CLINICAL PSYCHOLOGY
SCIENCE AND PRACTICE

# Empirically Supported Treatment: Recommendations for a New Model

David F. Tolin, The Institute of Living and Yale University School of Medicine
Dean McKay, Fordham University
Evan M. Forman, Drexel University
E. David Klonsky, University of British Columbia
Brett D. Thombs, Jewish General Hospital and McGill University

**Over the 20 years since the criteria for empirically supported treatments (ESTs) were published, standards for synthesizing evidence have evolved and more systematic approaches to reviewing the findings from intervention trials have emerged. Currently, the APA is planning the development of treatment guidelines, a process that will likely take many years. As an intermediate step, we recommend a revised set of criteria for ESTs that will utilize existing systematic reviews of all of the available literature, and recommendations that address the methodological quality, outcomes, populations, and treatment settings included in the literature.**

***Key words:*** **clinical significance, empirically supported treatment, GRADE tool, systematic reviews.** ***[Clin Psychol Sci Prac, 2015]***

## CONSIDERATIONS IN THE EVALUATION OF EMPIRICALLY SUPPORTED TREATMENTS: ARE EMPIRICALLY SUPPORTED TREATMENTS STILL RELEVANT?

Over two decades ago, the Society of Clinical Psychology (Division 12 of the American Psychological Association [APA]), under the direction of then President David Barlow, first published criteria for what were initially termed "empirically validated psychological treatments" (Task Force on Promotion and Dissemination of Psychological Procedures, 1993) and later termed "empirically supported psychological treatments" (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). The identification of *empirically supported treatments* (ESTs) has had substantial impact in psychology and related mental health disciplines. One immediately tangible effect of the movement to identify ESTs has been the validation of procedures for specific psychological problems, and the dissemination of that information to practitioners, consumers, and other stakeholders on the web (www.psychologicaltreatments.org).

Since (and perhaps in part due to) that early work, the quantity of treatment outcome studies has increased dramatically. A search of PsycINFO for the terms "randomized controlled trial" or "randomised controlled trial" (November 23, 2014) yielded only 20 citations for the year 1995, compared to 123 in 2000, 427 in 2005, and 950 in 2010. Among this increase in randomized controlled trials (RCTs), we see a wide range of therapeutic approaches being evaluated for efficacy. Since the development of the original list of ESTs, most of which were cognitive-behavioral treatments, efficacy trials for psychodynamic therapy (Milrod et al., 2007), transference-focused psychotherapy (Yeomans, Levy, & Caligor, 2013), family-based therapy (Lock et al., 2010), and interpersonal psychotherapy (e.g., Parker, Parker, Brotchie, & Stuart, 2006) have

Address correspondence to David F. Tolin, Ph.D., Anxiety Disorders Center, The Institute of Living, 200 Retreat Avenue, Hartford, CT 06106. E-mail: david.tolin@hhchealth.org.

doi:10.1111/cpsp.12122

appeared, to name a diverse few. The result has been a greater emphasis on empiricism among approaches that previously lacked a history of accumulating research support. This increase in diverse outcome research has shifted the debate among practitioners of different theoretical persuasions from mere assertions of theory to a consideration of empirical evidence.

The quality of available research evidence has also increased substantially over the past 20 years. Detailed and stringent guidelines have now been published regarding the execution and reporting of methodologically sound treatment outcome studies (Moher, Schulz, & Altman, 2001), and leading psychology journals such as the *Journal of Consulting and Clinical Psychology* require that manuscripts adhere to such guidelines (retrieved November 23, 2014, from http://www.apa.org/pubs/journals/ccp/index.aspx). These changes have led to a greater emphasis on study quality. Given the emphasis on establishing procedures as empirically supported, guidebooks have been published that carefully document how to design sound therapy research investigations (e.g., Areán & Kraemer, 2013). Recently, a review of trials of psychodynamic and cognitive-behavioral therapies, using a rating scale of various aspects of methodological quality and study reporting (Kocsis et al., 2010), concluded that study quality and reporting have been significantly improving over the past four decades (Thoma et al., 2012).

The EST movement has led to changes in how students are trained in clinical practice. Although training programs still have a wide degree of latitude, EST lists help guide curricula and inform syllabi. Most prominently, the APA Commission on Accreditation's Guidelines and Procedures (2013) encourages programs to train students in assessment and treatment procedures based on empirically supported methods, encourages placement in training settings that employ empirically supported approaches, and encourages internship training sites to include methods of demonstrating that interns possess intermediate to expert-level knowledge in ESTs.

Finally, the development of lists of ESTs has resulted in greater protections for the public. By developing a list of established and empirically supported interventions, treatment-seeking individuals are now better able to learn about and seek out information on well-validated treatments for specific disorders and problem areas. This increased consumer education encourages clinicians who might otherwise not have practiced in an empirically supported manner to acquire the necessary skills to begin offering scientifically based treatments. Perhaps the most ambitious illustration of the impact of the movement toward scientifically tested treatments on clinical practice are the National Institute of Clinical Excellence standards in the United Kingdom (NICE; Baker & Kleijnen, 2000), established to ensure that clinicians practice specific and accepted empirically based interventions for different psychological conditions (see http://guidance.nice.org.uk/Topic/MentalHealthBehavioural). Similarly, the Veterans Health Administration, which serves nearly 6 million veterans in the United States, has undertaken a complete overhaul of its mental health practices and is implementing a systemwide dissemination of empirically based treatments for posttraumatic stress disorder, depression, and serious mental illness (Ruzek, Karlin, & Zeiss, 2012).

Importantly, the early work on ESTs was an important catalyst for the APA's relatively recent emphasis on *evidence-based practice* (EBP). EBP is a broad template of activities that include assessment, case formulation, relationship factors, and treatment decisions that will assist the clinician to work with a patient to achieve the best possible outcome. In 2006, a Presidential Task Force of the American Psychological Association (APA Presidential Task Force on Evidence-Based Practice, 2006) adapted the Institute of Medicine's (2001) definition of evidence-based medicine, defining EBP as practice that integrates three sources of information: patient characteristics, clinical expertise, and the best available research evidence.

It might well be asked, given the broad movement in psychology and other health disciplines toward EBP, whether identification of ESTs is still a necessary task. We argue that it is, perhaps now more than ever. The "three-legged stool" of research evidence, patient characteristics, and clinician expertise leaves room for debate about the relative importance of each; however, we suggest that EBP is best approached as starting from the perspective of ESTs—that is, for any given problem, what treatment or treatments have proven efficacious? This scientific information is then interpreted and potentially adapted based on clinician expertise and

patient characteristics. Thus, where treatment selection is concerned, EBP might be thought of as an approach to ESTs, filtering that scientific information through the clinician's and patient's "lenses" (Djulbegovic & Guyatt, 2014; Tolin, 2014).

As a brief example, a clinician may want to select a treatment approach for an impoverished African American man with a presenting complaint of depression, as well as a significant drinking problem. Most likely, no published list of ESTs will match this situation perfectly. However, using the "filter system" of EBP may lead to a helpful solution. Examination of the available ESTs for depression alerts the clinician to the fact that behavioral activation has strong empirical support in the treatment of depression (Lejuez, Hopko, & Hopko, 2001; Lewinsohn, Biglan, & Zeiss, 1976; Martell, Addis, & Jacobson, 2001). The contributing research, however, did not address the present patient's characteristics such as socioeconomic status, race, and the presence of a co-occurring substance use disorder. The clinician would therefore rely on expertise and additional research to understand how an EST such as behavioral activation might be adapted in a manner that successfully addresses these issues. These modifications might include specific cultural adaptations (Benish, Quintana, & Wampold, 2011; Griner & Smith, 2006; van Loon, van Schaik, Dekker, & Beekman, 2013) or the addition (either concurrently or sequentially) of an EST for drinking problems such as behavioral couples therapy (O'Farrell, Cutter, Choquette, Floyd, & Bayog, 1992) or contingency management (Petry, Martin, Cooney, & Kranzler, 2000). The treatment(s) must also be delivered competently in a way that successfully engages the patient, thus requiring a high level of clinical competency and cross-cultural awareness. The process starts, however, with identification of a specific EST. To make informed decisions, patients and clinicians must be aware of the available scientific evidence, and the degree of confidence that can be placed in that evidence.

#### WHY DOES THE LIST NEED TO BE REVISED?

Many authors, including those broadly in agreement with the EST concept in theory, have raised significant concerns about how ESTs are currently defined. Many of the critiques of the EST movement point to problems in how research evidence is synthesized and evaluated. The original Division 12 report on ESTs delineated specific criteria (see Table 1) by which a treatment would be regarded as "probably efficacious" or "well established" (Chambless & Hollon, 1998; Chambless & Ollendick, 2001; Task Force on Promotion and Dissemination of Psychological Procedures, 1993), and these criteria are still being used today. In brief, to meet the highest standard of "well established," a treatment must be supported by (a) at least two independently conducted, well-designed studies or (b) a large series of well-designed and carefully controlled single-case design experiments. To meet the standard of "probably efficacious," a treatment must be supported by at least one well-designed study or a small series of single-case design experiments.

