Clinical Psychology Review 34 (2014) 519–530

Contents lists available at ScienceDirect

Clinical Psychology Review

Is the Dodo bird endangered in the 21st century? A meta-analysis of treatment comparison studies

David K. Marcus *, Debra O'Connell, Alyssa L. Norris, Abere Sawaqdeh

Department of Psychology, Washington State University, Pullman, WA 99164-4820, United States

HIGHLIGHTS

- A meta-analysis of treatment comparison studies published between 1996 and 2012
- Small, significant differences between treatments for primary outcomes
- No treatment differences when secondary outcomes are assessed.
- CBT is slightly more effective than alternative treatments for primary outcomes.

ARTICLE INFO

Article history:
Received 26 February 2014
Received in revised form 13 August 2014
Accepted 14 August 2014
Available online 23 August 2014

Keywords:
Dodo bird hypothesis
Treatment equivalence
Psychotherapy research
Meta-analysis
Empirically supported treatments

ABSTRACT

The Dodo bird hypothesis asserts that when bona fide treatments are compared they yield similar outcomes and this hypothesis is consistent with a common factors or contextual model of psychotherapy. Wampold et al. (1997), the most recent comprehensive meta-analysis to test the Dodo bird hypothesis, yielded consistent evidence of treatment equivalence. However, some of Wampold et al.'s analytic strategies, such as using multiple effect sizes from the same study and prioritizing long-term follow-up, may have obscured treatment differences. The current meta-analysis updated Wampold et al. by analyzing studies published in the subsequent 16 years ($k = 51$). Separate effect sizes were calculated for primary outcomes versus secondary outcomes, at termination and follow-up. Contrary to the Dodo bird hypothesis, there was evidence of treatment differences for primary outcomes at termination. Furthermore, cognitive-behavioral treatments may be incrementally more effective than alternative treatments for primary outcomes. Consistent with the Dodo bird hypothesis, there was little evidence of treatment differences for the secondary outcomes at termination and follow-up. There are small, statistically significant differences between bona-fide treatments when the specific targets of those treatments are assessed, but not when secondary outcomes are assessed, providing mixed support for both specific factors and contextual models of psychotherapy.

Contents

* Corresponding author. Tel.:+1 509 335 7750; fax: +1 509 335 5043.
E-mail address: david.marcus@wsu.edu (D.K. Marcus).

http://dx.doi.org/10.1016/j.cpr.2014.08.001
0272-7358/

1. Introduction

The Dodo bird hypothesis asserts that when bona fide treatments are compared they yield roughly equal outcomes. Meta-analyses of treatment comparison studies have generally yielded findings consistent with the Dodo bird hypothesis (e.g., Smith & Glass, 1977), but the most recent comprehensive meta-analysis supporting the Dodo bird hypothesis was conducted 17 years ago (Wampold, Mondin, Moody, Stich, et al., 1997). A meta-meta-analysis that synthesized the findings from these treatment comparison meta-analyses also mostly supported the Dodo bird hypothesis (Luborsky et al., 2002). Despite the meta-analytic support for the Dodo bird hypothesis, many psychotherapy researchers have questioned these conclusions on methodological (e.g., Crits-Christoph, 1997; Hunsley & Di Giulio, 2002) and conceptual grounds (e.g., Chambless, 2002). The Dodo bird debate addresses an essential question about *how* psychotherapy works, which may account for why there are such "diametrically opposed interpretations" of this literature (Barlow, 2002, p. 1).

Meta-analyses of studies comparing psychotherapy to various control conditions have demonstrated that overall, psychotherapy is more effective than no treatment or placebo controls (Lambert & Bergin, 1994; Shadish, Matt, Navarro, & Phillips, 2000). As Goldfried (2013) noted, "Six decades after Eysenck (1952) pointed out that there was no good evidence that psychotherapy had a positive impact on patients' lives, there has been the accumulation of a truly impressive body of research evidence to indicate that it indeed does" (p. 865). Reviews and meta-analyses of the treatment outcome literature become more controversial when the question shifts from *whether* to *how* psychotherapy is effective, or if some psychotherapies are more effective than other psychotherapies. One way to address the question of how psychotherapy works is through studies that directly compare two or more treatments to each other. If, for example, Treatment A is more effective than Treatment B, this finding would suggest that there is something specific to Treatment A that contributes to the outcome. Such a finding would support a medical model of psychotherapy, in which specific therapeutic techniques derived from a basic understanding of the condition being treated are responsible for a treatment's efficacy. For example, Barlow's (2004b) concept of "psychological treatments" would be consistent with this view of how psychotherapy functions. On the other hand, if Treatments A and B yield similar outcomes, it is possible that different mechanisms lead to similar improvements, or that it is the factors common to both treatments that result in the improvement (Imel & Wampold, 2008). If most treatment comparison studies yield small differences, a common factors explanation is more parsimonious and more plausible than there being many different mechanisms that all happen to yield similar outcomes. Although different treatments are likely to be grounded in different theoretical models and to use different techniques, from a common factors or contextual model perspective (Wampold, 2007), these models and techniques are primarily important, not as specific mechanisms of change, but because they provide a plausible rationale for the therapy to the client and the therapist (Frank & Frank, 1991; Imel & Wampold, 2008). Thus, if different treatments and techniques are similarly plausible they may also be similarly effective.

In their original Dodo bird literature review, Luborsky, Singer, and Luborsky (1975) counted the outcomes from studies that compared different schools of therapy to one another (e.g., client centered therapy to other psychotherapies, behavior therapy to other psychotherapy) and concluded that there was little evidence that one school of therapy was consistently superior to another. Subsequent Dodo bird meta-analyses generally followed one of two strategies. Some meta-analyses computed the average effect sizes for each school of therapy compared to control groups and then compared these effect sizes (Smith & Glass, 1977). The limitations of this approach have been well-documented (Shadish & Sweeney, 1991; Wampold, Mondin, Moody, Stich, et al., 1997) and lie in the potential confounds that are likely to occur when comparing studies that, for example, treated patients drawn from different populations, used therapists with differing qualifications and experience, and used different outcome measures. Other meta-analyses (e.g., Robinson, Berman, & Neimeyer, 1990) included analyses of studies that directly compared one treatment to another, but again aggregated the studies by school of therapy (e.g., cognitive versus general verbal), which raises other problems. Very different treatments may be combined into the same general type. Some examples include the following: (a) Shapiro and Shapiro (1982) combined psychodynamic and humanistic psychotherapy into the same treatment type, (b) Robinson et al. (1990) created a general verbal therapy category, and (c) there are a variety of different therapies that all fall within the behavioral treatment type such as exposure, relaxation, and social skills training, each of which involves different mechanisms and may not be equally effective. Furthermore, different techniques may be differentially effective for different disorders. In summary, these early Dodo bird studies are in many ways equivalent to conducting a meta-analysis to answer the question "which is more effective, antidepressants or antibiotics?"

Aside from possible partisan allegiance to a particular theoretical orientation or curiosity about whether one school of therapy is more effective than another, there was a practical reason why these early Dodo bird studies classed treatments into types for their meta-analyses. The researchers needed a way to organize the columns of results or, in other words, a direction for their effect sizes. For example, for studies comparing behavioral therapy to psychodynamic therapy, column A could be behavioral therapy, column B could be psychodynamic therapy, and a positive effect size would then indicate that behavioral therapy was superior to psychodynamic therapy. Without organizing the treatments into categories, the specific treatment assigned to each column in the meta-analysis would be arbitrary and the average effect size would approximate zero. The alternative of placing the more effective treatment in column A and the less effective treatment in column B is not viable because it would capitalize on chance and overestimate the magnitude of treatment differences.

Wampold, Mondin, Moody, Stich, et al. (1997) solved the problem of how to test the Dodo bird hypothesis without aggregating treatments into particular types or schools. In doing so they addressed the more basic question of whether comparisons of bona fide therapies yield differences in effectiveness regardless of which types of treatments are compared, and not the question of whether one school of psychotherapy is superior to another. The logic of the Wampold et al. meta-analysis bears some discussion, because it has seemingly been misunderstood by a variety of commentators (e.g., Hunsley & Di Giulio, 2002). Suppose that there was a Dodo bird hypothesis in pharmacology, with some researchers asserting that all medications yield equivalent outcomes due to common factors (e.g., the act of swallowing a pill, the medication was prescribed by a physician with professional expertise, the patient's positive expectation). Furthermore, suppose that there was a body of

research comparing various medications to one another for various conditions (e.g., an antidepressant to an antihistamine for treating seasonal allergies, an antibiotic to a chemotherapy drug for treating strep throat). With no a priori method for deciding which drug treatment should be subtracted from which drug treatment, the average outcome aggregated across all of the studies would approximate zero, but consider the distribution of the scores. Some comparisons would yield effect sizes close to zero, maybe because both treatments were ineffective (e.g., an antibiotic to an antidepressant for treating lung cancer) or because they were similarly effective (e.g., penicillin versus a sulfa drug for treating a bacterial infection). Yet, due to sampling and measurement error, unless the sample sizes were extremely large, even these studies would rarely yield effect sizes of exactly zero. On the other hand, there would also be a number of very large effect sizes (positive or negative depending on what was subtracted from what) when an effective treatment was compared to an ineffective or iatrogenic treatment (e.g., an antibiotic versus a chemotherapy drug for treating lung cancer, an antibiotic versus a chemotherapy drug for treating tuberculosis). If the heterogeneity of the effect sizes is more varied than would be expected simply due to chance/error, then the pharmaceutical Dodo bird hypothesis could be rejected (even with an average effect size across studies of zero). Thus, such an analysis would not address the prima facie absurd question of whether antibiotics are better than antihistamines, but would instead be a true test of the Dodo bird hypothesis by addressing the question of whether it generally matters which particular drug is prescribed, or, in the case of psychotherapy, which treatment is used. In other words, there would be no need for antibiotics to always perform better than other drugs for all conditions (or behavioral treatments to always be better than psychodynamic treatments) to reject the Dodo bird hypothesis, there would simply need to be meaningful heterogeneity among the studies.

