The present qualitative and quantitative review examines the efficacy of CBT versus non-CBT treatments for anxiety disorders and suggests that a prior failure (Barndeseth et al., 2013, p. 395) to detect differences in outcome can easily be attributed to excessive error variance in the analysis, essentially burying the “signal” under “noise.” Several sources of error variance are identified, including type of comparison therapy, diagnosis, what outcome is being measured, and study quality. It is suggested that the “signal” of CBT versus other psychotherapies can easily be seen or not seen, depending on what one chooses to analyze. Recommendations are made for further study about the relative efficacy of different psychological treatments.

**Key words:** anxiety disorders, cognitive-behavioral therapy, meta-analysis, psychotherapy. [Clin Psychol Sci Prac 21:351–362, 2014]

Any student of statistics or research methodology can attest to the fact that it is easier to retain a null hypothesis (e.g., there is no difference between intervention X and intervention Y) than it is to affirm an alternative hypothesis (e.g., there is a difference between intervention X and intervention Y). To retain a null hypothesis, one only need increase the degree of error introduced into the analysis. As the degree of noise increases, the signal becomes increasingly difficult to detect.

There are several intuitively obvious ways that a researcher could inadvertently decrease the signal-to-noise ratio in his or her analysis, thus increasing the likelihood of Type II error (erroneously retaining a null hypothesis). First, instead of analyzing a single, well-specified construct, he or she might mix together multiple measures of very different constructs. For example, in a meta-analysis of depression treatments, he or she might lump measures of depression together with measures of anxiety, measures of substance use, measures of psychosis, and so on. Second, he or she could blur the distinctions between unique interventions, either by comparing highly similar interventions to each other or by lumping a heterogeneous group of treatments together for comparison purposes (e.g., instead of comparing intervention X to intervention Y, he or she could compare intervention X to “everything else”). Third, he or she could group studies with widely different methodologies together; for example, comparing one study that employed an intent-to-treat analysis with others that analyzed data from treatment completers only. Fourth, he or she could fail to account for the methodological quality of the studies used in the meta-analysis, allowing methodologically poor studies to be treated equally with methodologically strong studies. In short, the more “noise” in the sample of studies used in the meta-analysis, the more likely a Type II error becomes.

This signal-to-noise issue is evident in the perennial “dodo bird” debate about the efficacy of psychotherapy. Some authors have argued for the null (noise) hypothesis that there are no differences in the efficacy of various forms of psychotherapy. Others have argued for the alternative (signal) hypothesis, that some psychotherapies are more effective than others. Meta-analyses have been used by researchers on both
sides of this debate, with some showing an advantage for behavioral and cognitive-behavioral therapies (Shadish, Matt, Navarro, & Phillips, 2000; Shapiro & Shapiro, 1982) and others not (Smith & Glass, 1977; Wampold et al., 1997).

Several authors have critiqued the meta-analyses favoring the null hypothesis, pointing out several of the methodological concerns, described above, that decreased the signal-to-noise ratio, including mixing together multiple measures of very different constructs (Crits-Christoph, 1997; Howard, Krause, Saunders, & Kopta, 1997), blurring the distinctions between unique interventions (Crits-Christoph, 1997; Hunsley & Di Giulio, 2002; Wilson & Rachman, 1983), grouping studies with widely different methodologies together (Crits-Christoph, 1997), and failing to account for the methodological quality of the studies used in the meta-analysis (Wilson & Rachman, 1983).