Given the proliferation of clinical research over the past two decades, the improved quality of clinical research, and the adoption of more sophisticated meth-

**Table 1.** Current definitions of "well established" and "probably efficacious" treatments (adapted from Chambless et al., 1998)

| | |
|---|---|
| **Well Established** | |
| I | At least two good between-group design experiments demonstrating efficacy in one or more of the following ways: |
| | A Superior (based on statistical significance alone) to pill or psychological placebo or to another treatment. |
| | B Equivalent to an already established treatment in experiments with adequate statistical power, considered to be approximately 30 per group. |
| | OR |
| II | A large series of single-case design experiments ($n > 9$) demonstrating efficacy. These experiments must have: |
| | A Used good experimental designs and |
| | B Compared the intervention to another treatment as in IA. |
| | Further criteria for both I and II: |
| III | Experiments must be conducted with treatment manuals. |
| IV | Characteristics of the client samples must be clearly specified. |
| V | Effects must have been demonstrated by at least two different investigators or investigating teams. |
| Probably Efficacious | |
| I | Two experiments showing the treatment is superior (based on statistical significance alone) to a waiting-list control group. |
| | OR |
| II | One or more experiments meeting all criteria for well-established treatments except V (demonstration by independent investigator teams). |
| | OR |
| III | A small series of single-case design experiments ($n > 3$) meeting well-established treatment criteria II, III, and IV. |

ods for research synthesis and evaluation, we concur with many critics who have suggested that the current criteria are outdated (see Table 2). The evaluation based on two studies sets an unacceptably low bar for efficacy, may not account for mixed findings, and risks creating a misrepresentative and highly selective impression of efficacy (Borkovec & Castonguay, 1998; Henry, 1998; Herbert, 2003). For example, if two studies find evidence that a given treatment is efficacious, five studies find the treatment is no better than placebo, and 10 studies find that the treatment is worse than placebo, the current criteria for a designation of a "well-established" EST would be satisfied. This is not a hypothetical scenario, and many bodies of treatment evidence include some studies with statistically significant results favoring a treatment and other studies that report null or even negative findings. This is a problem that occurs across areas of research, and its influence has been well documented in the evidence on pharmaceutical products, where a clear bias for trials favorable to a sponsored product has been demonstrated (Lexchin, Bero, Djulbegovic, & Clark, 2003; Lundh, Sismondo, Lexchin, Busuioc, & Bero, 2012). Registration of clinical trials (e.g., at www.clinicaltrials.gov) is increasingly emphasized to address this problem, not only for pharmaceutical studies but also for studies of psychological interventions, although poor adherence to registration policies and poor quality of trial registrations have been problematic (Riehm, Azar, & Thombs, 2015).

The exclusive focus on symptom reduction risks ignoring other potentially important clinical outcomes, such as functional impairment (Dobson & Beshai, 2013), despite the fact that functional concerns are a leading reason for individuals to seek treatment (Hunt & McKenna, 1993). Although symptom reduction and improvements in functioning are significantly correlated, there can be a mismatch after treatment (see Vatne & Bjorkly, 2008, for review). Thus, it is possible that a treatment is highly effective at reducing specific target symptoms, and yet the patient fails to achieve desired clinical outcomes such as improved social or occupational functioning. Therefore, a number of scholars have cautioned against the overreliance of symptom-based evaluations of efficacy and have instead urged consideration of wellness, quality of life, well-being, and functionality (Cowen, 1991; Hayes, 2004; Seligman, 1995). We propose that symptom reduction no longer be considered the *sine qua non* of treatment outcome. Symptom reduction is important in deter-

**Table 2.** Common critiques of the EST movement and suggested changes

| Area | Critiques | Proposed Changes |
|---|---|---|
| Concerns about the strength of treatment | • Inadequate attention to null or negative findings<br>• Reliance on statistical, rather than clinical, significance<br>• Inadequate attention to long-term outcomes<br>• Potentially significant variability in study quality | • Emphasize systematic reviews rather than individual studies<br>• Separate strength of effect from strength of evidence<br>• Grade quality of studies<br>• Consider clinical significance in addition to statistical significance<br>• Consider long-term efficacy in addition to short-term efficacy |
| Concerns about selecting among multiple treatment options | • Within a given EST category, there is little basis for choosing one over another<br>• Lack of clarity about whether empirical support translates to a recommendation | • Present quantitative information about treatment strength<br>• Make specific recommendations based on clinical outcomes and the quality of the available research |
| Concerns about the relevance of findings | • Inadequate attention to functional outcomes<br>• Inadequate attention to effectiveness in nonresearch settings or with diverse populations | • Include functional or other health-related outcomes as well as symptom outcomes<br>• Address generalization of research findings to nonresearch settings and diverse populations |
| Concern about unclear active treatment ingredients and the proliferation of manuals for specific diagnoses | • Listing of packaged treatments rather than empirically supported principles of change<br>• Emphasis on specific psychiatric diagnoses | • Evaluate and encourage dismantling research to identify empirically supported principles of change<br>• De-emphasize diagnoses and emphasize syndromes/mechanisms of psychopathology |

mining the efficacy of a treatment, but the value of symptom reduction is greatly diminished if functional improvement is not also demonstrated. Functional outcomes address domains of psychosocial functioning, which may include work attendance or performance, school attendance or performance, social engagement, or family functioning. Several measures of such functional outcomes have been published, including the Sheehan Disability Scale (Sheehan, 2008), Leibowitz Self-rating Disability Scale (Schneier et al., 1994), Work and Social Adjustment Scale (Mundt, Marks, Shear, & Greist, 2002), Range of Impaired Functioning Tool (Leon et al., 1999), and the functional subscales of the Outcomes Questionnaire (Lambert et al., 1996), in addition to a wide array of performance-based functional tests from disciplines such as industrial/organizational psychology. The value of specific measures in the evidence review will depend on their psychometric properties and direct relevance to the clinical problem being treated.

Quality of life (QOL) is a less well-defined construct (Gill & Feinstein, 1994), which is problematic for many trials of psychological treatment, given its apparently strong overlap with depression (Keltner et al., 2012). We therefore concur with Muldoon, Barger, Flory, and Manuck (1998) that objective functioning and subjective appraisals of well-being be considered separately. Nevertheless, there is increasing interest in QOL as an outcome measure in trials of psychological treatments, particularly in the United Kingdom (e.g., Layard & Clark, 2014), and its inclusion in treatment guidelines should be considered carefully going forward.

There is, at present, no clear way to establish whether a treatment has proven effective with diverse populations or in more clinically representative settings (Beutler, 1998; Goldfried & Wolfe, 1996, 1998; Gonzales & Chambers, 2002; Norcross, 1999; Seligman, 1996). Concerns about the transportability of treatment include the fact that patients seen in routine clinical practice might be more complex or heterogeneous than those in efficacy-oriented RCTs, that willingness to be randomized to treatments may be a confounding factor that diminishes sample representativeness, and that the therapists used in efficacy RCTs are more highly trained, specialized, monitored, or structured than are those working in routine clinical settings. The issue of treatment generalizability is complex. Patients seen in clinical settings do not necessarily appear more complex or severe than those seen in clinical trials; in one study of clinical outpatients deemed ineligible for depression research trials, the most common reasons for exclusion were partial remission of symptoms at intake and insufficient severity or duration of symptoms. Importantly, of those meeting criteria for major depression, none were excluded due to Axis I or Axis II comorbidity (Stirman, Derubeis, Crits-Christoph, & Rothman, 2005).

Evidence for differential efficacy of treatments administered in research versus clinical settings is mixed. In some cases, randomized and nonrandomized patients receiving similar treatments appear to do equally well (Franklin, Abramowitz, Kozak, Levitt, & Foa, 2000), whereas in other cases, treatments administered in a research setting yield outcomes superior to the same treatments administered in a clinical setting (Gibbons, Stirman, DeRubeis, Newman, & Beck, 2013; Kushner, Quilty, McBride, & Bagby, 2009). The reasons for a possibly stronger response in research trials are unclear, but could include factors such as therapist training and fidelity monitoring, setting time limits for treatment, and providing feedback to clinicians and patients on treatment progress.

Many have called for a greater emphasis on *effectiveness research,* which focuses primarily on the generalizability of the treatment to more clinically representative situations. We therefore suggest that the evaluation of ESTs attend not only to the efficacy of a treatment in research settings, but also in terms of that treatment's *effectiveness in nonresearch settings*. Criteria that could be considered include more diagnostically complex patients, effectiveness with nonrandomized patients, effectiveness when used by nonacademic practitioners, and utility in open-ended, flexible practice.

The internal validity and degree of research bias in clinical trials are not adequately addressed, potentially making the results prone to false-positive results (Luborsky et al., 1999; Wachtel, 2010). Internal validity relates to the degree to which a given trial likely answers the research question being evaluated correctly or free from bias. Bias is systematic error that can lead to underestimation or overestimation of true treatment

effects (Higgins & Green, 2008). It is not usually possible to know with precision the degree to which design flaws may have influenced results in a given treatment trial, but elements of trial design have been shown to be related to bias. In RCTs, generally, design weaknesses related to allocation concealment, blinding, and randomization methods may be expected to influence effect estimates, particularly when outcomes are subjective (Savovic et al., 2012), which is the case in most trials of psychological treatments (Wood et al., 2008). An additional example is the researcher allegiance effect (Gaffan, Tsaousis, & Kemp-Wheeler, 1995; Luborsky et al., 1999). The presence of researcher allegiance does not necessarily imply bias (Hollon, 1999; Leykin & DeRubeis, 2009); however, it is a risk factor that has been shown empirically to be associated with some probability of bias. Financial conflict of interest, a demonstrated source of publication bias in pharmaceutical studies (Friedman & Richter, 2004; Lexchin et al., 2003; Perlis et al., 2005), may also be considered in rating risk of bias (Bero, 2013; Roseman et al., 2011, 2012), although conflict of interest may be harder to identify and quantify in studies of psychological treatments.

#### DOES THE WORLD NEED ANOTHER LIST OF ESTS?

Even though, as we argue, it remains of vital importance to identify ESTs, one might ask whether another list would be beneficial to the field. We suggest that a well-designed list could be of great import, filling noticeable gaps in the available knowledge. Three alternative systems with which readers are likely to be familiar include the NICE standards in the United Kingdom (Baker & Kleijnen, 2000), the Practice Guidelines published by the American Psychiatric Association (e.g., 2009, 2010), and the Veterans Administration/Department of Defense Clinical Practice Guidelines (e.g., Veterans Health Administration, 2004, 2009). These systems are available immediately and have the advantage of addressing both psychological and pharmacological treatments. However, the breadth of these systems is also a limitation for psychologists. As broad guidelines, they lack the depth of information that clinical psychologists or other psychotherapy practitioners would need to make informed treatment decisions. For example, in the NICE guidelines for panic disorder (National Institute for Clinical Excellence, 2011), clinicians are advised to use cognitive-behavioral therapy (CBT). We would not disagree with this recommendation; however, NICE provides little means for understanding what kind of CBT is most helpful or the strength of various interventions. Thus, although existing guidelines are comprehensive and immediately available, we argue that there is room for an alternative source of information for consumers of research on psychological treatments. As the Society of Clinical Psychology has been at the forefront of identifying and disseminating ESTs for the past two decades and is one of the most prominent organizations dedicated to psychological ESTs in particular, it is logical for this group to take the lead in this next phase of treatment evaluation.