Wampold, Mondin, Moody, Stich, et al.'s (1997) innovation was to use the *Q* test of homogeneity to test the Dodo bird hypothesis. For each study, each treatment was randomly assigned to column A and column B (yielding an average effect size of approximately zero), but the test of the Dodo bird hypothesis was the *Q* statistic. A statistically significant *Q* indicates that the studies are heterogeneous, which would be counter to the Dodo bird hypothesis. In contrast, a small and nonsignificant *Q* indicates that the treatment comparison studies yielded similar outcomes that clustered around zero and thus any treatment differences found in particular studies were likely due to chance. Wampold et al. computed four different *Q* statistics using different combinations of effect sizes. All of them indicated very small amounts of heterogeneity (the largest being 13%) and none of them were statistically significant.

Although the use of the *Q* statistic was an important innovation, some of Wampold et al.'s other analytic strategies were questionable and may have minimized actual treatment differences. Many of the analyses Wampold et al. ran included multiple effect sizes gleaned from the same sample, violating the meta-analytic assumption of independence (and potentially reducing the overall amount of heterogeneity). Furthermore, Crits-Christoph (1997) questioned the inclusion of treatment studies that used college student volunteers, which may have limited external validity. He also questioned Wampold et al.'s decision to combine follow-up and termination assessments (or give priority to follow-up assessments) when computing effect sizes because patients may receive additional therapy during the follow-up period, attenuating genuine treatment differences. Similarly, Crits-Cristoph criticized Wampold et al.'s strategy of including both primary (i.e., measures of the symptoms/problems being treated) and secondary (e.g., quality of life measures) outcome measures when computing effect sizes, noting that treatments that reduce the targeted symptoms may be clinically superior to those that do not, even if both treatments yield similar outcomes on quality of life measures. Finally, Crits-Cristoph suggested that when the studies in the Wampold et al. meta-analysis that compared cognitive-behavioral therapy (CBT) to non-CBT using actual patients (not college student volunteers) were examined, there was evidence that CBT was superior to alternative treatments. On this last point, Wampold, Mondin, Moody, and Ahn (1997) countered that this "cognitive contrast" (p. 227) involved cherry picking studies and variables that were most favorable to CBT, likely resulting in Type I error.

Whereas the question of treatment equivalence and specific mechanisms for pharmaceuticals is unequivocally answered (if it was even ever asked), the question of treatment equivalence in psychotherapy is no more settled now than it was when Luborsky et al. (1975) published their Dodo bird paper. If anything, this controversy has grown more intense with the report by Task Force on Promotion and Dissemination of Psychological Procedures (1995) that initiated the empirically supported treatment (EST) movement. The report included a list of empirically validated treatments. The Task Force recommended that these treatments be taught in graduate programs and internships, and that policy makers, third party payers, and the general public be informed about these treatments. The Task Force and subsequent publications (e.g., Chambless & Ollendick, 2001) were clear that that the aim of the list was not to compare treatments or determine the best treatment for particular disorders (e.g., a number of therapies have been listed as ESTs for major depression), but to ensure that clinical psychologists use ESTs in their practice. However, both advocates and critics of the EST movement seem to agree that if the Dodo bird hypothesis is correct and all bona fide yield similar outcomes, then "the identification of ESTs would be a much less important exercise" (Chambless & Ollendick, 2001, p. 704) and the appropriate response would be to "cease the unwarranted emphasis on ESTs" (Messer & Wampold, 2002, p. 24). Thus, compelling evidence either supporting or refuting the Dodo bird hypothesis would have implications for psychotherapy research, including whether research efforts should be directed toward discovering and validating specific techniques, or whether factors common to most therapies, such as the therapeutic alliance and the installation of hope, should be the primary focus of psychotherapy research. Such findings also have implications for practice, including the extent to which therapists should rely on (or even be compelled to use) treatment manuals in their practices. In a paper critical of the Dodo bird hypothesis, Rounsaville and Carroll (2002) argued that if the Dodo bird hypothesis is correct then "psychotherapy should be delivered by the least highly trained lowest paid practitioners" (p. 17) and that clinical training programs should put a greater emphasis on recruiting empathic students who are skilled at engaging people, and less of an emphasis on intellectual aptitude and academic achievements.

Wampold, Mondin, Moody, Stich, et al.'s (1997) meta-analysis was published just two years after the first Task Force report, and therefore did not include studies that were conducted subsequent to this report. By developing criteria for what qualifies as an EST and updating the list on a regular basis (e.g., Chambless et al., 1998), the Task Force raised the stakes for documenting empirical support for specific treatments, and may have influenced the ways in which treatment comparisons were conducted. For example, the EST movement may have encouraged research on under-studied treatments (i.e., "unvalidated" treatments), so that commonly used treatments that had not been subjected to outcome studies could qualify as ESTs. Also, by requiring that different research teams conduct studies supporting the treatment, the Task Force may have increased the variety of researchers studying particular therapies, encouraging studies by researchers who were not also the developers of the treatment.

Subsequent to the Wampold, Mondin, Moody, Stich, et al. (1997) meta-analysis, there have also been some noteworthy changes in the requirements that many journals have for conducting and reporting the findings from treatment outcome studies. Specifically, most journals now require a flowchart adapted from the CONSORT Group (Altman et al., 2001) that details the number of participants at each stage of the study (e.g., number screened, number randomized). The American Psychological Association journals also require that clinical trials meet

the Journal Article Reporting Standards criteria (JARS; APA Publications & Communications Board Working Group on Journal Article Reporting Standards, 2008). Both the EST movement and the changes in reporting standards may have influenced the findings from treatment comparison studies, and thus the Dodo bird verdict. For example, the Task Force criterion stipulating that evidence for an EST can include outcome studies that show that the treatment is equivalent to an established treatment may lead to the publication of more studies with null findings that fail to find differences between treatments. The CONSORT and JARS criteria may also make it more difficult for researchers to manipulate results to accentuate treatment differences. For example, studies should include intent-to-treat analyses and not just completer analyses, and researchers are expected to report the results from all of the outcome measures they used, not just those that yielded statistically significant results. By updating the Wampold et al. meta-analysis, it is possible to see if these changes in the field of psychotherapy research are associated with changes in the treatment comparison literature.

1.1. The cognitive contrast

Closely connected with both the Dodo bird hypothesis and the EST debate are the questions of whether CBT is superior to other treatments (Crits-Christoph, 1997), and whether the EST movement is biased toward CBT treatments (Task Force on Promotion & Dissemination of Psychological Procedures, 1995). There may be greater justification for aggregating treatments into a CBT category than some of the aggregations that were performed in earlier meta-analyses (e.g., combining psychodynamic and humanistic therapies). All CBT treatments share basic assumptions about the inter-relations between thoughts, behaviors, and pathology. They all include cognitive restructuring as a central treatment component and most also include psycho-education and behavioral interventions.

Tolin (2010) revisited the cognitive contrast hypothesis in a meta-analysis of 26 studies that directly compared CBT to other bona fide treatments. Overall, he found evidence that CBT was superior to other bona fide treatments at reducing primary symptoms at both termination ($d = .22$) and at six-month follow-up ($d = .47$). More specifically, across disorders CBT yielded better outcomes than psychodynamic therapies ($d = .28$ at termination; $d = .50$ at six-months). Across comparison treatments, CBT was superior to other psychotherapies for treating anxiety ($d = .43$) and depression ($d = .21$) at termination. The results were mixed when CBT was compared to other treatments on secondary outcomes. Baardseth et al.'s (2013) re-analysis of the studies in Tolin's meta-analysis yielded similar effect sizes for the primary outcomes, but no differences between CBT and the comparison treatments for the secondary outcomes. In a second meta-analysis, Baardseth et al. examined the results from 13 anxiety disorder treatment studies that compared CBT to another therapy. In this meta-analysis, the effect size for primary outcomes was smaller ($g = .14$) and not statistically significant. Furthermore, CBT treatments yielded almost identical results to non-CBT treatments on the secondary outcome measures ($g = -.03$). Thus, there appears to be mixed empirical support for the relative efficacy of CBT, particularly depending on the outcome measures utilized.