Tolin’s (2010) recent meta-analysis of the efficacy of cognitive-behavioral therapy (CBT) versus other psychotherapies \( (N = 1,981) \) attempted to increase the signal-to-noise ratio in several ways. First, rather than mixing together multiple measures of different constructs, outcome measures were categorized according to what was being measured (primary symptoms, co-morbid symptoms, general functioning, social adjustment, global symptoms, quality of life, or self-concept). Similarly, disorders or clinical concerns being treated were categorized (academic concerns, anxiety disorders, depressive disorders, developmental disorders, eating disorders, habit disorders, marital distress, personality disorders, psychotic disorders, subclinical concerns, substance use disorders, or unspecified) and examined separately. Second, unlike previous meta-analyses (Wampold et al., 1997), efforts were made to compare CBT, which was categorized according to the presence of specific interventions (relaxation, exposure, behavioral rehearsal, cognitive restructuring, or operant procedures) only to bona fide alternative forms of psychotherapy (as defined by Wampold et al., 1997, and Westen, Novotny, & Thompson-Brenner, 2004), which were categorized according to what kind of comparison psychotherapy was being used (interpersonal therapy, psychodynamic therapy, or supportive therapy); furthermore, CBT-to-CBT comparisons were excluded. Third, the impact of study quality and methodological adequacy were assessed, according to the “gold standard” criteria proposed by Foa and Meadows (1997). Finally, the impact of researcher allegiance to one school of psychotherapy or another was examined, using allegiance ratings collected directly from the principal investigators of each study.

Results of the Tolin (2010) meta-analysis suggested that across diagnoses, CBT was superior to psychodynamic therapy, although not interpersonal or supportive therapies, at posttreatment and at follow-up for measures of the primary symptoms of the disorders being treated. Methodological strength of studies was not associated with larger or smaller differences between CBT and other therapies. Researchers’ self-reported allegiance was positively correlated with the strength of CBT’s superiority; however, when controlling for allegiance ratings, CBT was still associated with a significant advantage. The superiority of CBT over alternative therapies was evident only among patients with anxiety or depressive disorders.

More recently, Baardseth et al. (2013) published a critique of the Tolin study. Although agreeing with the statistical results, Baardseth et al. raised two intriguing conceptual questions: What is CBT? and How should outcomes be measured? Regarding the first question, there is certainly room for debate about what constitutes a theoretical orientation in psychotherapy (one could argue, for example, that pure cognitive restructuring is sufficiently different from certain behavioral interventions that they should not be classified together). Baardseth et al. did not attempt to define CBT, but did survey CBT practitioners \( (N = 91, \text{ a } 29\% \text{ survey response rate}) \) for their opinions on the topic. This methodology led to some potentially important differences in categorization; for example, in the Baardseth et al. meta-analysis, eye movement desensitization and reprocessing (EMDR) and present-centered therapy (PST) were categorized as non-CBT therapies, whereas in the Tolin meta-analysis, they were considered variants of CBT. The debate about the effective ingredients of EMDR is long-standing; many (e.g., Devilly, 2002; Lohr, Lilienfeld, Tolin, & Herbert, 1999; McNally, 1999) have suggested that imaginal exposure and cognitive restructuring, well-established components of CBT (Goldfried & Davison, 1994; Masters, Burish, Hollon, & Rimm, 1987), are...
the active mechanisms of that treatment. PST is described by the authors of clinical trials at times as being a nonspecific control treatment (Schnurr, Friedman, Lavori, & Hsieh, 2001), and at other times as a form of problem-solving training (McDonagh et al., 2005); to the extent that PST emphasizes problemsolving training, this would place it squarely in the realm of a well-established CBT intervention (Goldfried & Davison, 1994; Nezu, Nezu, & Lombardo, 2003).

The second question raised by Baardseth et al. (2013) is how outcomes of clinical trials should be defined. In most clinical trials, the primary outcome measure is of the disorder being treated. For example, in a study of panic disorder, the primary outcome would be a measure of the severity of panic disorder, such as the Panic Disorder Severity Scale (Shear et al., 1997). However, Baardseth et al. (2013) make the point that perhaps comparison psychotherapy trials should not focus exclusively on measures of the disorder being treated, stating, “CBT typically targets specific symptoms whereas other treatments may not” (p. 402) and “It may well be that CBT is able to effect change in disorder-specific symptom measures vis-à-vis other treatments, but not show any superiority with regard to nondisorder specific (i.e., nontargeted or global) measures” (p. 397). “Global” measures presumably refer to overall measures of psychological distress or impairment, without regard to a specific diagnosis. For example, in a study of panic disorder, a global measure might be a measure of the severity of overall illness, such as the Clinician’s Global Impression (Guy, 1976), or it might be a measure of quality of life, such as the Quality of Life Inventory (Frisch, Cornell, Villanueva, & Retzlaff, 1992). It is reasonable to expect that any treatment should impact overall mental health and quality of life. However, for purposes of treatment comparison, it is noted that such measures likely have less overall response (signal) and more error variance (noise), potentially obscuring real differences between the treatments.