In recent years, the APA Advisory Steering Committee for the Development of Clinical Practice Guidelines was formed to provide research-based recommendations for the psychological treatment of particular disorders (Hollon et al., 2014). When in place, guideline development panels, under the direction of the Steering Committee, will oversee the development of clinical practice guidelines. A number of steps that have been proposed by the Advisory Steering Committee to generate patient-focused, scientifically based, clinically useful guidance point the way toward steps that should be taken for a much-needed update of EST standards. Two of them, in particular, should be central to modernizing EST standards: (a) the evaluation of all existing evidence via high-quality systematic reviews, which include (i) evaluation of relevance to clinical practice, including treatment fidelity; (ii) an assessment or risk of bias; and (iii) other considerations, including evaluation of multiple clinical outcomes, including functional, as well as symptom, outcomes; and (b) a committee-based appraisal of the evidence, using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system (Atkins et al., 2004; Guyatt et al., 2006, 2008) to assess the quality of relevant evidence and degree to which benefits are established in excess of potential harms.

The proposed process by the APA Advisory Steering Committee for the Development of Clinical Practice Guidelines represents an important step forward in initiating a disorder-based guideline development process

for psychological treatments. This process, which parallels that used by the Institute of Medicine (2011a, 2011b), is expected to result in a transparent system of treatment recommendations for practitioners and consumers. However, it is an expensive and extremely time-consuming process, and it is unlikely that the Task Force will develop recommendations for a wide range of clinical problems in the immediate future. Indeed, the APA initiated a process for producing guidelines in 2010 and announced panels to develop guidelines for the treatment of obesity and posttraumatic stress disorder in 2012 and for depressive disorders in 2013, but has not yet generated any finished guidelines. Thus, there is an immediate need for dissemination of up-to-date, evidence-based guidance that can not only complement the work of the APA Task Force, but also provide practitioners with clear information about the strength of ESTs and the degree of confidence that can be derived from the available evidence.

#### TO WHAT EXTENT SHOULD WE FOCUS ON ESTABLISHED TREATMENTS, VERSUS PRINCIPLES OF CHANGE?

Over time, the field would likely benefit from a shift away from "named" or "packaged" treatments. The current EST list includes more recent multicomponent treatments that contain many different interventions within one treatment "package." CBT for fibromyalgia, as one example of a treatment currently identified as well established, is described as including education, relaxation, graded behavioral activation, pleasant activity scheduling, sleep hygiene, stress management, goal setting, structured problem solving, reframing, and communication skills (Bernardy, Fuber, Kollner, & Hauser, 2010). While the assessment of such treatment packages is a necessary step in identifying what works, such research does not allow for a determination of which aspects of the treatment are responsible for change (Borkovec & Castonguay, 1998; Gonzales & Chambers, 2002; Henry, 1998). That is, within a given treatment package, there is no way to determine which components of that treatment are therapeutically active or inert. As a result, practitioners are often unable to make informed decisions about which treatments to use (Herbert, 2003; Rosen & Davison, 2003; Westen, Novotny, & Thompson-Brenner, 2004), and many treatments may be weakened by ineffective components and/or work for reasons other than those proposed by the treatment developers (Lohr, Tolin, & Lilienfeld, 1998).

An emphasis on identifying the active ingredients of change need not exclude factors associated with the therapeutic relationship. Indeed, many have suggested that the therapeutic relationship accounts for greater variance in clinical outcomes than do those aspects of the therapy that are described as "techniques" (Blatt & Zuroff, 2005; Henry, 1998; Lambert & Barley, 2001; Norcross, 1999). Relationship-oriented therapist behaviors are themselves subject to empirical scrutiny.

A pressing question, however, is whether there is enough research to date to make meaningful recommendations to practitioners, consumers, and other stakeholders based solely on empirically supported principles of change. We suggest that the field is approaching that target, but has not yet arrived. Certainly, there is much work being done in this area (e.g., Castonguay & Beutler, 2006); however, in our opinion, the field has not yet amassed a body of evidence that would adequately address the multiple concerns of patients seen in clinical settings. As just one example, a recent review concluded that the mechanisms of prolonged exposure (PE) for posttraumatic stress disorder (PTSD), which is a well-studied and fairly straightforward treatment, remain unclear (Zalta, 2015). It would be difficult, therefore, to evaluate only mechanism-based processes at this time, although we believe that such research should be emphasized going forward.

#### HOW SHOULD WE HANDLE TREATMENTS WITH CONFLICTING EVIDENCE?

As noted previously, a primary limitation of the existing criteria is that it allows reviewers to select two positive studies, while potentially ignoring studies with null or even negative outcomes. In our view, the only defensible strategy is a systematic (quantitative) review that takes into account all of the available research evidence, rather than selecting a limited number of positive studies. This is the approach that has been proposed by the APA Advisory Steering Committee for the Development of Clinical Practice Guidelines (Hollon et al., 2014). Twenty years ago, there were not enough controlled research trials, in many cases, for such a process to be feasible. Today, however, the field has seen a marked increase in published research, making larger-scale reviews possible.

### HOW MUCH WEIGHT SHOULD WE AFFORD IMMEDIATE VERSUS LONG-TERM EFFICACY OF TREATMENTS?

Both short-term and long-term outcomes of psychological treatment are important. Short-term outcomes are frequently the strongest and give the best estimate of the immediate efficacy of the treatment. However, it is quite possible that a given treatment is effective in the short term but not at a time point well after treatment discontinuation (i.e., participants exhibited signs of relapse). In some cases, this might reflect a basic weakness of the treatment, suggesting that its effects are not durable. In some other cases, it could be argued that the treatment is only effective so long as one remains in the treatment; so long as the treatment can be feasibly delivered on a maintenance basis, this is not necessarily a fatal flaw. For example, while many have pointed out that gold standard cognitive-behavioral treatments for obesity have short-term effects (most people eventually gain back their lost weight), others point out that a continuous care model is both feasible and better suited to the problem of overeating (Perri, Sears, & Clark, 1993). In still other cases, a lack of long-term efficacy may reflect the presence of competing issues (e.g., chronic psychosocial stressors) that complicate the long-term prognosis despite an apparently successful treatment, suggesting the need for supplemental intervention. Alternatively, it is possible that a treatment might show only modest clinical effects at immediate posttreatment, but outcomes become stronger after treatment discontinuation (sleeper effects) due to memory consolidation effects, skill practice effects, naturalistic reinforcement, or other factors. Consumers, practitioners, and policymakers should be able to evaluate both short- and long-term treatment effects as part of a systematic review.

### HOW SHOULD WE ADDRESS EFFICACY VERSUS EFFECTIVENESS?

Many authors have questioned whether the results of RCTs conducted in clinical research settings will translate to more clinically representative settings such as private practice, community mental health centers, and hospitals (Beutler, 1998; Goldfried & Wolfe, 1996, 1998; Gonzales & Chambers, 2002; Norcross, 1999; Seligman, 1996). Concerns about the transportability of treatment include the fact that patients seen in routine clinical practice might be more complex or heterogeneous than those in efficacy-oriented RCTs, that willingness to be randomized to treatments may be a confounding factor that diminishes sample representativeness, and that the therapists used in efficacy RCTs are more highly trained, specialized, monitored, or structured than are those working in routine clinical settings. Many have therefore called for a greater emphasis on *effectiveness research,* which focuses primarily on the generalizability of the treatment to more clinically representative situations.

We suggest that treatments should be evaluated from both an efficacy and effectiveness perspective. Specifically, it is important to identify treatments that are not only efficacious in research-based settings but have also demonstrated evidence of effectiveness in more typical clinical settings. Criteria that could be considered include more diagnostically complex patients, effectiveness with nonrandomized patients, effectiveness when used by nonacademic practitioners, and utility in open-ended, flexible practice.

### HOW SHOULD TREATMENT COSTS AND BENEFITS BE WEIGHED?

There is, unfortunately, no quantitative "gold standard" for determining whether or not a treatment is cost-effective. Nevertheless, cost-effectiveness considerations must be taken into account. Two treatments may show similar clinical effects, but if one treatment is clearly more costly to consumers, third-party payers, or society (e.g., the treatment requires a very large number of sessions, long duration, or hospitalization), then this should be taken into consideration. It would be prohibitive to conduct a full cost-benefit analysis of every psychological treatment, but a reasonable panel of reviewers should be able to upgrade or downgrade a treatment based on obvious strengths or weaknesses in cost or patient burden.

### WHAT STRENGTH OF EFFECT SHOULD BE CONSIDERED "GOOD"?

Various attempts to define cutoffs of "good response" have been proposed. Cohen (1988), for example, suggested that effect sizes ($d$) of 0.2, 0.5, and 0.8 be considered small, moderate, and large effects, respectively. Others have proposed varying definitions of treatment

response and remission (Andreasen et al., 2005; Doyle & Pollack, 2003; Frank et al., 1991; McIntyre, Fallu, & Konarski, 2006; Simpson, Huppert, Petkova, Foa, & Liebowitz, 2006), usually operationalized as a cutoff score on a standardized measure. Similarly, many have called for the use of reliable change (demonstration that reduction on a measure is greater than would be expected to occur at random) and clinically significant change (variously described as posttreatment scores no longer in the pathological range, posttreatment scores in the normal range, or posttreatment scores that are closer to the normal range than the pathological range) as outcome criteria (Jacobson, Follette, & Revenstorf, 1984; Lambert & Bailey, 2012). Some have used the criterion of good end-state functioning (e.g., Feeny, Zoellner, & Foa, 2002), reflecting scores in the normal range on a variety of different measures, not solely measures of the disorder being treated. From a population-based perspective, some have suggested the use of statistics such as number needed to treat (NNT), reflecting the number of patients needed to treat to observe one improvement.