1.2. The present study

The current meta-analysis updated Wampold, Mondin, Moody, Stich, et al. (1997) by analyzing studies that have been published in the subsequent 17 years in the same journals that they reviewed. Like Wampold et al.'s meta-analysis, our primary analysis used the *Q* statistic to assess the degree of heterogeneity across studies that compared bona fide treatments. However, we also attempted to address some of the criticisms of Wampold et al. by (a) analyzing primary and secondary outcomes separately, (b) analyzing the outcomes at termination and follow-up separately, (c) only including studies that used actual patients, and (d) including only one effect size per sample per meta-analysis. Clinicians and researchers may disagree about the relative value of symptom-specific targeted outcomes or more general quality of life measures as well as about whether post-treatment or longer-term outcomes are more important when comparing therapies. Thus, by providing separate meta-analyses for each of these sets of outcomes, readers can draw their own conclusions about treatment equivalence and the Dodo bird hypothesis. Given these potential improvements on Wampold et al.'s pioneering approach to studying the Dodo bird hypothesis, and subsequent developments in the field, including the EST movement, we believe that an updated Dodo bird meta-analysis is warranted and has the potential to inform ongoing research and practice efforts.

Additionally, because a large majority of treatment comparison studies compared CBT to another form of therapy, a secondary aim of the current paper was to examine the cognitive contrast hypothesis using a set of studies that were selected using different search criteria than those used by either Tolin (2010) or Baardseth et al. (2013). Once again, for these analyses, separate meta-analyses were conducted for primary and secondary outcomes and at termination and follow-up. We also attempted to replicate Tolin's moderator analyses, which found that CBT was superior to psychodynamic therapy and was more effective than other therapies for treating anxiety and depression.

2. Method

2.1. Identification of studies

The current meta-analysis generally adopted the same inclusion criteria used by Wampold, Mondin, Moody, Stich, et al. (1997), with one additional criterion. Following Wampold et al., the study had to be published in *Behaviour Research and Therapy*, *Behavior Therapy*, *Journal of Consulting and Clinical Psychology*, *Journal of Counseling Psychology*, *Archives of General Psychiatry*, or *Cognitive Therapy and Research* between 1996 and 2012 (inclusive).[1] The study also had to compare at least two *bona fide* treatments.[2] To be considered a *bona fide* treatment, the treatment must target a clinically-relevant problem or issue and be tailored to the patient. It must also meet two of four conditions: (a) have a treatment manual, (b) cite an established therapeutic approach, (c) include active ingredients that have published citations, or (d) include a description that contains a reference to an established psychological process.

To be included, studies had to randomly assign clients to the treatment conditions. The included studies also had to provide the necessary information to calculate an effect size, ideally, group means and standard deviations. Following Wampold et al., component studies (additive or dismantling) were excluded, as were studies that compared a single active treatment to a treatment that only included common or nonspecific factors (i.e., placebo controls). Wampold et al. also required that the treatment be provided by a therapist with at least a master's degree, but we relaxed this criterion to include therapy provided by graduate students if they were supervised. Finally, in addition to the Wampold et al. criteria, we required that the patients, in addition to the treatment, be *bona fide*. In other words, samples composed of college student volunteers who participated for course credit were excluded, even if the participants reported having actual symptoms. This search yielded 48 articles that described the findings from 51 independent samples that

[1] Two of these articles reported the findings from follow-up or secondary analyses of studies that had been published in other journals during the 1996–2012 review period. In these two instances the findings from the original study were included in the meta-analysis even though they were not published in one of the six selected journals.

[2] None of the included studies had more than two active treatments, although a number of studies also included a placebo control group, which was not included in the analyses.

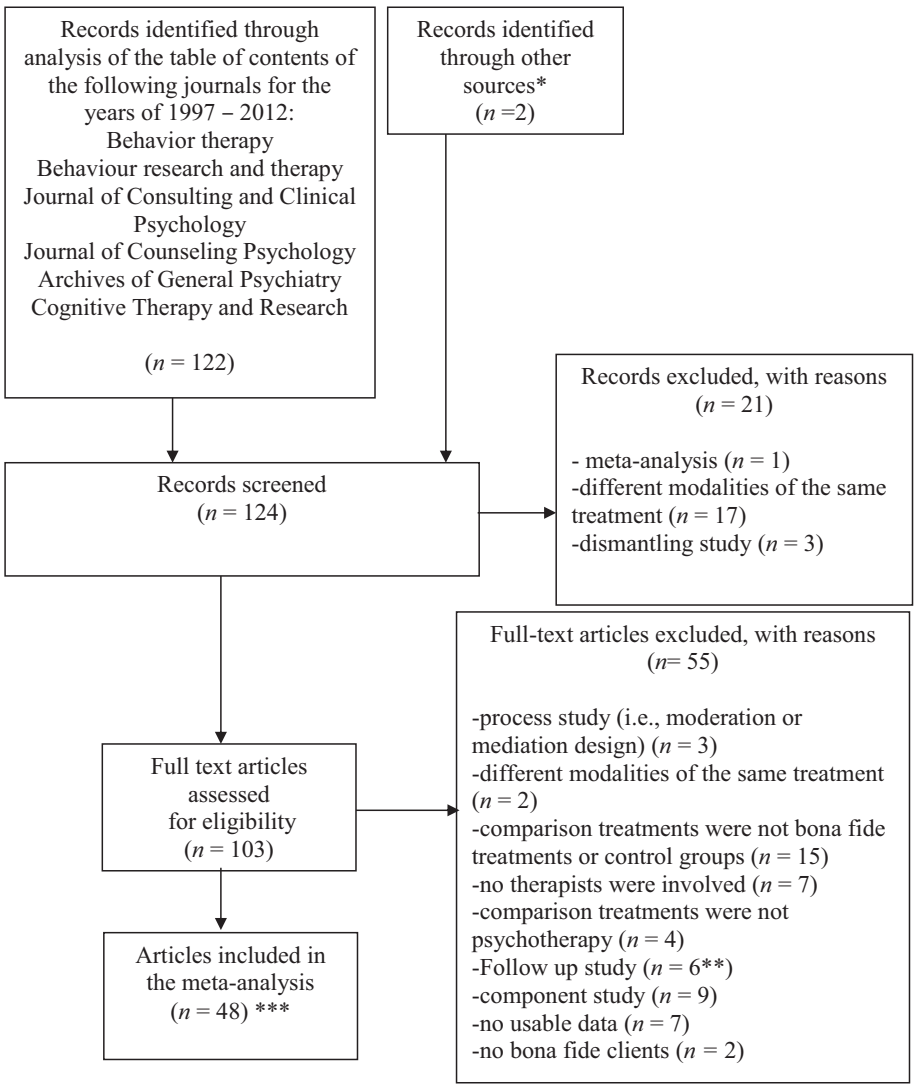

Fig. 1. Flow chart of study selection. *These were articles that reported the original outcome data from studies in the original search that had only reported follow-up data. **Four of the studies reported in these articles were already included in the meta-analysis and the other two were the records identified through other sources **3 articles provided data on two independent samples each, and were coded as two separate studies thereby adding 3 more studies to the analyses ($k = 51$).

included 8789 participants (Fig. 1).[3] The studies included in the meta-analysis are summarized in Appendix A.

2.2. Coding and analyses

The effect sizes were computed as Cohen's *d* (the mean of treatment group A minus the mean of treatment group B divided by the pooled standard deviation). To combine the results across studies, the effect sizes were weighted by the inverse variance so that studies with larger samples had greater weight when computing the average *d*. We used a random effects model to combine the effect sizes across studies, because, unlike a fixed effect model, a random effects model does not assume that variation across studies is only due to chance. A random effects model assumes that the included studies are only a subsample of all of the possible studies that could be conducted on the topic and allows researchers to generalize beyond the specific studies included in the meta-analysis (Field, 2003). Following the logic of Wampold, Mondin, Moody, Stich, et al. (1997), the treatments were randomly assigned as treatment A or B for each study, resulting in average effect sizes approximating zero. Outcome variables were classified as primary if they measured the symptoms or problems that were the direct target of the treatment (e.g., a depression measure for a depression treatment study). All other outcome variables were classified as secondary, including measures of symptoms other than those targeted by the treatment (e.g., a depression measure in a panic disorder treatment study), general quality of life measures, and measures of constructs associated with or believed to contribute to the etiology of the problem being targeted (e.g., a perfectionism measure in a bulimia treatment study). In order not to violate the independence assumption, when a study provided multiple primary (or secondary) outcome measures (including both self-report and clinician ratings[4]), an effect size was calculated for each measure and then they were averaged by adding the effect sizes and dividing by the number of outcome measures.