“Nontargeted” measures, also advocated by Baardseth et al. (2013), presumably refer to measures of disorders that are not the primary diagnosis being treated. For example, in a study of panic disorder, a nontargeted measure might be a measure of the severity of depression, such as the Beck Depression Inventory (Beck, Steer, & Brown, 1996). The logic of this argument is a bit difficult to follow, as it appears that Baardseth et al. (2013) are suggesting that if we measure outcomes other than the disorder we are trying to treat, CBT and non-CBT treatments may not differ in efficacy. If that is indeed the argument, it is a rather unique one, and not one evidently shared by the authors of the studies themselves. The titles of the articles sampled note that the treatments are “for chronic posttraumatic stress disorder” (McDonagh et al., 2005, p. 515), for “PTSD treatments” (Taylor et al., 2003, p. 330), “for generalized anxiety disorder” (Durham et al., 1994, p. 315), “for PTSD rape victims” (Rothbaum, Astin, & Marsteller, 2005, p. 607), “for posttraumatic stress disorder” (Schnurr et al., 2003, p. 481; Schnurr et al., 2007, p. 820), “for patients with panic disorder” (Shear, Houck, Greeno, & Masters, 2001, p. 1993), “for traumatic stress” (Ironson, Freund, Strauss, & Williams, 2002, p. 113), “in generalized anxiety disorder” (Leichsenring et al., 2009, p. 875), “in the amelioration of posttraumatic stress disorder” (Devilly & Spence, 1999, p. 131), “for social phobia” (Borge et al., 2008, p. 991), “in the treatment of posttraumatic stress disorder” (Power et al., 2002, p. 299), and “treatment of PTSD” (Lee, Gavriel, Drummond, Richards, & Greenwald, 2002, p. 1071).

Baardseth et al. (2013) then produced a new meta-analysis of CBT versus non-CBT treatments for anxiety disorders. Thirteen clinical trials, published between 1991 and 2011, were sampled. Results showed no differences between CBT treatments and “bona fide” non-CBT treatments across disorder-specific and nondisorder-specific symptom measures (most treatments, regardless of treatment type, were associated with strong pretreatment to posttreatment effect sizes). Presented below is a reanalysis of the studies sampled in the Baardseth et al. study.

METHOD
Data Source and Study Ratings
The present meta-analysis used the 13 CBT versus non-CBT studies sampled by Baardseth et al. (2013; see Table 1, from Baardseth et al.). Study data were extracted by the author, with queries to study authors as needed. Four doctoral-level independent raters who
were researcher-clinicians specializing in anxiety disorders with experience in the conduct of clinical outcome research reviewed the Methods sections of the articles, with two raters reviewing each article. During the initial rating, disagreement occurred on 97 of 448 (21.7%) items. Disagreements were resolved through mutual discussion. On three (0.7%) items, the disagreement could not be resolved through mutual discussion and was resolved by the author. Specific ratings included the following.

Methodological quality, according to the “gold standards” listed by Foa and Meadows (1997), including (a) clearly defined target symptoms (diagnosis and minimum severity required for study entry, inclusion/exclusion criteria discussed), (b) reliable and valid measures (rated for each outcome measure used in the study), (c) use of evaluators unaware of treatment condition, (d) mention of assessor training, (e) manualized, specific treatment programs (i.e., a manual was used and followed), (f) unbiased assignment to treatment (either random or stratified), and (g) treatment adherence (both assessed and reported to be adequate).