These methods (many of which overlap considerably) all have their individual strengths and weaknesses. Ultimately, however, there is no clear consensus in the field to tell us how strong of an effect must be observed before we pronounce a treatment to be efficacious. In our view, the degree to which treatment effects are considered clinically meaningful is highly dependent on contextual factors such as the disorder being treated and the goals of treatment. In a case of (for example) mild depression treated on an outpatient basis, full remission and good end-state functioning might be considered appropriate targets, and one might be skeptical of a treatment that fails to achieve those goals. On the other hand, for chronically psychotic patients seen in residential or day treatment, improvements in psychosocial functioning, regardless of the presence of psychotic symptoms, might be considered an appropriate goal, and full remission would not be reasonably expected. Brief inpatient interventions for suicidality may have as their aim the reduction of suicidal ideation and behavior, but not necessarily the remission of depression. Interventions with medical populations might aim to improve compliance with treatment regimens, but not necessarily address the symptoms of a psychological disorder directly. Thus, a "clinically meaningful" treatment result for one group and purpose might not be suitable for another group and purpose. The conclusion that a treatment is "efficacious" therefore is a subjective process that requires human decision-making.

### A PROPOSED SYSTEM OF TREATMENT EVALUATION FOR THE SOCIETY OF CLINICAL PSYCHOLOGY

As described previously, the proposed process of systematic evaluation by the APA Advisory Steering Committee for the Development of Clinical Practice Guidelines represents a clear move in the right direction. However, we argue that there remains a need, both due to the time-consuming nature of the APA process and due to the specific needs of clinical psychologists and consumers for evidence-based decision-making, for the Society of Clinical Psychology to create a new system by which scientific evidence of treatment efficacy can be evaluated and disseminated in a clear, transparent, and cost-effective manner that prioritizes the empirical basis of psychological treatments. The system we propose here is consistent with the methods that will be used by the APA Task Force (Hollon et al., 2014), but requires less time and therefore can provide more rapid dissemination of findings and recommendations. The most time-consuming aspect of the APA Task Force will be the systematic review of research findings. That process could be greatly sped up by using existing, published systematic reviews of the literature. Since the original EST criteria were developed, systematic reviews and meta-analyses are now available for most interventions, and for many of these, the Task Force will be able to use high-quality reviews that have already been published in order to expedite its work.

We note as well that although many of the existing clinical trials and systematic reviews are based on participants selected according to diagnostic criteria (e.g., those listed in the *Diagnostic and Statistical Manual of Mental Disorders* [5th ed.; *DSM-5*; American Psychiatric Association, 2013]), there is no requirement that they do so. Indeed, the reliability and validity of the *DSM* and the medical taxonomy implied therein have been critiqued as a basis for psychotherapy research (Fensterheim & Raw, 1996; Henry, 1998). Over the coming

years, we encourage clinical psychology researchers to focus on distinct, empirically derived *syndromes* of psychopathology (which can range from mild to severe), rather than on categorical diagnoses. Such a shift would comport well with the Research Domain Criteria (RDoC) project currently underway within the National Institute of Mental Health (Insel et al., 2010), although the specific RDoC dimensions may or may not be those chosen as targets for psychotherapy research. That shift would also likely decrease the EST movement's reliance on a large number of treatment manuals, a process to which many authors, even those supportive of the broad EST movement, object (e.g., Fonagy, 1999; Goldfried & Eubanks-Carter, 2004; Levant, 2004; Norcross, 1999; Wachtel, 2010; Westen et al., 2004). Understanding the core dimensions of pathology and the treatments that target this dimension would create a much simpler, more intuitive, and more practitioner-friendly system.

The proposed system takes into account the recommendations of APA work groups (American Psychological Association, 1995, 2002), suggesting that treatment guidelines should (a) be based on broad and careful consideration of the relevant empirical literature, (b) take into consideration the level of methodological rigor and clinical sophistication of the research, (c) take comparison conditions into account, (d) consider available evidence regarding patient-treatment matching, (e) specify the outcomes the intervention is intended to produce, (f) identify known patient variables that influence the utility of the intervention, (g) take the setting of the treatment into account, (h) note possible adverse effects, and (i) take treatment cost into account.

### STEP 1: EXAMINATION OF SYSTEMATIC RESEARCH REVIEWS

We propose that candidate treatments be evaluated on the basis of existing (published or unpublished) quantitative reviews by a Task Force operating under the direction of the Committee on Science and Practice, the group that has overseen the identification of ESTs over the past two decades. The process of member selection should be transparent, with an open nominating process, public listing of member names, and organizational measures to ensure diversity of member backgrounds. Following the recommendations of the APA work groups (American Psychological Association, 1995, 2002), review panels should (a) be composed of individuals with a broad range of documented expertise, (b) disclose actual and potential conflict of interest, (c) maintain a climate of openness and free exchange of views, and (d) have clearly defined processes and methods.

When an individual nominates a treatment for evaluation, the nominator may provide existing reviews or may create a new review for this purpose, although all reviews will be evaluated carefully for thoroughness and risk of bias (see below). Published or unpublished systematic reviews that are not deemed to meet rigorous quality standards will not be considered for EST designation. Recently conducted reviews (i.e., within the past 2 years) will be required unless the evidence in an older review is robust and a strong case can be made that it is unlikely that there are recent developments that would influence the evaluation of the body of evidence for or against a treatment. The effectiveness of a given treatment can be evaluated (a) based on comparisons to known and quantifiable inactive control conditions including (i) wait list, (ii) pill placebo, and (iii) psychological placebo or (b) by comparing alternative psychological treatments.

#### Evaluating the Quality of Systematic Reviews

There are a number of ways to determine whether a systematic review has been conducted with sufficient transparency and rigor to provide confidence that its results are comprehensive and reflect the best possible evidence. The Cochrane Handbook (Higgins & Green, 2008) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Liberati et al., 2009) are well-respected systems for evaluation; the Task Force will use, at least in its initial efforts, the AMSTAR checklist (Shea, Bouter, et al., 2007; Shea, Grimshaw, et al., 2007; Shea et al., 2009) as described above and shown in the online supplement. The AMSTAR checklist is not scored; therefore, there is no cutoff at which a review is considered reliable; rather, the items on the checklist will be used to inform the group's subjective decision of when a systematic review is of sufficient quality and reported sufficiently well to be used by the Division 12 Task Force (Table 3).

**Table 3.** Summary of the proposed Division 12 procedure for evaluating empirically supported treatments

| Step | Process | Details |
|---|---|---|
| Step 1 | Systematic review | • Treatment is nominated<br>• Existing systematic review is evaluated according to:<br>  ○ PICOTS (population, intervention, comparison, outcomes, timeline, setting)<br>  ○ Risk of bias (low, unclear, high) |
| Step 2 | Committee-based evidence review | • GRADE (Grading of Recommendations Assessment, Development, and Evaluation) recommendation by committee: very strong, strong, weak |

In some cases, a systematic review may combine trials in which the treatments differed from each other in one or more ways, such as the manner in which an intervention was applied, the characteristics of the provider, or the length of treatment or follow-up. In some cases, elements of treatment might be added to or subtracted from certain studies. Such modifications across studies create a dilemma for the reviewers, who must determine whether there is sufficient similarity among the studies to consider them all to be testing the same essential treatment. Some degree of clinical heterogeneity must be anticipated and allowed, or else there would be very few meaningful groupings of studies for review. However, the degree to which there is clinical heterogeneity that negatively impacts the interpretability of a single quantitative result must be carefully considered before a meta-analysis is considered by the Task Force (Ioannidis, 2008). A standard part of the review should include agreement among reviewers that all of the selected studies are similar enough that they can be considered to reflect a single treatment.

The use of systematic reviews does not preclude the inclusion of single-case designs, as these designs, when using appropriate experimental control, can establish causality (Horner et al., 2005) in a manner comparable to RCTs, although the smaller number of subjects may limit the generalizability of findings. Methods have been developed for calculating effect sizes of such studies and conducting Bayesian and multilevel modeling (see Shadish, 2014, for a summary). Assessment of the quality of single-subject designs could employ published quality indicators (Horner et al., 2005; Smith et al., 2007), in a manner that parallels the procedures used to evaluate RCTs. Consistent with current approaches to evidence synthesis, however, we do not recommend that evidence from only single-subject designs be used as the basis of recommendations, which should rely largely on synthesis of data from larger clinical trials.

#### Evaluation of Relevance to Clinical Practice

An important component for ensuring the external validity of systematic reviews is the definition of structured review questions. The mnemonic PICOTS refers to the explicit delineation of trials that are eligible for consideration in the systematic review based on the population that received the treatment (P); the intervention delivered (I); the comparison, such as another active treatment or an inactive control (C); outcomes that are assessed (O); the timeline (e.g., 12 weeks, 6 months, or long-term) (T); and setting of treatment, for example, inpatient psychiatry versus outpatient community clinics (S). To ensure external validity or generalizability, the Task Force should insist that a clear PICOTS statement is included in the systematic review, clearly defining the population of interest, the intervention, the comparisons considered, outcomes examined, and timing of outcome assessment.

In addition, the systematic review should evaluate the degree to which trials included in the review took steps to ensure treatment fidelity. Bellg et al. (2004) provide a thorough discussion of elements of treatment fidelity and steps that can be taken to enhance treatment fidelity in trials of behavior change studies. In the context of systematic reviews, there are no standard instruments for assessing steps taken to ensure treatment fidelity in included trials. Elements that were included in Chambless and Hollon's (1998) original EST definition, and that continue to be evaluated in evidence reviews, are therapist qualifications and training, the use of a treatment manual, and monitoring of the degree to which the treatment is implemented according to the manual.