To calculate inter-rater reliability, a random sample of 15 of the 48 articles (79 outcome measures) were independently coded by all four co-authors. For 71 of the 79 (87.7%) measures there was complete

[3] One study provided follow-up data, but did not provide outcome data at termination.

[4] We also conducted separate supplementary analyses for the self-report and clinician rated outcomes.

agreement about whether it was a primary or secondary outcome, in four instances (4.9%) three of the four co-authors agreed, and in four cases (4.9%) two raters rated the measure as a primary outcome and two rated it as a secondary outcome. In the eight instances when there was not unanimous independent agreement, the co-authors discussed the measure until unanimous agreement was reached. Disagreements typically occurred when it was unclear whether the measure was diagnostic of the problem being treated. For example, a study on the treatment of binge eating disorder (Munsch, Meyer, & Biedert, 2012) included body mass index (BMI) as an outcome variable. Two of the coders initially rated BMI as a primary outcome and two rated it as a secondary outcome. Because binge eating disorder can occur in individuals with normal weight and BMI is not a symptom of this disorder (American Psychiatric Association, 2013), after discussion we classified BMI as a secondary measure, even though the study authors had classified it as a primary measure.

To test the Dodo bird hypothesis and determine whether effect sizes were from a single population, a *Q* test was calculated. A statistically significant *Q* value indicates that the effect sizes were heterogeneous. I^2 "describes the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins, Thompson, Deeks, & Altman, 2003, p. 558). An I^2 value of 0 indicates no heterogeneity. I^2 values of 25% or less are considered low levels of heterogeneity, and values of 50% are considered moderate (Higgins et al., 2003). In the current meta-analysis, the I^2 value indicates the extent to which differences in outcome reflect differences between the treatments that were compared (not just sampling error). Thus, the larger the value of I^2, the greater the heterogeneity among the effect sizes, and the less support there is for the Dodo bird hypothesis. All of the analyses were conducted using Lipsey and Wilson's (2001) SPSS statistical programs and Viechtbauer's (2010) "metafor" package for R.

Effect sizes and tests of homogeneity were computed at pre-treatment, termination, and follow-up for both primary and secondary measures, resulting in six separate meta-analyses. Based on random assignment, the pre-treatment effect sizes should be homogeneous, and a statistically significant *Q* statistic for pre-treatment data would suggest that there was either something wrong with the data set or with the logic of using the *Q* statistic to test the Dodo bird hypothesis. In other words, at pretreatment, before any interventions were implemented, the effect sizes should cluster around zero and there should not be statistically significant heterogeneity across studies (i.e., differences between treatment groups should be entirely due to sampling error). Because six months was the modal follow-up time (24 of the 39 studies that reported follow-up data reported a six-month follow-up), when studies reported several follow-up data points, the point closest to six months was used to compute the follow-up effect size.

Wampold, Mondin, Moody, Stich, et al. (1997) also calculated an upper bound for treatment differences by examining the absolute values of the effect sizes from each treatment comparison study, which yielded an average effect of .21. The problem they correctly identified with this approach is that it capitalizes on chance and overestimates the average difference between two treatments. The pretreatment effect sizes calculated in the current study, which, based on random assignment, should approximate zero, can provide a useful point of comparison. A meta-analysis of the absolute values of these pre-treatment effects was compared to the absolute values of the termination and follow-up effect sizes. Furthermore, the standard deviation of the pre-treatment effect sizes was used to generate Monte Carlo data to approximate how likely the absolute values of the termination and follow-up effect sizes would occur simply due to chance.

Additional cognitive contrast meta-analyses were conducted comparing cognitive or CBT treatments to other treatments. Unlike the Dodo bird analyses, these meta-analyses followed standard meta-analytic procedures because treatment A was always CBT and treatment B was the alternative treatment. Treatments were classified as CBT if the original study authors identified them as cognitive therapy or CBT. All other therapies (including behavioral therapy) were classified as other treatments. All four co-authors independently coded the articles for whether the study compared CBT to a non-CBT treatment. There was unanimous agreement for 45 of the 48 articles (93.75%) and the remaining three articles were discussed until there was unanimous agreement. For the 38 articles describing studies that compared CBT to a non-CBT treatment, 20 of the articles referred to the treatment as CBT, 14 as cognitive therapy, two as cognitive restructuring, one as schema-focused therapy, and one as trauma-focused therapy. The CBT provided in all 38 of these articles included cognitive restructuring. Additionally, 35 of the articles explicitly noted that psycho-education was a component of the CBT treatment (the other three may have included psycho-education, but failed to mention it). Finally, 34 of the articles included a behavioral component as part of the treatment, most typically behavioral experiments to test alternative beliefs, exposure exercises, or skills training. The four studies that did not include a behavioral component compared cognitive therapy to a behavioral treatment and intentionally excluded any explicitly behavioral interventions. Therefore all of these treatments shared cognitive restructuring (and most likely psycho-education) as a central component of therapy in common.

Effect sizes were computed so that a positive *d* indicated that CBT yielded a better outcome. To assess possible publication bias, we graphed funnel plots with the effect sizes plotted on the x-axis and standard errors on the y-axis (with the largest values at the bottom). We then used a regression test (Egger, Smith, Schneider, & Minder, 1997) to examine the funnel plot asymmetry, which could indicate publication bias. Finally, we used the trim and fill procedure (Duval & Tweedie, 2000) to impute presumed missing values to make the funnel plot symmetrical and recalculate the effect size adding these imputed values. A large change in the recalculated effect size (i.e., 95% confidence intervals that no longer overlap) suggests the possibility of publication bias.

3. Results

3.1. Description of the included studies

The median number of sessions provided was 12, ranging from one study that examined a single-session-treatment to seven studies that examined treatments ranging from 20 to 35 sessions (and one study with 312 sessions; Giesen-Bloo et al., 2006). The treatment sessions ranged in length from 15 to 180 min ($M = 79.1$, $SD = 38.6$). Most ($k = 47$) of the studies treated adult samples, with three studies treating children or adolescents, and one study including both adults and children. Six of the studies included exclusively female participants, two studies were exclusively male, and the remainder ($k = 43$) were mixed. Across studies, the samples were 66% female. Most of the studies provided individual therapy ($k = 37$), ten provided group therapy, and four provided both individual and group therapy. The most commonly treated disorders were eating disorders ($k = 9$), depression ($k = 6$), and post-traumatic stress disorder ($k = 7$). Seventeen studies treated other anxiety and anxiety-related disorders, including generalized anxiety disorder ($k = 4$), panic disorder with or without agoraphobia ($k = 4$), social anxiety ($k = 4$), specific phobia ($k = 2$), obsessive–compulsive disorder ($k = 2$), and unspecified anxiety ($k = 1$). The remaining 12 studies treated a variety of disorders and problems (Appendix A).

Most of the studies compared CBT to some other therapy ($k = 41$), most frequently to a type of behavior therapy ($k = 21$). Six studies compared CBT to interpersonal therapy (IPT), five compared CBT to acceptance and commitment therapy (ACT), three studies compared CBT to psychodynamic therapy, and six studies compared CBT to other treatments [present-centered therapy (2 studies), self-system therapy, stress inoculation/stress management therapy, problem solving therapy, and 12-step treatment]. Eight of the remaining studies compared two non-CBT treatments to one another, and two compared one form of CBT to another form of CBT.

3.2. Homogeneity of effect sizes

Because the signs for each treatment were randomly assigned, the average effect sizes for the six meta-analyses were all close to zero, ranging from −.016 to .044 (Table 1). As expected, at pre-treatment the Q statistics for both the primary [$Q\,(48) = 27.11$, $p > .99$, $I^2 = 0\%$] and the secondary [$Q\,(37) = 11.45$, $p > .99$, $I^2 = 0\%$] outcome measures were not statistically significant. In fact, given that the Q values were smaller than the degrees of freedom, the effect sizes were entirely homogeneous (i.e., $I^2 = 0\%$). In contrast there was a statistically significant and moderate level of heterogeneity at termination for the primary outcomes, $Q\,(49) = 107.48$, $p < .001$, $I^2 = 54.41\%$. There was, however, a very small, non-significant amount of heterogeneity for the secondary outcomes at termination, $Q\,(37) = 43.85$, $p = .20$, $I^2 = 13.34\%$. The same pattern emerged for the outcomes at follow-up, except that there appeared to be less heterogeneity than at termination (and there were fewer studies that reported follow-up outcomes). The primary outcomes yielded a small-to-moderate, but statistically significant level of heterogeneity $Q\,(38) = 63.02$, $p = .007$, $I^2 = 39.70\%$. The secondary outcomes were homogeneous $Q\,(31) = 31.09$, $p = .46$, $I^2 = 0.28\%$.