For each outcome measure, the construct being measured, categorized as (a) primary symptom severity (measuring one or more symptoms of the principal diagnosis), (b) comorbid symptoms (symptoms of a specific psychiatric disorder that are not part of the principal diagnosis), (c) global symptom severity (an overall measure of psychiatric symptoms, not limited to a specific diagnosis or cluster of symptoms), (d) functioning or quality of life (an overall measure of functional impairment, or a broad-based measure of functioning, that is not limited to a specific domain of functioning), or (e) mechanism (a measure that is not designed to be a clinical outcome measure, but rather relates to a specific model of treatment and is expected to show different mechanisms across treatments).

Data Synthesis

Outcome measures were categorized according to the five categories described above. When multiple outcome measures were used within a single category (e.g., for a panic disorder study, two measures of panic disorder severity), an aggregate effect size was used (Rosenthal & Rubin, 1986). Studies were also divided according to whether an intent-to-treat or completer analysis was used.

Data were analyzed using Comprehensive Meta-Analysis v.2.2 software. For each comparison of continuous data (e.g., mean scores on a measure of psychopathology), Hedge’s $g$ was calculated, weighted by sample size. Hedge’s $g$ is a statistically corrected variation of Cohen’s $d$, for which 0.2, 0.5, and 0.8 are conventionally understood to represent small, medium, and large effects, respectively (Cohen, 1988). For studies reporting dichotomous outcomes (i.e., treatment responder versus nonresponder), odds ratios (OR) were calculated and converted to $g$. The 95% confidence interval (CI), and statistical significance ($p$) were calculated for each effect size estimate. To examine

<table>
<thead>
<tr>
<th>Clinical Trial</th>
<th>N</th>
<th>Disorder</th>
<th>CBT Treatment</th>
<th>Non-CBT Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borge et al. (2008)</td>
<td>73</td>
<td>Social anxiety</td>
<td>Cognitive therapy</td>
<td>Interpersonal psychotherapy</td>
</tr>
<tr>
<td>Devilly and Spence (1999)</td>
<td>23</td>
<td>PTSD</td>
<td>Cognitive-behavior trauma treatment</td>
<td>EMDR</td>
</tr>
<tr>
<td>Durham et al. (1994)</td>
<td>64</td>
<td>GAD</td>
<td>Cognitive therapy</td>
<td>Psychoanalytic therapy</td>
</tr>
<tr>
<td>Ironson et al. (2002)</td>
<td>19</td>
<td>PTSD</td>
<td>Prolonged exposure</td>
<td>EMDR</td>
</tr>
<tr>
<td>Lee et al. (2002)</td>
<td>24</td>
<td>PTSD</td>
<td>Stress inoculation prolonged exposure therapy</td>
<td>EMDR</td>
</tr>
<tr>
<td>Leichsenring et al. (2009)</td>
<td>57</td>
<td>GAD</td>
<td>Cognitive-behavioral therapy</td>
<td>Short-term psychodynamic therapy</td>
</tr>
<tr>
<td>McDonagh et al. (2005)</td>
<td>51</td>
<td>PTSD</td>
<td>Cognitive-behavioral therapy</td>
<td>Present-centered therapy</td>
</tr>
<tr>
<td>Power et al. (2002)</td>
<td>48</td>
<td>PTSD</td>
<td>Exposure with cognitive restructuring</td>
<td>Present-centered therapy</td>
</tr>
<tr>
<td>Rothbaum et al. (2005)</td>
<td>40</td>
<td>PTSD</td>
<td>Prolonged exposure</td>
<td>EMDR</td>
</tr>
<tr>
<td>Schnurr et al. (2003)</td>
<td>325</td>
<td>PTSD</td>
<td>Trauma-focused group therapy</td>
<td>Present-centered group therapy</td>
</tr>
<tr>
<td>Schnurr et al. (2007)</td>
<td>284</td>
<td>PTSD</td>
<td>Prolonged exposure</td>
<td>Present-centered therapy</td>
</tr>
<tr>
<td>Shear et al. (2001)</td>
<td>66</td>
<td>Panic</td>
<td>Cognitive-behavioral therapy</td>
<td>Emotion-focused therapy</td>
</tr>
<tr>
<td>Taylor et al. (2003)</td>
<td>41</td>
<td>PTSD</td>
<td>Exposure therapy</td>
<td>EMDR</td>
</tr>
</tbody>
</table>

Note. CBT = cognitive-behavioral therapy; PTSD = posttraumatic stress disorder; GAD = generalized anxiety disorder; EMDR = eye movement desensitization and reprocessing.
differences between effect size estimates, the mixed-effects between-group heterogeneity ($Q$) was calculated. Random-effects models were used.