#### Assessing Risk of Bias

The original EST criteria (Chambless et al., 1998) operationalized methodological adequacy as including the use of a treatment manual, a well-characterized sample, and random assignment to treatment and

control conditions. Since these criteria were published, however, standards for evaluating both the external and internal validity of treatment trials have evolved substantially, and there are now several widely accepted methods of determining methodological adequacy that should be considered. We recommend that authors of systematic reviews assess validity using the Cochrane Risk of Bias Tool (Higgins et al., 2011). This tool, widely regarded as the standard for evaluating risk of bias in RCTs included in systematic reviews, provides a rating system and criteria by which individual RCTs are evaluated according to the potential sources of bias related to (a) adequate allocation sequence generation; (b) concealment of allocation to conditions; (c) blinding of participants, personnel, and outcome assessors; (d) incomplete outcome data; (e) selective outcome reporting; and (f) other sources of bias (see online supplement). Adequate sequence allocation ensures that study participants were appropriately randomized to study conditions. Allocation concealment means that the random assignment is implemented in a way that cannot be predicted by participants or key study personnel. Blinding of key study personnel and outcome assessors ensures that those personnel in a position to affect outcome data are unaware of participants' study condition. Blinding of participants indicates that the participants themselves are unaware of study condition. Blinding of participants is not commonly used (and is often not possible) in trials of psychotherapy. In many cases, such as when a treatment group is compared to a nontreatment group, this would be reflected as a methodological limitation common to studies of psychological treatments. However, the Cochrane system allows a "low risk of bias" determination on this item when the outcome and outcome measurement are not likely to be influenced by lack of blinding, or outcome assessment was blinded and the nonblinding of participants was unlikely to introduce bias. Blinding of participants, or at least to study aims and hypotheses, would be possible in comparison trials between two psychological treatments; full blinding of participants has been noted in some studies of computerized cognitive bias modification training (e.g., Amir, Beard, Burns, & Bomyea, 2009). Appropriate handling of incomplete (missing) outcome data due to attrition during the study or to exclusions from the analysis helps ensure that the outcome analyses adequately represent the outcomes of the sample. Examination of selective outcome reporting helps identify whether important (possibly nonsignificant) findings were omitted from the report of the study (Higgins & Green, 2008). Whether or not clinical trials are registered and, if so, ascertaining whether published outcomes are consistent with registered outcomes is an important step in a systematic review (Milette, Roseman, & Thombs, 2011; Thombs, Kwakkenbos, & Coronado-Montoya, 2014).

Across all dimensions, trials are rated as *high risk of bias, unclear risk of bias,* or *low risk of bias*. Cochrane advocates that systematic reviews assess the potential influence on outcomes of each of these dimensions separately and recommends against attempting to generate a single score or rating of overall bias (Higgins & Green, 2008). Summary scores tend to confound the quality of reporting with the quality of trial conduct, to assign weights to different items in ways that are difficult to justify, and to yield inconsistent and unpredictable associations with intervention effect estimates (Greenland & O'Rourke, 2001; Juni, Witschi, Bloch, & Egger, 1999).

Both individual trials and systematic reviews can be judged as having low, unclear, or high risk of bias (see online supplement). A systematic review would be graded to be at low risk of bias when the conclusions from the review are based on evidence judged to be at low risk of bias, according to the GRADE dimensions described above. Note that this grading system differs markedly from those originally proposed by the Division 12 Task Force (e.g., Chambless et al., 1998). Two well-conducted studies are no longer considered sufficient; this system would now require that the conclusions of the systematic review are based on studies deemed to be of high quality.

Assessment of risk of bias requires human judgment (Higgins et al., 2011), and, unfortunately, there is no quantitative algorithm that will consistently lead to reliable and valid assessment. Thus, there will always be room for disagreement and debate about the merits of individual studies and about the quality of research across studies for a given treatment. Assessment of whether a particular methodological concern in a trial creates a risk of bias requires both knowledge of the trial methods and a judgment about whether those

methods are likely to have led to a risk of bias. The Cochrane Risk of Bias Tool, at least, makes the decision process transparent and provides accessible guidance for how decisions should be made (Higgins & Green, 2008).

#### Additional Considerations for the Evaluation of Systematic Reviews and Recommendations for Implementation

Systematic reviews will be examined for both short-term and long-term outcomes. Long-term outcomes will generally be defined as outcomes collected some time after treatment discontinuation; however, we recognize that some treatments may include a low-intensity "maintenance" phase that continues for a long time after the more acute phase; outcomes during the maintenance phase might be appropriate for consideration as long-term effects. Effects for both symptom reduction and functional outcomes will be coded, relying on validated measures that are appropriate for the population and treatment under study. Finally, the review will note whether the treatment has demonstrated effectiveness (e.g., more diagnostically complex patients, effectiveness with nonrandomized patients, effectiveness when used by nonacademic practitioners, and utility in open-ended, flexible practice) in addition to efficacy.

### STEP 2: COMMITTEE-BASED EVIDENCE REVIEW USING THE GRADE TOOL

The systematic review, having been graded for risk of bias, must then be translated into practical recommendations that will address the concerns of a broad range of patients, presenting problems, clinicians, and clinical settings. As it is unlikely that any statistical algorithm will ever be able to provide such guidance consistently, the process of recommending treatments must ultimately be a product of human judgment. The systematic review will provide raw information about the strength of clinical effects, as well as the risk of bias of the studies evaluating the treatment. In addition to those basic assessments, a determination of whether psychological treatments should be recommended to clinicians, consumers, and other stakeholders must be based on the strength and quality of existing evidence and a comparison of the likely benefits versus burden, cost, and potential harms of the treatment. The best strategy one can use in such a situation is to provide a clear framework to guide the decision-making process, and to make the process as transparent as possible so that the public can understand how these judgments were made.

A number of different strategies have been employed by guideline developers to attempt to make clear the strength of evidence and recommendations, although the most widely used system is the GRADE system (Atkins et al., 2004; Guyatt et al., 2008). The aim of the GRADE system is to rate quality of evidence and strength of recommendations in a manner that is explicit, comprehensive, transparent, and pragmatic. Factors that are taken into account in making these decisions include the methodological quality of the evidence that supports estimates of benefits, costs, and harms; the importance of the outcome that the treatment improves; the magnitude of the treatment effect and the precision of its estimate; the burden, costs, and potential risks associated with the therapy; and other consumer values that might be expected to influence their decision process.

#### Using the GRADE System for Treatment Recommendations

The GRADE system rates evidence quality as *high, moderate,* or *low*. Evidence is judged to be *high quality* when reviewers can be highly confident that the true effect lies close to that of the estimate of the effect. For example, evidence is judged as high quality if all of the following apply:

1. There is a wide range of studies included in the analyses with no major limitations.
2. There is little variation between studies.
3. The summary estimate has a narrow confidence interval.

Evidence is judged to be *moderate quality* when reviewers conclude that the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. For example, evidence is judged as moderate quality if any of the following applies:

1. There are only a few studies, and some have limitations but not major flaws.
2. There is some variation between studies, or the confidence interval of the summary estimate is wide.

Evidence is judged to be *low quality* when the true effect may be substantially different from the estimate of the effect. For example, evidence is judged as low quality if any of the following applies:

1. The studies have major flaws.
2. There is important variation between studies.
3. The confidence interval of the summary estimate is very wide.

In the GRADE system to determine quality of evidence, evidence based on RCTs begins as high-quality evidence, but such evidence could be downgraded based on concerns such as study limitation, inconsistency of results, indirectness of evidence, imprecision, and reporting bias. Other types of studies begin as lower-quality evidence, but may be upgraded if merited on a case-by-case basis.

The GRADE process typically results in a *weak* or a *strong* recommendation. For the psychotherapy evaluation, we suggest that the GRADE system be modified to include a third category. A three-tier system would better correspond to the current reality that few existing trials of psychological treatments have assessed functional and disability outcomes, despite the fact that such outcomes may be more important than symptom outcomes. Thus, based on evidence from the submitted systematic review and meta-analysis, we recommend that the Task Force use an adapted GRADE process and make one of three recommendations for the empirical support of a psychological treatment: *weak, strong,* or *very strong*. Treatments not meriting at least a *weak* recommendation (e.g., no systematic review is available, or the outcomes of treatment studies do not satisfy the minimal criteria for a weak recommendation) will be described simply as lacking sufficient evidence of efficacy. The criteria for these recommendations are shown in Table 4.

The GRADE recommendations are hierarchical; treatments are ranked according to the highest level of recommendation obtained. A *very strong* recommendation is made when there is high-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated, as well as a clinically meaningful effect on functional outcomes, with significant improvement noted at immediate posttreatment and at a follow-up (treatment discontinuation) interval of not less than 3 months, with relatively little risk of harm and reasonable resource use, and there is at least one well-conducted study that has demonstrated effectiveness of that treatment in nonresearch settings (e.g., settings that provide routine clinical care, such as community mental health centers, inpatient or outpatient treatment facilities, health maintenance organizations, or private practices). We recognize that this level of recommendation may be largely aspirational at this time, although some treatments will merit a *very strong* recommendation at present. In other cases, the establishment of this level of recommendation sets a bar for the planning of future treatment outcome studies.