Although the statistically significant heterogeneity for the primary measures at termination appears to counter the Dodo bird hypothesis, this moderate amount of heterogeneity may have been due to a few studies that yielded large group differences. We used the "leave1out" function in the metafor program (Viechtbauer, 2010), which reruns the meta-analysis omitting one study at a time. The Q statistics (all with 48 df) ranged from to 95.02 to 107.48 (all $ps < .001$), so no single study accounted for the heterogeneity in this meta-analysis. In contrast, one study may have contributed to the statistically significant heterogeneity at follow-up for the primary outcomes. Clark et al. (2006) yielded the largest treatment difference in primary outcome at follow-up ($d = 1.14$). If this study is omitted from the meta-analysis, the Q statistic becomes marginal at follow-up, $Q\,(37) = 51.87$, $p = .053$, $I^2 = 28.67\%$. The Q statistic remained statistically significant if any of the other studies were omitted, with values ranging from 55.17 ($p = .03$) to 63.02 ($p = .005$).

Finally, we reran the analyses separately for the self-report and clinician rating based outcomes, with largely similar results. For the primary outcomes, at termination both the self-reported outcomes, $Q\,(33) = 62.25$, $p = .0015$, $I^2 = 46.99\%$ and the clinician-rated outcomes, $Q\,(31) = 71.59$, $p < .001$, $I^2 = 56.70\%$, were heterogeneous. At follow-up, the self-reported primary outcomes remained heterogeneous, $Q\,(25) = 52.96$, $p = .0009$, $I^2 = 52.79\%$, but the clinician-rated outcomes were no longer heterogeneous, $Q\,(23) = 26.66$, $p = .27$, $I^2 = 13.72\%$. The follow-up data from the Clark et al. (2006) study that was an outlier was based on self-reports, which explains why the meta-analysis of the self-report data at follow-up was heterogeneous, whereas the follow-up of the clinician ratings was not. The self-reported secondary outcomes remained homogeneous at both termination and follow-up and there were not enough studies ($k = 8$) that provided clinician ratings of secondary outcomes to warrant analyses.[5]

Table 1
Results of the Dodo bird meta-analyses.

	k	d	se	Q	I^2	|d|
Pre-treatment						
Primary	49	−.016	.027	27.11	0%	.094
Secondary	38	−.011	.038	11.45	0%	.092
Termination						
Primary	50	−.015	.053	107.48***	54.41%	.288
Secondary	38	.005	.043	43.85	13.34%	.188
Follow-up						
Primary	39	.044	.048	63.02**	39.70%	.168
Secondary	32	.020	.043	31.09	0.28%	.179

** $p < .01$.
*** $p < .0001$.

3.3. Upper bound analyses

To assess the upper bound of the treatment differences, the absolute values of the effect sizes were aggregated. This method capitalizes on chance and overestimates the true differences between treatments, which can be seen by examining the average of these effect sizes for the pre-treatment data. Because all of the studies used random assignment, the pre-treatment group differences should roughly equal zero, but the average of the absolute values was .09 for both the primary and secondary outcomes. The average of the absolute values for the termination effect sizes (.29 for the primary outcomes and .19 for the secondary outcomes) and follow-up effect sizes (.17 for the primary outcomes and .18 for the secondary outcomes) appeared larger.

Because the absolute values of the post-treatment and follow-up effect sizes were by definition greater than zero, it did not make sense to subject them to standard meta-analytic significance tests. Instead, we used the pre-treatment data to generate Monte Carlo data. The weighted standard deviation of the effect sizes for the primary measures at pre-treatment (when the signs were randomly assigned) was .14. We created 50,000 simulated data sets with a mean of zero (the presumed mean when signs are randomly assigned), a standard deviation of .14, and a sample size of 50 (the number of studies that reported primary measure outcomes at termination). We next calculated the means of the absolute values of the numbers in each data set. The grand mean of the 50,000 mean scores in this Monte Carlo data set was .11, which should therefore approximate the value that would occur simply due to chance when averaging the absolute values of the differences between pairs of studies. Only .001% of the data sets had mean values greater than .16, so the average of the absolute values of .29 for the effect sizes for the primary measures at termination was statistically significant, $p < .001$. Substituting a sample size of 39 (the number of studies that reported primary outcome data at follow-up), the 99.99th percentile was .165, so the average absolute value at follow-up (.168) was also significantly greater than zero ($p < .001$). If the outlier study is omitted, the average effect size becomes .138, which just exceeds the critical value for the 95th percentile (.135). For the secondary outcome data, the same Monte Carlo method was used, substituting the weighted standard deviation of the secondary outcome measures at pre-treatment (.128) and the corresponding sample sizes. The absolute value *d*s for the secondary measures at termination (.19) and at follow-up (.18) were both statistically significant ($p < .001$).

3.4. Moderator analyses

Because there was statistically significant heterogeneity among the effect sizes for the primary measures at termination, we conducted meta-analytic equivalents to the regression (Hedges & Olkin, 1985) to test potential continuous moderators or Hedges (1982) meta-analytic analogue to the analysis of variance (ANOVA) to test categorical moderators. These analyses were conducted using the absolute values of the effect sizes to examine whether any of these moderator variables were associated with larger treatment differences. Neither sample size [$\beta = -.17$; $B = -.0004$ ($SE = .0004$); $Z = -1.02$, $p = .31$], year of publication [$\beta = .063$; $B = .0025$ ($SE = .0073$); $Z = .34$, $p = .74$], percent of the sample that was female [$\beta = .15$; $B = .111$ ($SE = .121$); $Z = .91$, $p = .36$], nor whether the treatment was administered in an individual or group format [$Q_B\,(1) = 1.09$, $p = .30$; $Q_W\,(45) = 31.85$, $p = .93$] was associated with larger treatment differences for the primary outcome measures at termination. There was a trend toward studies with younger samples yielding larger treatment differences for the

[5] A table with the effect sizes for each study separated by self-report versus clinician ratings is available from the first author.

primary outcome measures at termination [$\beta = -.32$; $B = -.007$ ($SE = .0039$); $Z = -1.85$, $p = .06$].

3.5. The cognitive contrast

The two studies that compared one form of CBT to another form of CBT were excluded from these analyses. For the primary outcome measures at termination, the average *d* across the 40 studies that compared CBT to another treatment was .16, which was small, but statistically significant, 95% *CI* = .07–.26; $Z = 3.38$, $p = .0007$. There was also a small, but statistically significant amount of heterogeneity among these studies, $Q\ (39) = 54.87$, $p = .047$, $I^2 = 28.92\%$. There was little evidence that these results were influenced by publication bias. The regression test for funnel plot asymmetry was not statistically significant ($z = -1.18$, $p = .24$). The trim and fill procedure did impute four missing studies, but they were on the right side of the graph and adding these data points did not statistically increase the estimated effect size ($d = .21$; *CI* = .11–.31). Finally, the "leave1out" procedure indicated that no single study disproportionately influenced the results, with *d*s ranging from .15 to .18 (all *p*s < .003) when individual studies were omitted from the meta-analysis.

The superiority of CBT for the primary measures was maintained at follow-up with an average *d* of .16 across the 31 studies that provided follow-up data, 95% *CI* = .06–.26; $Z = 3.09$, $p = .002$. There was also statistically significant heterogeneity among these studies, $Q\ (30) = 46.03$, $p = .03$, $I^2 = 34.83\%$. The regression test for funnel plot asymmetry was not statistically significant ($z = 1.15$, $p = .25$) and the trim and fill procedure indicated that there were no missing studies. Rerunning the analysis leaving out a study at a time had little effect on the results with *d*s ranging from .13 to .19 (all *p*s < .006)

In contrast, for the secondary outcome measures at termination, the average *d* across the 32 studies that compared CBT to another treatment was only .07, which was not statistically significant, 95% *CI* = −.01–.16; $Z = 1.64$, $p = .10$. There was little heterogeneity among these studies, $Q\ (31) = 33.81$, $p = .33$, $I^2 = 8.32\%$. The regression test for funnel plot asymmetry was not statistically significant ($z = -.08$, $p = .93$). The trim and fill procedure imputed no missing studies. At follow-up, the results for the secondary outcome measures were almost identical, $d = .08$ ($k = 26$; 95% *CI* = −.01–.18; $Z = 1.72$, $p = .08$). There was no heterogeneity among these studies, $Q\ (25) = 23.13$, $p = .57$, $I^2 = 0\%$. The regression test for funnel plot asymmetry was not statistically significant ($z = -.21$, $p = .83$) and the trim and fill procedure indicated that there were no missing studies.