RESULTS

Replicating Baardseth and Colleagues’ Results

Because Baardseth et al. (2013) evidently combined completer and intent-to-treat data, the same procedure was followed here for replication purposes. Following Baardseth’s methods, for those studies providing both intent-to-treat and completer analyses, only the intent-to-treat data were included in this analysis, as this would provide a more conservative estimate of treatment response. For disorder-specific measures, the overall effect size ($g$) was 0.230 ($p = .16$, a small effect that is not markedly different from Baardseth and colleagues’ reported $g = 0.14$). For all other measures combined, the effect size ($g$) was 0.177 ($p = .32$), which is slightly more discrepant from Baardseth and colleagues’ reported $g = 0.16$. When only measures of comorbid psychopathology (symptoms of a specific psychiatric disorder that are not part of the principal diagnosis) were included, the effect size ($g$) was 0.102 ($p = .49$), a figure that is quite close to Baardseth and colleagues’ reported $g = 0.14$.

Taking a Closer Look at Baardseth and Colleagues’ Results

As discussed in the Introduction, lumping measures of different constructs together likely contributes to excessive “noise” in the analysis, which could bury the signal. The same is likely true of combining completer and intent-to-treat results. As shown in Figure 1, for disorder-specific measures, in the treatment completer sample, the pooled effect size ($g$) was 0.41 ($p = .08$), a finding nearly identical to that obtained by Tolin (2010) despite the fact that different studies were used. For all other measures combined, the pooled effect size was 0.22 ($p = .43$). Significant between-study heterogeneity was found (for symptom-specific measures, $Q_{11} = 69.23$, $p < .001$; for all other measures, $Q_{10} = 87.33$, $p < .001$), suggesting that results were not uniform across studies.

Six studies (Borge et al., 2008; Leichsenring et al., 2009; McDonagh et al., 2005; Schnurr et al., 2003, 2007; Shear et al., 2001) used intent-to-treat analyses. For measures of primary symptom severity, these studies had a nonsignificant pooled effect size of 0.30 ($p = .18$). Although the small number of studies precludes more fine-grained analysis, it is noted that two studies, one in which CBT and interpersonal therapy were combined with residential treatment (Borge et al., 2008), and one of which was a treatment for PTSD in survivors of childhood sexual abuse (McDonagh et al., 2005), favored the alternative treatment (although the completer sample in the McDonagh et al. study favored CBT), whereas the other four favored CBT. For all other measures, the pooled effect size was 0.38 ($p = .15$).

Effects Vary According to Comparison Therapy

Closer examination of Figure 1 suggests that much of the heterogeneity of effect size estimates can be attributed, at least in part, to the kind of psychotherapy used as a comparison condition. As shown in Table 2, for the completer samples, pooled effect size for CBT versus other treatment on measures of primary symptom severity was greatest for studies using present-centered therapy (large effect), followed by emotion-focused psychotherapy (large effect) and psychodynamic therapy (moderate effect). By contrast, the overall effect of CBT versus EMDR was not significant. When intent-to-treat and completer samples were combined, a significant and moderate effect was found for CBT versus psychodynamic therapy.

Effects Vary According to Diagnosis

Table 2 also shows that outcomes for measures of primary symptom severity differed according to the principal diagnosis. For the completer sample, pooled effect size for CBT versus other treatment was greatest for the lone study of patients with panic disorder (large effect), followed by generalized anxiety disorder (moderate effect). By contrast, the overall effect of CBT versus EMDR was not significant. When intent-to-treat and completer samples were combined, a significant and moderate effect was found for CBT versus psychodynamic therapy.
Figure 1. Individual and group effect size (g, with 95% CI) for studies of CBT versus other psychotherapies for anxiety disorders, measuring symptoms of the main disorder being treated (top), and all other measures (bottom).
with generalized anxiety disorder (confounded with psychodynamic therapy).