A *strong* recommendation, which will be more readily attainable for many treatments at this time, requires the presence of moderate- to high-quality evidence

**Table 4.** Modified GRADE recommendations for psychological treatments based on systematic reviews (adapted from Guyatt et al., 2008)

| Recommendation | |
|---|---|
| Very strong recommendation | All of the following:<br>• There is high-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated.<br>• There is high-quality evidence that the treatment produces a clinically meaningful effect on functional outcomes.<br>• There is high-quality evidence that the treatment produces a clinically meaningful effect on symptoms and/or functional outcomes at least 3 months after treatment discontinuation.<br>• At least one well-conducted study has demonstrated effectiveness in nonresearch settings. |
| Strong recommendation | At least one of the following:<br>• There is moderate- to high-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated.<br>• There is moderate- to high-quality evidence that the treatment produces a clinically meaningful effect on functional outcomes. |
| Weak recommendation | Any of the following:<br>• There is only low- or very low-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated.<br>• There is only low- or very low-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated as well as on functional outcomes.<br>• There is moderate- to high-quality evidence that the effect of the treatment, although statistically significant, may not be of a magnitude that is clinically meaningful. |

that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated, *or* on functional outcomes, again, with a clear positive balance in consideration of benefits versus possible harms and resource use. Evidence of external effectiveness of generalizability is not required for this level of recommendation.

*Weak* recommendations, which are not necessarily intended to discourage the use of treatments, are made when there is only low- or very low-quality evidence that the treatment produces a clinically meaningful effect on symptoms of the disorder being treated and/or functional outcomes, or when the evidence suggests that the effects of the treatment may not be clinically meaningful (although they may be statistically significant). In the case of a *weak* recommendation, it is not clear that gains from treatment warrant the resources involved, and patient preferences will be central in determining whether engaging in the treatment is the best possible decision.

#### Taking Contextual Factors Into Account

It would be prohibitive, on several levels, for the Task Force to explicitly require comparative effectiveness analyses of all possible treatments or analyses of cost-effectiveness. However, when there are obvious concerns, the committee should be able to incorporate them into the recommendation. This might occur, for instance, in contextualizing the clinical meaningfulness of a treatment effect when there are other psychological treatments that have well-documented and much larger effects. Similarly, if a treatment generates an effect that is similar to other well-studied treatments, but requires a very large number of sessions or length of time to generate the same effect at a much higher cost, then the Task Force may take this into consideration.

The Task Force may take into account the purported mechanism or active ingredient(s) of treatment and may upgrade or downgrade the recommendation based on the quality of evidence supporting that mechanism or ingredient(s). It is conceptually difficult to standardize this consideration into the criteria, as admittedly the mechanisms of many efficacious treatments are unclear. However, to the extent that a given treatment is based on a specific purported mechanism or relies strongly on a particular treatment ingredient, the board can and should consider whether those assertions are supported. Single-case designs are often particularly useful for such purposes. Such consideration would help reduce the risk of allocating resources to elements of treatment that are inert or worse. Below, we describe a longer-term plan for identifying active therapeutic ingredients.

Although most ESTs appear effective when applied to minority groups with specific disorders (e.g., Miranda et al., 2005), it cannot be automatically assumed that an EST that is effective for the majority population will be equally effective among minority groups. Therefore, it is important that research on treatment efficacy and effectiveness attend to the generalizability of effects across diverse populations. At this time, it would be difficult to require a documentation of efficacy or effectiveness across minority groups, given the many nuances associated with assessing, treating, and modifying treatments for different populations. Furthermore, it would likely be counterproductive to identify a treatment as appropriate for minority populations unless all such populations had been studied. We therefore recommend that nominators of treatments identify specific studies demonstrating efficacy or effectiveness within a particular underrepresented group and that such findings be highlighted in the presentation of the treatment and by the Task Force when recommendations are made.

## CONCLUSIONS AND FUTURE DIRECTIONS

The EST movement has, overall, provided positive direction for clinical psychology. However, several valid criticisms of the process have been offered. In this article, we propose a new approach for identifying ESTs and for recommending specific psychological treatments to practitioners, consumers, and other stakeholders. Twenty years after the original Division 12 Task Force report, such an update is long overdue. Although clinical psychology once led the way in articulating how a treatment should be determined to be empirically supported (and although many other healthcare fields still look to those original criteria for guidance), advances in the field of evidence-based medicine have rendered the old criteria obsolete.

In this article, we propose a two-stage process by which the Society of Clinical Psychology/Division 12

may help bridge the gap between the current, outdated EST criteria and the planned treatment guidelines from APA. The aim is to begin to evaluate treatments in a manner that parallels and will support the methods proposed by APA, but in a manner that lends itself to more rapid dissemination of scientific findings to those who would benefit most from them. We propose that the process of identifying one or two positive studies for a treatment ceases, and that in its place we begin evaluating systematic reviews of the treatment outcome literature, weighting them according to the risk of bias in the studies contributing the review. We further recommend that instead of labeling treatments as "well established" or "probably efficacious," as is currently done under the current system, we translate the research findings into clear recommendations of *very strong, strong,* or *weak,* using well-established, widely accepted, and transparent grading guidelines. These steps, which can be implemented immediately, will greatly improve the quality of information that is disseminated.

As mentioned earlier, the APA Presidential Task Force on Evidence-Based Practice (2006) defines EBP as consisting of three components of information: best available research evidence, clinical expertise, and patient characteristics. In our view, these three components play different critical roles in clinical decision-making (e.g., Tolin, 2014), in which the best available research evidence forms the basis of clinical decisions and is interpreted, adjusted, and implemented through clinical expertise and patient characteristics. A skilled evidence-based practitioner will first identify the EST that most closely matches the concerns presented by a given patient. One EST is selected over the others by examining the available research evidence that shows the strength of the treatment and the quality of evidence. ESTs may also need to be adapted or augmented, based on patient characteristics such as comorbid psychopathology, situational factors, or cultural and demographic features. Such selection, adaptation, and augmentation procedures derive from the expertise of the clinician, guided wherever possible by the best scientific evidence (with the understanding that such research will rarely line up perfectly with the clinical problem). It is noted in this model that clinical expertise and patient characteristics do not trump the best available research evidence, nor should all three factors be considered an "either-or" selection. That is, skillful EBP does not involve selecting a treatment based on research evidence *or* on the clinician's expertise *or* on patient characteristics. Rather, the best available research evidence (including ESTs) forms the basis of clinical judgment, with additional selection and modification based on clinical expertise and patient characteristics. The modifications to how ESTs are evaluated and disseminated proposed in this article are hoped to help EBP practitioners reach appropriate conclusions based on the best available clinical science.

## REFERENCES


American Psychiatric Association. (2009). *Practice guideline for the treatment of patients with panic disorder* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (2010). *Practice guideline for the treatment of patients with major depressive disorder* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

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

American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist*, *57*, 1052–1059. doi:10.1037//0003-066X.57.12.1052

Amir, N., Beard, C., Burns, M., & Bomyea, J. (2009). Attention modification program in individuals with generalized anxiety disorder. *Journal of Abnormal Psychology*, *118*(1), 28–33. doi:10.1037/a0012589

Andreasen, N. C., Carpenter, W. T., Jr., Kane, J. M., Lasser, R. A., Marder, S. R., & Weinberger, D. R. (2005). Remission in schizophrenia: Proposed criteria and rationale for consensus. *American Journal of Psychiatry*, *162*, 441–449. doi:10.1176/appi.ajp.162.3.441

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*, 271–285. doi:10.1037/0003-066X.61.4.271

Areán, P. A., & Kraemer, H. C. (2013). *High-quality psychotherapy research*. New York, NY: Oxford University Press.

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., ... GRADE Working Group. (2004). Systems for grading the quality of evidence and the strength of

recommendations I: Critical appraisal of existing approaches: The GRADE Working Group. *BMC Health Services Research*, *4*(1), 38. doi:10.1186/1472-6963-4-38

Baker, M., & Kleijnen, J. (2000). The drive towards evidence-based health care. In N. Rowland & S. Goss (Eds.), *Evidence-based counseling and psychological therapies: Research and applications* (pp. 13–29). New York, NY: Routledge.

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., & Ory, M., … Treatment Fidelity Workgroup of the NIH-BCC. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*, *23*, 443–451. doi:10.1037/0278-6133.23.5.443

Benish, S. G., Quintana, S., & Wampold, B. E. (2011). Culturally adapted psychotherapy and the legitimacy of myth: A direct-comparison meta-analysis. *Journal of Counseling Psychology*, *58*, 279–289. doi:10.1037/a0023626

Bernardy, K., Fuber, N., Kollner, V., & Hauser, W. (2010). Efficacy of cognitive-behavioral therapies in fibromyalgia syndrome—A systematic review and meta-analysis of randomized controlled trials. *Journal of Rheumatology*, *37*, 1991–2005. doi:10.3899/jrheum.100104

Bero, L. A. (2013). Why the Cochrane risk of bias tool should include funding source as a standard item. *Cochrane Database Systematic Review*, *12*, ED000075. doi:10.1002/14651858.ED000075

Beutler, L. E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting and Clinical Psychology*, *66*, 113–120. doi:10.1037/0022-006X.66.1.113

Blatt, S. J., & Zuroff, D. C. (2005). Empirical evaluation of the assumptions in identifying evidence based treatments in mental health. *Clinical Psychology Review*, *25*, 459–486. doi:10.1016/j.cpr.2005.03.001

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. doi:10.1037/0022-006X.66.1.136

Castonguay, L. G., & Beutler, L. E. (2006). *Principles of therapeutic change that work*. New York, NY: Oxford University Press.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., … Woody, S. (1998). Update on empirically validated therapies. II. *The Clinical Psychologist*, *51*(1), 3–16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18. doi:10.1037/0022-006X.66.1.7

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716. doi:10.1146/annurev.psych.52.1.685

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Commission on Accreditation. (2013). *Guidelines and principles for accreditation of programs in professional psychology*. Washington, DC: American Psychological Association.