Given that the primary outcomes were heterogeneous at both termination and follow-up, it was possible to examine whether CBT's superiority varied depending on the treatment to which it was compared and whether it varied depending on the problem being treated. At termination, the omnibus meta-analytic analogue to the ANOVA (Hedges, 1982) for the type of comparison treatment was not statistically significant, $Q_B\ (4) = 6.29$, $p = .18$; $Q_W\ (35) = 48.57$, $p = .06$. Whereas CBT was statistically superior to IPT, and psychodynamic therapy, it was not statistically more effective than behavioral therapies, ACT, or various miscellaneous therapies (see Table 2 for effect sizes). At follow-up, the omnibus meta-analytic analogue to the ANOVA for the type of comparison treatment was statistically significant, $Q_B\ (4) = 14.87$, $p = .005$; $Q_W\ (26) = 31.16$, $p = .22$. However, CBT was only statistically superior to psychodynamic therapy. CBT was not statistically more effective than IPT, behavioral therapies, ACT, or miscellaneous therapies (Table 2).

At termination, the omnibus meta-analytic analogue to the ANOVA for the type of problem treated was not statistically significant, $Q_B\ (3) = 5.39$, $p = .15$; $Q_W\ (36) = 46.26$, $p = .12$. However, whereas CBT was statistically superior to other treatments for treating anxiety, eating disorders, and various miscellaneous disorders, it was not more effective for treating depression (Table 2). The omnibus meta-analytic analogue to the ANOVA for the type of problem treated was not statistically significant at follow-up, $Q_B\ (3) = 3.78$, $p = .29$; $Q_W\ (27) = 29.73$, $p = .33$. At follow-up, CBT was statistically superior for treating depression, but not anxiety, eating disorders, or miscellaneous disorders (Table 2). In the case of the two studies that provided both termination and follow-up data for depression, there was no difference between CBT and non-CBT treatments at termination ($d = -.04$, $p = .86$) even though CBT was superior at follow-up.

Table 2
Moderator analyses for the cognitive contrast for primary outcome measures.

	Termination			Follow-up		
	k	*d*	*p*	*k*	*d*	*p*
Alternative treatment						
IPT	6	.23	.006	5	.15	.12
Psychodynamic	3	.38	.0003	2	.46	.0001
Behavioral	21	.11	.11	17	.13	.07
ACT	5	.14	.20	2	.10	.55
Other	5	.10	.26	5	.004	.92
Problem type						
Anxiety	20	.12	.04	18	.12	.07
Eating disorders	6	.29	.002	5	.18	.11
Depression	6	−.006	>.99	2	.50	.01
Other	8	.25	.003	6	.11	.18

Note. For each effect size, cognitive behavioral therapy (CBT) is being compared to the alternative so that a positive *d* value indicates that CBT is superior to the alternative treatments. IPT = Interpersonal Therapy; ACT = acceptance and commitment therapy.

4. Discussion

We began the paper with Barlow's (2002) observation that the literature is filled with diametrically opposed interpretations of the Dodo bird research, and that one's interpretation of this literature is strongly associated with one's beliefs about how psychotherapy works. Our current findings may also account for how psychotherapy researchers can draw such differing conclusions from meta-analyses of treatment comparison studies. When specific treatment outcomes are assessed at the termination of therapy, some bona fide treatments are more effective than other bona fide treatments, which provides some qualified support for the specific factors/EST position. These effect sizes remained heterogeneous even when possible outliers were examined, and the absolute value of these effect sizes was statistically greater than the absolute value of the effect sizes at pre-treatment. The upper bound of this treatment difference was .29, but given that the upper bound of pre-treatment randomization is about .09, the actual effect size for the difference between two treatments is certainly smaller than .29 when not capitalizing on chance. Considering that when CBT was compared to other treatments, the average effect size at termination for targeted outcomes was .16 (which is very close to the difference between the upper bound estimates at pretreatment and termination; .09–.29), this value may be considered a reasonable approximation of the difference between bona fide treatments at termination. An effect size of approximately .20 is small in terms of Cohen's (1988) guidelines, but it still may be clinically meaningful. If, following Rosenthal and Rubin's (1982) binomial effect size display approach, success is defined as the median outcome score on the primary measures at termination, an estimated effect of .20 ($r = .10$) means that the more effective treatment will have a 55% success rate compared to a 45% success rate for the less effective treatment. To the extent that the aim of psychotherapy is to reduce the severity of targeted symptoms at the end of treatment (e.g., Barlow, 2004b), it appears that some treatments are incrementally more effective than other treatments.

The Dodo bird hypothesis had greater support with respect to secondary outcomes at both termination and follow-up and also perhaps in terms of primary outcomes at follow-up. Exactly how much credence to afford to the significance testing for these variables depends on one's

views about how outliers should be analyzed and about the ad hoc Monte Carlo analyses that were conducted. Regardless of whether these effect sizes were statistically different from zero, it is clear that the magnitude of these effects was very small, with upper bound estimates never exceeding .19, compared to upper bound estimates at pre-treatment of approximately .09. In fact, aside from the results for the primary outcomes at termination, the results from this meta-analysis were strikingly similar to Wampold, Mondin, Moody, Stich, et al.'s (1997) Dodo bird meta-analysis, even though the two meta-analyses shared no studies in common and used different analytic strategies. The average *d* for the absolute values of the primary outcomes at follow-up and the secondary outcomes at termination and follow-up ranged from .17 to .18 ($M = .18$), which is quite similar to the absolute value *d*s reported by Wampold et al. (range = .18–.21; $M = .20$). Also, like Wampold et al.'s overall findings, the secondary outcomes in the current meta-analysis were homogeneous at both termination and follow-up. Although there was statistically significant heterogeneity for the primary outcomes at follow-up, this heterogeneity was largely due to one outlier study. Overall, it appears that bona fide treatments yield similar outcomes with respect to quality of life measures and other measures of psychopathology that were not the specific target of the treatment. To the extent that clients are seeking therapy for something more than or different from specific symptom relief (e.g., personality change, improved quality of life, personal growth), bona fide treatments generally appear to yield similar outcomes.

There are two likely reasons why the statistically significant post-treatment primary outcomes that we found were obscured in Wampold, Mondin, Moody, Stich, et al.'s (1997) meta-analysis. First, Wampold et al. combined primary and secondary outcomes and studies often report more secondary outcomes than primary outcomes (e.g., depression treatment studies may include two or three measures of depression, but often also include measures of anxiety, general distress, quality of life, etc.), so secondary outcomes with smaller effect sizes could have masked statistically significant primary outcomes. Second, depending on the analysis, Wampold et al. either combined post-treatment and follow-up outcomes (which may have included multiple follow-ups) or gave priority to the follow-up outcomes. Therefore, it is not surprising that Wampold et al.'s results were more similar to our results for the follow-up and secondary measures than to the primary measures at termination.

Given the statistically significant heterogeneity among the primary outcomes at termination, it may be worthwhile to examine the studies that yielded the largest effect sizes. A study by Deckersbach, Rauch, Buhlmann, and Wilhelm (2006) yielded the largest effect size ($d = 1.42$), which was almost twice as large as the next largest difference. They found that habit reversal therapy was more effective for reducing tic severity in adults with Tourette syndrome than supportive therapy. The next three studies, with effect sizes ranging between .71 and .86, found that CBT was superior to either applied relaxation (Clark et al., 2006) or to meditation (Koszycki, Benger, Shlik, & Bradwejn, 2007) for the symptoms of social phobia, and that CBT was superior to applied relaxation for treating panic attacks (Arntz & Van Den Hout, 1996). In each of these four studies, a highly symptom-focused treatment (habit reversal or CBT) was more effective than a less focused treatment (supportive therapy, meditation, or applied relaxation) at reducing a very specific symptom (tics or panic attacks) or a relatively specific symptom (social phobia). In contrast, of the ten studies that treated depression or GAD, none yielded an effect size greater than .55. These findings are consistent with Chambless' (2002) suggestion that there may be higher levels of treatment equivalence for some disorders such as depression and greater treatment differences for disorders like panic disorder. In other words, there may be Dodo disorders and non-Dodo disorders. Similarly, Westen, Novotny, and Thompson-Brenner (2004) noted that short-term targeted treatments may be most effective for treating disorders characterized by specific discrete symptoms (e.g., panic disorder), but may be less effective for treating "generalized affect states" (p. 655) or more characterological conditions such as depression or GAD.

There was also a trend toward larger treatment sizes for primary outcomes at termination for younger samples. Given that most of the studies in the present meta-analysis ($k = 47$) used adult samples and that this finding was of marginal statistical significance ($p = .06$), this association should be interpreted with caution. This finding is, however, consistent with previous meta-analyses of the child and adolescent treatment literature which has yielded evidence supporting the role of specific factors in treatment outcomes for targeted problems (Weisz, Weiss, Han, Granger, & Morton, 1995), and evidence that ESTs are modestly more effective than standard care for child and adolescent clients (Weisz, Jensen-Doss, & Hawley, 2006; Weisz et al., 2013). In contrast, there was no association between year of publication and the magnitude of the effect sizes for primary outcomes at termination, suggesting that changes in reporting standards have not had an effect on the results from treatment comparison studies.