Effects Vary According to Outcome

Finally, Table 2 shows that the effect of CBT versus other psychotherapies varies according to what outcomes are being measured. For the completer samples, pooled effect size for CBT versus other treatment was greatest for measures of global illness severity (large effect), followed by measures of primary symptom severity (small effect), measures of functioning/quality of life (small effect), and treatment mechanisms (small effect). By contrast, the pooled effect for measures of comorbid psychopathology was quite small and nonsignificant. When intent-to-treat and completer samples were combined, a significant and moderate effect was found for measures of global illness severity.

<table>
<thead>
<tr>
<th>Table 2. Relationship between comparison therapy, diagnosis, and outcome measure and pooled effect size of CBT versus other psychotherapies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completer Samples</strong></td>
</tr>
<tr>
<td><strong>Number of Comparisons</strong></td>
</tr>
<tr>
<td><strong>Comparison therapy</strong></td>
</tr>
<tr>
<td>Present-centered therapy</td>
</tr>
<tr>
<td>Emotion-focused therapy</td>
</tr>
<tr>
<td>Psychodynamic therapy</td>
</tr>
<tr>
<td>Eye movement desensitization and reprocessing therapy</td>
</tr>
<tr>
<td>Interpersonal therapy</td>
</tr>
<tr>
<td><strong>Diagnosis</strong></td>
</tr>
<tr>
<td>Panic disorder</td>
</tr>
<tr>
<td>Generalized anxiety disorder</td>
</tr>
<tr>
<td>Posttraumatic stress disorder</td>
</tr>
<tr>
<td>Social phobia</td>
</tr>
<tr>
<td><strong>Outcome measure</strong></td>
</tr>
<tr>
<td>Global illness severity</td>
</tr>
<tr>
<td>Primary symptom severity</td>
</tr>
<tr>
<td>Functioning/quality of life</td>
</tr>
<tr>
<td>Treatment mechanisms</td>
</tr>
<tr>
<td>Comorbid symptoms</td>
</tr>
</tbody>
</table>

Note. *p < .05; **p < .001.

*When a study reported both the completer and intent-to-treat (ITT) results, only the ITT results were included in the meta-analysis.

Effects Vary According to Study Quality

Methodological strength indices as described by Foa and Meadows (1997), and the pooled effect size for completer samples with positive and negative ratings on those items, are shown in Table 3. Of note, there was an insufficient number of studies meeting the Foa and Meadows (1997) criterion of description of assessor training. Of the remaining items, the greatest discrepancy was seen between studies that did and did not use evaluators unaware of treatment condition, followed by studies that did and did not use random assignment, comparisons that did and did not use reliable and valid measures, and studies that did or did not have clearly defined target symptoms. In each case, the studies that received a positive rating for methodological strength showed a greater pooled effect of CBT versus other psychotherapy. No difference was seen between studies that did and did not use manualized, specific treatment programs in both conditions or studies that did and did not assess adequate treatment fidelity.

DISCUSSION

As discussed in the Introduction, the greater the noise in a meta-analysis, the less likely one is to detect the signal. The meta-analysis by Baardseth et al. (2013), and their subsequent affirmation of the “dodo bird” verdict (the null hypothesis that the kind of psychotherapy being practiced has no impact on clinical outcomes), is an excellent example of this phenomenon. Baardseth et al. reported small to nonexistent differences in outcome between CBT and other psychotherapies for anxiety disorders. However, as demonstrated here, the data from Baardseth and colleagues’ own meta-analysis show that the signal of CBT versus other psychotherapies can easily be seen or not seen, depending on various methodological factors. The present analysis replicates the previous finding by Tolin (2010) that patients receiving and completing CBT fare significantly better at posttreatment than do patients receiving and completing other psychotherapies. This advantage is evident for patients receiving present-centered therapy, emotion-focused therapy, or psychodynamic therapy and for patients diagnosed with panic disorder or with generalized anxiety disorder. However, the limited number of studies for any particular diagnosis other than PTSD precludes definitive
conclusions about how CBT compares to non-CBT psychotherapy for that particular condition.