Cowen, E. L. (1991). In pursuit of wellness. *American Psychologist*, *46*, 404. doi:10.1037/0003-066X.46.4.40

Djulbegovic, B., & Guyatt, G. H. (2014). Evidence-based practice is not synonymous with delivery of uniform health care. *JAMA*, *312*, 1293–1294. doi:10.1001/jama.2014.10713

Dobson, K., & Beshai, S. (2013). The theory-practice gap in cognitive behavioral therapy: Reflections and a modest proposal to bridge the gap. *Behavior Therapy*, *44*, 559–567. doi:10.1016/j.beth.2013.03.002

Doyle, A. C., & Pollack, M. H. (2003). Establishment of remission criteria for anxiety disorders. *Journal of Clinical Psychiatry*, *64*(Suppl. 15), 40–45.

Feeny, N. C., Zoellner, L. A., & Foa, E. B. (2002). Treatment outcome for chronic PTSD among female assault victims with borderline personality characteristics: A preliminary examination. *Journal of Personality Disorders*, *16*(1), 30–40. doi:10.1521/pedi.16.1.30.22555

Fensterheim, H., & Raw, S. D. (1996). Psychotherapy research is not psychotherapy practice. *Clinical Psychology: Science and Practice*, *3*, 168–171. doi:10.1111/j.1468-2850.1996.tb00067.x

Fonagy, P. (1999). Achieving evidence-based psychotherapy practice: A psychodynamic perspective on the general acceptance of treatment manuals. *Clinical Psychology: Science and Practice*, *6*, 442–444. doi:10.1093/clipsy.6.4.442

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., … Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder: Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, *48*, 851–855. doi:10.1001/archpsyc.1991.01810330075011

Franklin, M. E., Abramowitz, J. S., Kozak, M. J., Levitt, J. T., & 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. doi:10.1037/0022-006X.68.4.59

Friedman, L. S., & Richter, E. D. (2004). Relationship between conflicts of interest and research results. *Journal of General Internal Medicine*, *19*(1), 51–56. doi:10.1111/j.1525-1497.2004.30617.x

Gaffan, E. A., Tsaousis, I., & Kemp-Wheeler, S. M. (1995). Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *63*, 966–980. doi:10.1037/0022-006X.63.6.966

Gibbons, C. R., Stirman, S. W., DeRubeis, R. J., Newman, C. F., & Beck, A. T. (2013). Research setting versus clinic setting: Which produces better outcomes in cognitive therapy for depression? *Cognitive Therapy and Research*, *37*, 605–612. doi:10.1007/s10608-012-9499-7

Gill, T. M., & Feinstein, A. R. (1994). A critical appraisal of the quality of quality-of-life measurements. *JAMA*, *272*, 619–626. doi:10.1001/jama.1994.03520080061045

Goldfried, M. R., & Eubanks-Carter, C. (2004). On the need for a new psychotherapy research paradigm: Comment on Westen, Novotny, and Thompson-Brenner (2004). *Psychological Bulletin*, *130*, 669–673; author reply 677-683. doi:10.1037/0033-2909.130.4.669

Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained alliance. *American Psychologist*, *51*, 1007–1016. doi:10.1037//0003-066X.51.10.1007

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. doi:10.1037/0022-006X.66.1.143

Gonzales, J. J., & Chambers, D. A. (2002). The tangled and thorny path of science to practice: Tensions in interpreting and applying "evidence." *Clinical Psychology: Science and Practice*, *9*, 204–209. doi:10.1093/clipsy.9.2.204

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, *2*, 463–471. doi:10.1093/biostatistics/2.4.463

Griner, D., & Smith, T. B. (2006). Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy (Chic)*, *43*, 531–548. doi:10.1037/0033-3204.43.4.531

Guyatt, G. H., Gutterman, D., Baumann, M. H., Addrizzo-Harris, D., Hylek, E. M., Phillips, B., … Schunemann, H. (2006). Grading strength of recommendations and quality of evidence in clinical guidelines: Report from an American College of Chest Physicians task force. *Chest*, *129*(1), 174–181. doi:10.1378/chest.129.1.174

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., … GRADE Working Group. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, *336*, 924–926. doi:10.1136/bmj.39489.470347.AD

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, *35*, 639–665. doi:10.1016/S0005-7894(04)80013-3

Henry, W. P. (1998). Science, politics, and the politics of science: The use and misuse of empirically validated treatment research. *Psychotherapy Research*, *8*(2), 126–140. doi:10.1093/ptr/8.2.126

Herbert, J. D. (2003). The science and practice of empirically supported treatments. *Behavior Modification*, *27*, 412–430. doi:10.1177/0145445503253836

Higgins, J. P., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., … Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, d5928. doi:10.1136/bmj.d5928

Higgins, J. P., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Hoboken, NJ: Wiley-Blackwell.

Hollon, S. D. (1999). Allegiance effects in treatment research: A commentary. *Clinical Psychology: Science and Practice*, *6*, 107–112. doi:10.1093/clipsy.6.1.107

Hollon, S. D., Arean, P. A., Craske, M. G., Crawford, K. A., Kivlahan, D. R., Magnavita, J. J., … Kurtzman, H. (2014). Development of clinical practice guidelines. *Annual Review of Clinical Psychology*, *10*, 213–241. doi:10.1146/annurev-clinpsy-050212-185529

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Child*, *71*(2), 165–179. doi:10.1177/001440290507100203

Hunt, S. M., & McKenna, S. P. (1993). Measuring quality of life in psychiatry. In S. R. Walker & R. M. Rossner (Eds.), *Quality of life assessment: Key issues in the 1990s* (pp. 343–354). Boston: Kluwer Academic.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., … Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*, 748–751. doi:10.1176/appi.ajp.2010.09091379

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academies Press.

Institute of Medicine. (2011a). *Clinical practice guidelines we can trust*. Washington, DC: National Academies Press.

Institute of Medicine. (2011b). *Finding what works in health care: Standards for systematic reviews*. Washington, DC: National Academies Press.

Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, *14*, 951–957. doi:10.1111/j.1365-2753.2008.00986.x

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Toward a standard definition of clinically significant change. *Behavior Therapy*, *17*, 308–311. doi:10.1016/S0005-7894(86)80061-2

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, *282*, 1054–1060. doi:10.1001/jama.282.11.1054

Keltner, J. R., Vaida, F., Ellis, R. J., Moeller-Bertram, T., Fitzsimmons, C., Duarte, N. A., … CHARTER Group. (2012). Health-related quality of life 'well-being' in HIV distal neuropathic pain is more strongly associated with depression severity than with pain intensity. *Psychosomatics*, *53*, 380–386. doi:10.1016/j.psym.2012.05.002

Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., Thase, M. E., … Leon, A. C. (2010). A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry*, *51*, 319–324. doi:10.1016/j.comppsych.2009.07.001

Kushner, S. C., Quilty, L. C., McBride, C. & Bagby, R. M. (2009). A comparison of depressed patients in randomized versus nonrandomized trials of antidepressant medication and psychotherapy. *Depression and Anxiety*, *26*, 666–673. doi:10.1002/da.20566

Lambert, M. J., & Bailey, R. J. (2012). Measures of clinically significant change. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 3: Data analysis and research publication*, (pp. 147–160). Washington, DC: American Psychological Association.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy*, *38*, 357–361. doi:10.1037/0033-3204.38.4.357

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K. M., Okiishi, J., Burlingame, G. M., … Reisinger, C. R. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.2)*. Wilmington, DE: American Professional Credentialing Services.

Layard, R., & Clark, D. M. (2014). *Thrive: The power of evidence-based psychological therapies*. London, UK: Penguin.

Lejuez, C. W., Hopko, D. R., & Hopko, S. D. (2001). A brief behavioral activation treatment for depression: Treatment manual. *Behavior Modification*, *25*, 255–286. doi:10.1177/0145445501252005

Leon, A. C., Solomon, D. A., Mueller, T. I., Turvey, C. L., Endicott, J., & Keller, M. B. (1999). The Range of Impaired Functioning Tool (LIFE-RIFT): A brief measure of functional impairment. *Psychological Medicine*, *29*, 869–878. doi:10.1017/S0033291799008570

Levant, R. F. (2004). The empirically validated treatments movement: A practitioner/educator perspective. *Clinical Psychology: Science and Practice*, *11*, 219–224. doi:10.1093/clipsy.bph075

Lewinsohn, P. M., Biglan, A., & Zeiss, A. M. (1976). Behavioral treatment for depression. In P. O. Davidson (Ed.), *Behavioral management of anxiety, depression, and pain* (pp. 91–146). New York, NY: Brunner/Mazel.

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ*, *326*, 1167–1170. doi:10.1136/bmj.326.7400.1167

Leykin, Y., & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating association from bias. *Clinical Psychology: Science and Practice*, *16*, 54–65. doi:10.1111/j.1468-2850.2009.01143.x

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, *6*(7), e1000100. doi:10.1371/journal.pmed.1000100

Lock, J., Le Grange, D., Agras, W. S., Moye, A., Bryson, S. W., & Jo, B. (2010). Randomized clinical trial comparing family-based treatment with adolescent-focused individual therapy for adolescents with anorexia nervosa. *Archives of General Psychiatry*, *67*, 1025–1032. doi:10.1001/archgenpsychiatry.2010.128

Lohr, J. M., Tolin, D. F., & Lilienfeld, S. O. (1998). Efficacy of eye movement desensitization and reprocessing: Implications for behavior therapy. *Behavior Therapy*, *29*, 123–156. doi:10.1016/S0005-7894(98)80035-X

van Loon, A., van Schaik, A., Dekker, J., & Beekman, A. (2013). Bridging the gap for ethnic minority adult outpatients with depression and anxiety disorders by culturally adapted treatments. *Journal of Affective Disorders*, *147*(1–3), 9–16. doi:10.1016/j.jad.2012.12.014

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., … Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, *6*, 95–106. doi:10.1093/clipsy.6.1.95

Lundh, A., Sismondo, S., Lexchin, J., Busuioc, O. A., & Bero, L. (2012). Industry sponsorship and research outcome. *Cochrane Database Systematic Review*, *12*, MR000033. doi:10.1002/14651858.MR000033.pub2

Martell, C. R., Addis, M. E., & Jacobson, N. S. (2001). *Depression in context: Strategies for guided action*. New York, NY: W. W. Norton.