4.1. The cognitive contrast

The current cognitive contrast meta-analysis and Tolin (2010) only had two overlapping studies and Tolin did not include studies that compared CBT to behavior therapy. Nevertheless, the results from the two meta-analyses were generally consistent. For primary outcome measures at termination, our average effect size was .16 and Tolin's was .22. Like Tolin, we also found that CBT was superior to other treatments at follow-up, although our effect size of .16 was considerably smaller than the .47 effect size reported by Tolin. The discrepancy was likely due to the composition of the comparison studies: 18 of the 20 follow-up studies analyzed by Tolin compared CBT to psychodynamic therapy, whereas only two of the follow-up studies in our meta-analysis compared CBT to psychodynamic therapy. These two studies yielded an average effect size ($d = .46$) that was similar to Tolin's follow-up results. Overall, there appears to be consistent evidence that CBT is superior to psychodynamic therapies for treating specific symptoms or targeted problems, both at termination and at follow-up. We also found that CBT was statistically more effective than other therapies at termination for treating anxiety, but again our effect size was considerably smaller than the value reported by Tolin (.12 versus .43). Because exposure-based behavioral treatments have proven effective for treating anxiety (Barlow, 2004a), this discrepancy may be due to our decision to separate CBT from behavior therapy in contrast to Tolin's decision to include behavior therapy within CBT.

Consistent with both Baardseth et al.'s (2013) re-analysis of Tolin's meta-analysis and their "comprehensive anxiety meta-analysis" (p. 399), we failed to find any differences between CBT and other treatments for secondary or non-disorder specific outcomes. Although we found that CBT was statistically superior to other treatments for treating anxiety, whereas Baardseth et al. did not, our effect size was actually slightly smaller than the effect size reported by Baardseth et al. ($d = .12$ in the current study, $g = .14$ in Baardseth et al.). Because our meta-analysis included more studies, it had greater statistical power to find a small but statistically significant effect.

4.2. Limitations

Although we believe that Wampold, Mondin, Moody, Stich, et al.'s (1997) strategy of using the *Q* statistic to examine the heterogeneity of the effect sizes from treatment outcome studies is an innovative method for testing the Dodo bird hypothesis, one limitation of this approach is that in many situations the *Q* statistic is underpowered to detect actual heterogeneity (Huedo-Medina, Sánchez-Meca, Marin-Martinez, & Botella, 2006). Had there been instances where there was moderate or large amounts of heterogeneity as indicated by I^2 along with *Q* values that were not statistically significant, this issue of power may have been a greater limitation. Similarly, despite the other

statistical concerns that have been identified regarding the original Wampold et al. Dodo bird study, with between 136 and 295 effect sizes per analysis, a lack of statistical power was not an issue for their meta-analysis. Regardless, future meta-analyses that rely on the *Q* statistic to test the study hypotheses should be sensitive to this issue of statistical power.

A frequently noted limitation of meta-analyses designed to test the Dodo bird hypothesis is that by mixing a variety of treatments for a variety of disorders, they are unable to address the specific question of which treatments are most effective for which disorders, thus possibly obscuring genuine specific effects (e.g., Chambless, 2002; DeRubeis, Brotman, & Gibbons, 2005; Howard, Krause, Saunders, & Kopta, 1997). Although it was possible to examine some moderators in the cognitive contrast meta-analysis, most of these analyses only involved a small subset of studies and there were certainly not enough studies to examine treatment by disorder interactions. The even more limited number of studies that provided follow-up data likely accounted for the inconsistencies between termination and follow-up: CBT was superior to other treatments for treating anxiety disorders and eating disorders, but not depression at termination, but at follow-up CBT was superior to other treatments for treating anxiety disorders and depression but not eating disorders. Still, these results are consistent with the findings that (a) CBT is superior to IPT for treating eating disorders at termination, but that at follow-up the two treatments are equivalent (Wilson, Grilo, & Vitousek, 2007), and (b) that CBT for depression can reduce relapse at follow-up (Hollon et al., 2005).

A related limitation derives from the research base for these Dodo bird meta-analyses, which is neither a reflection of the full range of psychotherapies provided in clinical practice nor the frequency with which such therapies are practiced. Thus, although the majority of practitioners identify themselves as psychodynamic (27%) or integrative (25%; Norcross & Rogan, 2013), only four studies compared psychodynamic therapy to another treatment. On the one hand, the continued discrepancy between the treatments that are most studied and the treatments that are most commonly used may continue to limit the usefulness of psychotherapy research (especially treatment comparison studies) for practitioners (Safran, Abreu, Ogilvie, & DeMaria, 2011; Westen et al., 2004). However, compared to CBT, psychodynamic therapy has not fared especially well in either the current meta-analysis or in Tolin (2010), which may not encourage additional research focused on these treatments. Furthermore, the treatments included in the current meta-analysis were almost exclusively short-term therapies, with a median of 12 therapy sessions and only one long-term treatment study. Thus, any conclusions about treatment equivalence or relatively small differences between bona fide treatments should be limited to the treatments being compared (most frequently CBT to behavior therapy), and should not be generalized to long-term treatments or used to advocate for the efficacy of untested therapies.

The limitations inherent in interpreting follow-up outcomes are worth noting. Although Westen et al. (2004) and others have criticized psychotherapy researchers for not reporting long-term outcome data, the results from follow-up assessments are often ambiguous. Participants may receive additional treatment during the follow-up period and life events may contribute to further improvement or relapse. Furthermore, attrition during the follow-up period may not be random (i.e., those who improve more may be more likely to participate in follow-up assessments). Therefore, the attenuation of treatment differences for the primary outcomes from termination to follow-up could reflect genuine treatment equivalence over time (e.g., IPT outcomes catching up to CBT outcomes for bulimia at follow-up), or could be a methodological artifact (e.g., clients receiving therapy during the follow-up period).

Another limitation is that this meta-analysis only included 51 studies, which is fewer than were included in Wampold et al. (yet more than Baardseth et al., 2013 or Tolin, 2010). It may be noteworthy that although we reviewed the same six journals as Wampold et al., the average number of included treatment comparison articles published per year in these journals declined from 4.4[6] in Wampold et al.'s meta-analysis to 3.0 in the current meta-analysis. This decline cannot be attributed to our decision to omit studies with college student volunteers because only two studies were omitted as a result of this criterion (Fig. 1). Furthermore, our inclusion criteria were more liberal than Wampold et al.'s with respect to including studies that used premaster's-level graduate student therapists. Instead, it seems likely that the number of treatment comparison studies is declining. It may be that the Task Force EST criteria have encouraged psychotherapy researchers to redirect their focus to placebo control studies, which can suffice for establishing that a psychotherapy is an EST. Critiques of the scientific value of treatment comparison studies may have also contributed to this decline. As Borkovec and Castonguay (1998) have argued, most treatment comparison studies include multiple confounds (e.g., are the two treatments provided with equivalent levels of quality?) that make the findings from treatment comparison difficult to interpret. For example, the National Institute of Mental Health Treatment of Depression Collaborative Research Protocol (Elkin et al., 1989), which may have been the most ambitious treatment comparison study, yielded findings that remain open to debate with some concluding that the study found that medication was superior to psychotherapy for treating severe depression (Klein & Ross, 1993) and others disputing this conclusion because of a site-by-treatment confound (Jacobson & Hollon, 1996).

Additionally, the failure of many treatment comparison studies to yield statistically significant differences between the treatments and the support that the Dodo bird hypothesis has received in previous meta-analyses may have also contributed to the decline of treatment comparison studies. Relatedly, treatment comparison studies are expensive in terms of costs, time, and effort, yet most are likely to be statistically underpowered. Thus, even using the least conservative estimate of $d = .30$, which capitalizes on chance, a treatment comparison study would require 175 patients *in each treatment condition* to achieve .80 power, yet the median total number of randomized participants in the 51 studies reviewed was 80. With only 40 participants per group, the median treatment comparison study achieved only .26 power. Only seven of the studies started by randomizing over 200 participants. Because an effect size of .30 is certainly an overestimate, it is clear that the vast majority of treatment comparison studies are underpowered to find treatment differences for targeted outcomes at termination, let alone at follow-up. Bell, Marcus, and Goodlad (2013) noted that dismantling and additive studies face a similar challenge with being underpowered.