Baardseth et al. (2013) asserted, “It may well be that CBT is able to effect change in disorder-specific symptom measures vis-à-vis other treatments, but not show any superiority with regard to nondisorder specific (i.e., nontargeted or global) measures” (p. 397), yet the present analysis of measures of global illness severity (i.e., overall psychiatric illness without regard to any specific diagnosis) actually yielded the largest effects, in direct contrast to Baardseth and colleagues’ assertion. Thus, it cannot be reasonably argued that CBT’s only advantage is for a narrow range of symptoms.

It is noted that there were some studies (three of 12 completer comparisons and five of 14 comparisons when combining intent-to-treat and completer analyses) that showed an advantage of other psychotherapies over CBT. It is therefore worthwhile to take a closer look at these studies. One (McDonagh et al., 2005) showed an advantage for CBT in the completer analysis but an advantage for present-centered therapy in the intent-to-treat analysis in patients with posttraumatic stress disorder. A higher dropout rate for CBT was noted in that study, although the authors note that this could have been due to a high dropout rate for one inexperienced therapist in the CBT condition (when that therapist’s data were excluded, dropout rates were equal between CBT and present-centered therapy). Another study (Borge et al., 2008) showed an advantage for interpersonal psychotherapy over CBT when both treatments were included in a residential treatment program.

The remaining three studies showing an advantage for other psychotherapies over CBT (Ironson et al., 2002; Lee et al., 2002; Power et al., 2002) were all studies of EMDR for patients with posttraumatic stress disorder. Thus, of the five comparisons favoring another treatment over CBT, four were for patients with PTSD. As noted previously, some authors have suggested (and dismantling research appears to support) that the active ingredients of EMDR are exposure and cognitive restructuring, core elements of CBT. Therefore, comparing EMDR to CBT would be equivalent to comparing one version of CBT to another, a worthwhile goal but not one that helps us resolve the “dodo bird” verdict, as others have noted (Crits-Christoph, 1997; Hunsley & Di Giulio, 2002; Wilson & Rachman, 1983).

Examination of these three studies also raises important issues of methodological quality. The greatest discrepancy in comparison effect size was seen between studies that did and did not use evaluators unaware of treatment condition; the three studies suggesting an advantage of EMDR over CBT did not use evaluators unaware of treatment condition. A substantial discrepancy was also seen among studies that did and did not use random assignment to treatments; two of the three studies suggesting an advantage of EMDR appear not to have used random assignment. Thus, Baardseth and colleagues’ (2013) failure to find a signal of differential treatment outcome can be attributed, at least in part, to problems that have plagued previous analyses (e.g., Wampold et al., 1997), including mixing together multiple measures of very different constructs, blurring the distinctions between unique interventions, grouping studies with widely different methodologies together, and failing to account for the methodological quality of the studies used in the meta-analysis. The failure to find a signal is hardly surprising and is highly vulnerable to Type II error.
That having been said, Baardseth et al. (2013) raise several issues that have significant merit. They raise an important point that clinical outcome research should examine not only dimensional change (e.g., decrease in panic disorder symptom severity), but also categorical outcomes such as remission and clinically significant change. Although categorical outcomes are less sensitive than dimensional ones (and are therefore more likely to bury the signal of one treatment versus another), it is important to determine whether a given treatment results in a patient being well, not simply significantly better. Baardseth et al. also correctly point out that some non-CBT psychotherapies have been demonstrated to be effective. It is worth noting that a stronger effect of treatment X compared to treatment Y does not necessarily imply that treatment Y has no value. Baardseth et al. are also correct in pointing out that the differences between CBT and other psychotherapies are small to medium, using Cohen’s (1988) definitions. Thus, even though CBT shows an advantage over other psychotherapies for patients with anxiety disorders, it is far from a “home run.” I concur with Baardseth et al. that it is more important to identify the unique, active mechanisms of therapeutic change than it is to decide whether one “packaged” treatment is superior to another.