McIntyre, R. S., Fallu, A., & Konarski, J. Z. (2006). Measurable outcomes in psychiatric disorders: Remission as a marker of wellness. *Clinical Therapeutics*, *28*, 1882–1891. doi:10.1016/j.clinthera.2006.11.007

Milette, K., Roseman, M., & Thombs, B. D. (2011). Transparency of outcome reporting and trial registration of randomized controlled trials in top psychosomatic and behavioral health journals: A systematic review. *Journal of Psychosomatic Research*, *70*, 205–217. doi:10.1016/j.jpsychores.2010.09.015

Milrod, B., Leon, A. C., Busch, F., Rudden, M., Schwalberg, M., Clarkin, J., … Shear, M. K. (2007). A randomized controlled clinical trial of psychoanalytic psychotherapy for panic disorder. *American Journal of Psychiatry*, *164*, 265–272. doi:10.1176/appi.ajp.164.2.265

Miranda, J., Bernal, G., Lau, A., Kohn, L., Hwang, W. C., & LaFromboise, T. (2005). State of the science on psychosocial interventions for ethnic minorities. *Annual Review of Clinical Psychology*, *1*, 113–142. doi:10.1146/annurev.clinpsy.1.102803.143822

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology*, *1*, 2. doi:10.7326/0003-4819-134-8-200104170-00011

Muldoon, M. F., Barger, S. D., Flory, J. D., & Manuck, S. B. (1998). What are quality of life measurements measuring? *British Medical Journal*, *316*, 542–545. doi:10.1136/bmj.316.7130.542

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. H. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *British Journal of Psychiatry*, *180*, 461–464. doi:10.1192/bjp.180.5.46

National Institute for Clinical Excellence. (2011). *Generalised anxiety disorder and panic disorder (with or without agoraphobia) in adults: Management in primary, secondary and community care (National Clinical Practice Guideline number CG113)*. London: British Psychological Society and Royal College of Psychiatrists.

Norcross, J. C. (1999). Collegially validated limitations of empirically validated treatments. *Clinical Psychology: Science and Practice*, *6*, 472–476. doi:10.1093/clipsy.6.4.472

O'Farrell, T. J., Cutter, H. S. G., Choquette, K. A., Floyd, F. J., & Bayog, R. D. (1992). Behavioral marital therapy for male alcoholics: Marital and drinking adjustment during the two years after treatment. *Behavior Therapy*, *23*, 529–549. doi:10.1016/S0005-7894(05)80220-5

Parker, G., Parker, I., Brotchie, H., & Stuart, S. (2006). Interpersonal psychotherapy for depression? The need to define its ecological niche. *Journal of Affective Disorders*, *95* (1–3), 1–11. doi:10.1016/j.jad.2006.03.019

Perlis, R. H., Perlis, C. S., Wu, Y., Hwang, C., Joseph, M., & Nierenberg, A. A. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry*, *162*, 1957–1960. doi:10.1176/appi.ajp.162.10.1957

Perri, M. G., Sears, S. F., Jr., & Clark, J. E. (1993). Strategies for improving maintenance of weight loss: Toward a continuous care model of obesity management. *Diabetes Care*, *16*(1), 200–209. doi:10.2337/diacare.16.1.200

Petry, N. M., Martin, B., Cooney, J. L., & Kranzler, H. R. (2000). Give them prizes, and they will come: Contingency management for treatment of alcohol dependence. *Journal of Consulting and Clinical Psychology*, *68*, 250–257. doi:10.1037//0022-006X.68.2.250

Riehm, K. E., Azar, M., & Thombs, B. D. (2015). Transparency of outcome reporting and trial registration of randomized controlled trials in top psychosomatic and behavioral health journals: A 5-year follow-up. *Journal of Psychosomatic Research*, *79*(1), 1–12. doi:10.1016/j.jpsychores.2015.04.010

Roseman, M., Milette, K., Bero, L. A., Coyne, J. C., Lexchin, J., Turner, E. H., & Thombs, B. D. (2011). Reporting of conflicts of interest in meta-analyses of trials of pharmacological treatments. *JAMA*, *305*, 1008–1017. doi:10.1001/jama.2011.257

Roseman, M., Turner, E. H., Lexchin, J., Coyne, J. C., Bero, L. A., & Thombs, B. D. (2012). Reporting of conflicts of interest from drug trials in Cochrane reviews: Cross sectional study. *British Medical Journal*, *345*, e5155. doi:10.1136/bmj.e5155

Rosen, G. M., & Davison, G. C. (2003). Psychology should list empirically supported principles of change (ESPs) and not credential trademarked therapies or other treatment packages. *Behavior Modification*, *27*, 300–312. doi:10.1177/0145445503027003003

Ruzek, J. I., Karlin, B. E., & Zeiss, A. M. (2012). Implementation of evidence-based psychological treatments in the Veterans Health Administration. In R. K. McHugh & D. H. Barlow (Eds.), *Dissemination of*

*evidence-based psychological treatments* (pp. 78–96). New York, NY: Oxford University Press.

Savovic, J., Jones, H. E., Altman, D. G., Harris, R. J., Juni, P., Pildal, J., … Sterne, J. A. (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Annals of Internal Medicine*, *157*, 429–438. doi:10.7326/0003-4819-157-6-201209180-00537

Schneier, F. R., Heckelman, L. R., Garfinkel, R., Campeas, R., Fallon, B. A., Gitow, A., … Liebowitz, M. R. (1994). Functional impairment in social phobia. *Journal of Clinical Psychiatry*, *55*, 322–331.

Seligman, M. E. (1995). The effectiveness of psychotherapy: The *Consumer Reports* Study. *American Psychologist*, *50*, 965–974. doi:10.1037/0003-066X.50.12.965

Seligman, M. E. (1996). Science as an ally of practice. *American Psychologist*, *51*, 1072–1079. doi:10.1037/0003-066X.51.10.1072

Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, *23*, 139–146. doi:10.1177/0963721414524773

Shea, B. J., Bouter, L. M., Peterson, J., Boers, M., Andersson, N., Ortiz, Z., … Grimshaw, J. M. (2007). External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS ONE*, *2*, e1350. doi:10.1371/journal.pone.0001350

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., … Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, 7, 10. doi:10.1186/1471-2288-7-10

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., … Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, *62*, 1013–1020. doi:10.1016/j.jclinepi.2008.10.009

Sheehan, D. V. (2008). Sheehan Disability Scale. In A. Rush, M. First, & D. Blacker, (Eds.), *Handbook of psychiatric measures* (2nd. ed., pp. 100–102). Washington, DC: American Psychiatric Publishing.

Simpson, H. B., Huppert, J. D., Petkova, E., Foa, E. B., & Liebowitz, M. R. (2006). Response versus remission in obsessive-compulsive disorder. *Journal of Clinical Psychiatry*, *67*, 269–276. doi:10.4088/jcp.v67n0214

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., … Wagner, A. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders*, *37*, 354–366. doi:10.1007/s10803-006-0173-3

Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Rothman, A. (2005). Can the randomized controlled trial literature generalize to nonrandomized patients? *Journal of Consulting and Clinical Psychology*, *73*, 127–135. doi:10.1037/0022-006X.73.1.127

Task Force on Promotion and Dissemination of Psychological Procedures. (1993). Training in and dissemination of empirically-validated psychological treatments: Report and recommendation. *The Clinical Psychologist*, *48*, 3–23.

Thoma, N. C., McKay, D., Gerber, A. J., Milrod, B. L., Edwards, A. R., & Kocsis, J. H. (2012). A quality-based review of randomized controlled trials of cognitive-behavioral therapy for depression: An assessment and metaregression. *American Journal of Psychiatry*, *169*, 22–30. doi:10.1176/appi.ajp.2011.11030433

Thombs, B. D., Kwakkenbos, L., & Coronado-Montoya, S. (2014). Trial registration in rheumatology: The next step. *Arthritis Care and Research*, *66*, 1435–1437. doi:10.1002/acr.22335

Tolin, D. F. (2014). Evidence-based practice: Three-legged stool or filter system? *The Clinical Psychologist*, *67*(3), 1–3.

Vatne, S., & Bjorkly, S. (2008). Empirical evidence for using subjective quality of life as an outcome variable in clinical studies: A meta-analysis of correlates and predictors in persons with a major mental disorder living in the community. *Clinical Psychology Review*, *28*, 869–889. doi:10.1016/j.cpr.2008.01.001

Veterans Health Administration, Department of Defense. (2004). VA/DoD clinical practice guideline for the management of post-traumatic stress. Version 1.0. Retrieved January 2004.

Veterans Health Administration, Department of Defense. (2009). VA/DoD clinical practice guideline for the management of major depressive disorder. Version 2.0. Retrieved June 2015.

Wachtel, P. L. (2010). Beyond "ESTs:" Problematic assumptions in the pursuit of evidence-based practice. *Psychoanalytic Psychology*, *27*, 251–272. doi:10.1037/a0020532

Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, *130*, 631–663. doi:10.1037/0033-2909.130.4.63

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Juni, P., Altman, D. G., … Sterne, J. A. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. *British Medical Journal*, *336*, 601–605. doi:10.1136/bmj.39465.451748.AD

Yeomans, F. E., Levy, K. N., & Caligor, E. (2013). Transference-focused psychotherapy. *Psychotherapy (Chic)*, *50*, 449–453. doi:10.1037/a0033417

Zalta, A. K. (2015). Psychological mechanisms of effective cognitive-behavioral treatments for PTSD. *Current Psychiatry Reports*, *17*, 560. doi:10.1007/s11920-015-0560-6

Received April 4, 2015; revised June 15, 2015; accepted June 19, 2015.

#### SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Table S1.** Domains for determining risk of bias in systematic reviews of psychotherapy outcome studies.

**Table S2.** Summary assessments of risk of bias.

**Table S3.** Scoring criteria for the Assessment of Multiple Systematic Reviews System.