4.3. Implications and conclusions

Critics of the Dodo bird hypothesis have argued that support for this hypothesis would have far reaching implications for clinical training and practice. The mixed findings from the current meta-analysis suggest a "both/and" rather than an "either/or" approach to the selection and training of prospective clinical psychologists. If, as appears to be the case, some issues and diagnoses require specific treatment techniques (e.g., habit reversal therapy for tic disorders, CBT for panic disorder), whereas others may respond equally well to a variety of interventions, ideally psychotherapists in training should be cognitively flexible and capable of assessing when specialized techniques would be beneficial and when common factors may be most important. Thus, psychotherapists need to be capable of critically reviewing the literature to determine whether there is good evidence favoring one treatment over another versus when they are treating "Dodo disorders." They also need to be capable of helping their clients articulate whether their main goal

[6] This value was estimated by counting the number of asterisked references in Wampold et al.'s Reference section (114). Wampold et al. did not report the number of articles included in the meta-analysis and 114 is smaller than the smallest number of effects in any of the meta-analyses they reported.

is symptom relief (primary outcomes) or more general life changes (secondary outcomes). Thus, programs should focus on recruiting students who are empathic, socially skilled, *and* intelligent. Ironically, whereas some (Rounsaville & Carroll, 2002) have argued that support for the Dodo bird would mean that minimally trained low-paid practitioners could provide mental health services, others (e.g., Strosahl, 1998), presuming that the Dodo bird hypothesis is false, have argued that minimally trained master's level therapists could be taught to simply follow a set of highly structured manuals. Our mixed findings may provide a powerful argument for why well-trained therapists who have competencies in both general clinical skills (i.e., common factors) and who are critical thinkers capable of evaluating the research literature may be best qualified to provide psychotherapy.

To the extent that the Dodo bird verdict is a proxy for the debate between a medical model of psychotherapy (i.e., common factors, like a good bedside manner, are important, but the specific active ingredients are the key to healing), and a contextual model (i.e., psychotherapy as "a cultural healing practice," Wampold, 2007, p. 860), the current meta-analysis provided some qualified support for each model. In support of specific ingredients, at termination some treatments were more effective than others for treating focused symptoms. In these circumstances CBT was statistically more effective than alternative treatments. However, in support of a contextual model, these effects were generally small. Furthermore, these treatment differences do not extend to secondary outcomes and usually dissipate at follow-up. Thus, although it would be irresponsible to withhold proven treatments when clients present seeking relief from specific symptoms such as panic attacks or tics, for most clients it is unlikely that the specific treatment manual used by the therapist will have a major impact on the treatment outcome, especially in the months following the termination of therapy. These conclusions remain strikingly similar to those reached by Luborsky et al. (1975) almost 40 years ago.

References

Altman, D.G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., & Lang, T. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*, *134*, 663–694.

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

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851.

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

Baardseth, T. P., Goldberg, S. B., Pace, B. T., Wislocki, A. P., Frost, N. D., Siddiqui, J. R., & Wampold, B. E. (2013). Cognitive-behavioral therapy versus other therapies: Redux. *Clinical Psychology Review*, *33*, 395–405.

Barlow, D. H. (2002). Editor's introduction. *Clinical Psychology: Science and Practice*, *9*, 1.

Barlow, D. H. (2004a). *Anxiety and its disorders: The nature and treatment of anxiety and panic* (2nd ed.). New York: Guilford Press.

Barlow, D. H. (2004b). Psychological treatments. *American Psychologist*, *59*, 869–878.

Bell, E. C., Marcus, D. K., & Goodlad, J. K. (2013). Are the parts as good as the whole? A meta-analysis of component treatment studies. *Journal of Consulting and Clinical Psychology*, *81*, 722–736.

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.

Chambless, D. L. (2002). Beware the Dodo bird: The dangers of overgeneralization. *Clinical Psychology: Science and Practice*, *9*, 13–16.

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

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

Clark, D.M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N., & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *74*, 568–578.

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

Crits-Christoph, P. (1997). Limitations of the Dodo bird verdict and the role of clinical trials in psychotherapy research: Comment on Wampold et al. (1997). *Psychological Bulletin*, *122*, 216–220.

Deckersbach, T., Rauch, S., Buhlmann, U., & Wilhelm, S. (2006). Habit reversal versus supportive psychotherapy in Tourette's disorder: A randomized controlled trial and predictors of treatment response. *Behaviour Research and Therapy*, *44*, 1079–1090.

DeRubeis, R. J., Brotman, M.A., & Gibbons, C. J. (2005). A conceptual and methodological analysis of the nonspecifics argument. *Clinical Psychology: Science and Practice*, *12*, 174–183.

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry*, *46*, 971–982.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, *2*, 105–124.

Frank, J.D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy* (3rd ed.). Baltimore: Johns Hopkins University Press.

Giesen-Bloo, J., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., van Asselt, T., & Arntz, A. (2006). Outpatient psychotherapy for borderline personality disorder: Randomized trial of schema-focused therapy vs transference-focused psychotherapy. *Archives of General Psychiatry*, *63*, 649–658.

Goldfried, M. R. (2013). What should we expect from psychotherapy? *Clinical Psychology Review*, *33*, 862–869.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, *7*, 119–137.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.

Hollon, S. D., DeRubeis, R. J., Shelton, R. C., Amsterdam, J.D., Salomon, R. M., O'Reardon, J. P., & Gallop, R. (2005). Prevention of relapse following cognitive therapy vs medications in moderate to severe depression. *Archives of General Psychiatry*, *62*, 417–422.

Howard, K. I., Krause, M. S., Saunders, S. M., & Kopta, S. M. (1997). Trials and tribulations in the meta-analysis of treatment differences: Comment on Wampold et al. (1997). *Psychological Bulletin*, *122*, 221–225.

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*, 193–206.

Hunsley, J., & Di Giulio, G. (2002). Dodo bird, phoenix, or urban legend? The question of psychotherapy equivalence. *The Scientific Review of Mental Health Practice*, *1*, 11–22.

Imel, Z., & Wampold, B. (2008). The importance of treatment and the science of common factors in psychotherapy. In S. D. Brown, & R. W. Lent (Eds.), *Handbook of counseling psychology* (pp. 249–262) (4th ed.). Hoboken, NJ: Wiley.

Jacobson, N. S., & Hollon, S. D. (1996). Cognitive–behavior therapy versus pharmacotherapy: Now that the jury's returned its verdict, it's time to present the rest of the evidence. *Journal of Consulting and Clinical Psychology*, *64*, 74–80.

Klein, D. F., & Ross, D. C. (1993). Reanalysis of the National Institute of Mental Health Treatment of Depression Collaborative Research Program general effectiveness report. *Neuropsychopharmacology*, *8*, 241–251.

Koszycki, D., Benger, M., Shlik, J., & Bradwejn, J. (2007). Randomized trial of a meditation-based stress reduction program and cognitive behavior therapy in generalized social anxiety disorder. *Behaviour Research and Therapy*, *45*, 2518–2526.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyana, T. P., Berman, J. S., Levitt, J. T., & Krause, E. D. (2002). The Dodo bird verdict is alive and well-mostly. *Clinical Psychology: Science and Practice*, *9*, 2–12.

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

Messer, S. B., & Wampold, B. E. (2002). Let's face facts: Common factors are more potent than specific therapy ingredients. *Clinical Psychology: Science and Practice*, *9*, 21–25.

Munsch, S., Meyer, A. H., & Biedert, E. (2012). Efficacy and predictors of long-term treatment success for cognitive-behavioral treatment and behavioral weight-loss treatment in overweight individuals with binge eating disorder. *Behaviour Research and Therapy*, *50*, 775–785.

Norcross, J. C., & Rogan, J.D. (2013). Psychologists conducting psychotherapy in 2012: Current practices and historical trends among Division 29 members. *Psychotherapy*, *50*, 490.

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

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166–169.

Rounsaville, B. J., & Carroll, K. M. (2002). Commentary on Dodo bird revisited: Why aren't we Dodos yet? *Clinical Psychology: Science and Practice*, *9*, 17–20.

Safran, J.D., Abreu, I., Ogilvie, J., & DeMaria, A. (2011). Does psychotherapy research influence the clinical practice of researcher–clinicians? *Clinical Psychology: Science and Practice*, *18*, 357–371.

Shadish, W. R., Matt, G. E., Navarro, A.M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*, 512–529.

Shadish, W. R., & Sweeney, R. B. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let Dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, *59*, 883–893.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, *92*, 581–604.

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

Strosahl, K. (1998). The dissemination of manual-based psychotherapies in managed care: Promises, problems, and prospects. *Clinical Psychology: Science and Practice*, *5*, 382–386.

Task Force on Promotion and Dissemination of Psychological Procedures (1995). Training in and dissemination of empirically-validated psychological procedures: Report and recommendations. *The Clinical Psychologist*, *48*(1), 3–23.

Tolin, D. F. (2010). Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review*, *30*, 710–720.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Wampold, B. E. (2007). Psychotherapy: *The* humanistic (and effective) treatment. *American Psychologist*, *62*, 857–873.

Wampold, B. E., Mondin, G. W., Moody, M., & Ahn, H. (1997). The flat earth as a metaphor for the evidence for uniform efficacy of bona fide psychotherapies: Reply to Crits-Christoph (1997) and Howard et al. (1997). *Psychological Bulletin*, *122*, 226–230.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, 'all must have prizes'. *Psychological Bulletin*, *122*, 203–215.

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist*, *61*, 671–689.

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A.M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. *JAMA Psychiatry*, *70*, 750–761.

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

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.

Wilson, G. T., Grilo, C. M., & Vitousek, K. M. (2007). Psychological treatment of eating disorders. *American Psychologist*, *62*, 199–216.