as a function of Axis-II comorbidity: Long-term responses on eating and psychiatric symptoms. *International Journal of Eating Disorders, 20,* 149–161.

Stiles, W. B., & Shapiro, D. A. (1989). Abuse of the drug metaphor in psychotherapy process-outcome research. *Clinical Psychology Review, 9,* 521–543.

Strupp, H. H., Horowitz, L. M., & Lambert, M. J. (Eds.). (1997). *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery*. Washington, DC: American Psychological Association.

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

Sullivan, H. S. (1953). *The interpersonal theory of psychiatry*. New York: Norton.

Tang, T., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive–behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67,* 894–904.

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

Thase, M. E., Simons, A. D., McGeary, J., Cahalane, J., Hughes, C., Harden, T., et al. (1992). Relapse after cognitive behavior therapy of depression: Potential implications for longer courses of treatment. *American Journal of Psychiatry, 149,* 1046–1052.

Thompson-Brenner, H., Glass, S., & Westen, D. (2003). A multidimensional meta-analysis of psychotherapy for bulimia nervosa. *Clinical Psychology: Science and Practice, 10,* 269–287.

Thompson-Brenner, H., & Westen, D. (2004a). *Accumulating evidence for personality subtypes in eating disorders: Differences in comorbidity, adaptive functioning, treatment response, and treatment interventions in a naturalistic sample.* Unpublished manuscript, Boston University, Boston.

Thompson-Brenner, H., & Westen, D. (2004b). *A naturalistic study of psychotherapy for bulimia nervosa: Comorbidity, outcome, and therapeutic interventions in the community*. Unpublished manuscript, Boston University, Boston.

van Blakom, A. J. L. M., van Oppen, P., Vermeulen, A. W. A., van Dyck, R., & Harne, C. M. V. (1994). A meta-analysis on the treatment of obsessive compulsive disorder: A comparison of antidepressants, behavior, and cognitive therapy. *Clinical Psychology Review, 14,* 359–381.

Wampold, B., Mondin, G., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research, 7,* 21–43.

Watson, D., & Clark, L. A. (1992). Affects separable and inseparable: On the hierarchical arrangement of the negative affects. *Journal of Personality and Social Psychology, 62,* 489–505.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1994). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology, 104,* 15–25.

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70,* 299–310.

Weinberger, J. (1995). Common factors aren't so common: The common factors dilemma. *Clinical Psychology: Science and Practice, 2,* 45–69.

Weinberger, J. (2000). *Why can't psychotherapists and psychotherapy researchers get along? Underlying causes of the EST—effectiveness controversy.* Unpublished manuscript, Adelphi University, Garden City, NY.

Weiss, B., Catron, T., & Harris, V. (2000). A 2-year follow-up of the effectiveness of traditional child psychotherapy. *Journal of Consulting and Clinical Psychology, 68,* 1094–1101.

Wenzlaff, R. M., & Eisenberg, A. R. (2001). Mental control after dysphoria: Evidence of a suppressed, depressive bias. *Behavior Therapy, 32,* 27–45.

West, J. C., Zarin, D. A., Peterson, B. D., & Pincus, H. A. (1998). Assessing the feasibility of recruiting a randomly selected sample of psychiatrists to participate in a national practice-based research network. *Social Psychiatry and Psychiatric Epidemiology, 33*(12), 620–623.

Westen, D. (1998a). Case formulation and personality diagnosis: Two processes or one? In J. Barron (Ed.), *Making diagnosis meaningful* (pp. 111–138). Washington, DC: American Psychological Association.

Westen, D. (1998b). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124,* 333–371.

Westen, D. (1999). Psychodynamic theory and technique in relation to research on cognition and emotion: Mutual implications. In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 727–746). New York: Wiley.

Westen, D. (2000). Integrative psychotherapy: Integrating psychodynamic and cognitive-behavioral theory and technique. In C. R. Snyder & R. Ingram (Eds.), *Handbook of psychological change: Psychotherapy processes and practices for the 21st century* (pp. 217–242). New York: Wiley.

Westen, D. (2002). Manualizing manual development. *Clinical Psychology: Science and Practice, 9,* 416–418.

Westen, D., & Arkowitz-Westen, L. (1998). Limitations of Axis II in diagnosing personality pathology in clinical practice. *American Journal of Psychiatry, 155,* 1767–1771.

Westen, D., & Harnden-Fischer, J. (2001). Personality profiles in eating disorders: Rethinking the distinction between Axis I and Axis II. *American Journal of Psychiatry, 165,* 547–562.

Westen, D., Heim, A. K., Morrison, K., Patterson, M., & Campbell, L. (2002). Simplifying diagnosis using a prototype-matching approach: Implications for the next edition of the DSM. In L. E. Beutler & M. L. Malik (Eds.), *Rethinking the DSM: A psychological perspective* (pp. 221–250). Washington, DC: American Psychological Association.

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899.

Westen, D., Moses, M. J., Silk, K. R., Lohr, N. E., Cohen, R., & Segal, H. (1992). Quality of depressive experience in borderline personality disorder and major depression: When depression is not just depression. *Journal of Personality Disorders, 6,* 382–393.

Westen, D., Muderrisoglu, S., Fowler, C., Shedler, J., & Koren, D. (1997). Affect regulation and affective experience: Individual differences, group differences, and measurement using a Q-sort procedure. *Journal of Consulting and Clinical Psychology, 65,* 429–439.

Westen, D., & Shedler, J. (1999a). Revising and assessing Axis II, Part 1: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156,* 258–272.

Westen, D., & Shedler, J. (1999b). Revising and assessing Axis II, Part 2: Toward an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156,* 273–285.

Westen, D., Shedler, J., Durrett, C., Glass, S., & Martens, A. (2003). Personality diagnosis in adolescence: DSM-IV Axis II diagnoses and an empirically derived alternative. *American Journal of Psychiatry, 160,* 952–966.

Westen, D., & Weinberger, J. (2003). *When clinical description becomes statistical prediction.* Unpublished manuscript, Emory University, Atlanta, GA.

Wiers, R. W., van Woerden, N., Smulders, F. T. Y., & De Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology, 111,* 648–658.

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120,* 3–24.

Wilson, G. T. (1998). Manual-based treatment and clinical practice. *Clinical Psychology: Science and Practice, 5,* 363–375.

Wilson, G. T. (1999). Rapid response to cognitive behavior therapy. *Clinical Psychology: Science and Practice, 6,* 289–292.

Wilson, G. T., Fairburn, C. G., & Agras, W. S. (1997). Cognitive-behavioral therapy for bulimia nervosa. In D. M. Garner & P. E. Garfinkel (Eds.), *Handbook of treatment for eating disorders* (2nd ed., pp. 67–93). New York: Guilford Press.

Wixom, J., Ludolph, P., & Westen, D. (1993). Quality of depression in borderline adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 32,* 1172–1177.

Wolf, E. M., & Crowther, J. H. (1992). An evaluation of behavioral and cognitive-behavioral group interventions for the treatment of bulimia nervosa in women. *International Journal of Eating Disorders, 11,* 3–15.

Zimmerman, M., McDermut, W., & Mattia, J. (2000). Frequency of anxiety disorders in psychiatric outpatients with major depressive disorder. *American Journal of Psychiatry, 157,* 1337–1340.

Zinbarg, R. E., & Barlow, D. H. (1996). Structure of anxiety and the anxiety disorders: A hierarchical model. *Journal of Abnormal Psychology, 105,* 181–193.

Received February 3, 2003
Revision received September 8, 2003
Accepted September 15, 2003 ■