Several limitations of the present analysis should be noted. First, the overall number of comparison studies sampled (13) is rather low, and it is acknowledged that subsequent research could well yield different results. This small number of studies becomes particularly problematic when attempting to examine the effect of CBT versus other therapies for specific anxiety disorders (e.g., there was only one study of panic disorder), or when attempting to compare CBT to specific other therapies (e.g., there were only two samples, both from a single study, comparing CBT to psychodynamic therapy). Similarly, in some cases, diagnosis and comparison therapy are highly confounded (e.g., the single study of panic disorder is also the single study of EFT). Indeed, the only diagnosis with more than a handful of studies is PTSD; of the nine PTSD studies, six compared CBT to EMDR (pooled $g = -0.059$), and three compared CBT to PST (pooled $g = 0.865$).

The present analysis suggests several directions for future research comparing various forms of psychotherapy. These include the following.

1. **More comparison research is needed.** Although the “dodo bird” verdict can be safely overturned, it is clear from the present literature review that there are not enough comparison studies to reach fine-grained conclusions about which treatments are most effective for which disorders. Here, due to the low number of studies, diagnosis and comparison condition were confounded (e.g., PTSD studies were confounded with EMDR comparisons, panic studies were confounded with EFT studies). This concern can only be resolved by testing a wider range of treatments on a wider range of conditions.

2. **Research should avoid labels that obscure what is actually carried out in treatment.** The labels used to describe a treatment may not accurately reflect the therapeutic ingredients. As noted previously, Baardseth et al. (2013) and the authors of two clinical trials using PST seem to reach different conclusions about what PST actually is: a nonspecific control treatment (Schnurr et al., 2001), a problem-solving training (McDonagh et al., 2005), or a bona fide non-CBT treatment (Baardseth et al., 2013). Similarly, labeling a treatment “EMDR” obscures the fact that this treatment includes imaginal exposure and cognitive restructuring, which are at the core of CBT for PTSD (Devilly, 2002; Lohr et al., 1999; McNally, 1999). Even the label “CBT” can be confusing, as this term refers to a range of behavioral and cognitive interventions such that one “CBT” may not resemble another “CBT” strongly. Rather than labeling treatments, the field would benefit from an accurate description of what is actually carried out in treatment (e.g., exposure, cognitive restructuring).
3. Clinical trials, and reviews thereof, should clearly specify the desired outcomes. It would be difficult to argue that clinical trials should not include measures of primary symptom severity (yet surprisingly, some do not). A clinical trial testing the efficacy of a PTSD treatment, for example, should include a reliable and valid measure of PTSD severity in its outcomes. However, it is also worth noting that even in the present limited sample of studies, different outcome measures performed somewhat differently. It could be argued, therefore, that important secondary outcomes for any clinical trial should include improved overall mental health (and therefore a reliable and valid measure of global illness severity should be used) and improved functioning or quality of life (and therefore a reliable and valid measure of functional capacity and/or quality of life should be used). This will help the field compare apples to apples when combining results of multiple trials.

4. Quantitative reviews should consider the methodological quality of clinical trials. As shown here, the signal-to-noise ratio improves significantly (therefore reducing likelihood of Type II error) when studies use evaluators unaware of treatment condition, assign patients randomly to treatments, and use outcome measures that are reliable and valid for the intended purpose. This point is not merely a reiteration of the “garbage in–garbage out” argument; rather, the present results underscore the fact that methodological “garbage” can actually lead researchers to erroneously retain null hypotheses, of which the “dodo bird” verdict is only one.

NOTES
1. The author thanks Dr. Barbara Rothbaum for providing study data.
2. The author thanks Drs. Elizabeth Davis, Gretchen Diefenbach, Christina Gillam, and Bethany Wootton for serving as raters.

REFERENCES


Received October 25, 2013; revised January 20, 2014; accepted January 20, 2014.