Person 2019, Vol. 23, No. 1, 56–75

Einleitung der Herausgeber zum Artikel „In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy[1]“

Jana Lammers

Hamburg

Nicht nur der PZA sieht sich aktuell wieder vermehrt mit angeblich empirisch untermauerten Behauptungen konfrontiert, andere Psychotherapieverfahren, insbesondere die CBT (Kognitiv-behaviorale Therapie), seien ihm in der Wirksamkeit überlegen. Die Herausgeber[2] wollen deshalb den Lesern der PERSON einen aktuellen Beitrag präsentieren, der diese Diskussion empirisch beleuchtet.

Bruce Wampold, bekannt als Mit-Autor des in 2018 in deutscher Sprache erschienenen Buches „Die Psychotherapie-Debatte“[3], ist Vertreter des „Kontextuellen Metamodells“. Dieses postuliert im Gegensatz zum „Medizinischen Metamodell“, dass die Beziehung (zwischen Patient und Behandler) therapeutisch ist.

Der folgende Artikel „In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy“, den wir im englischen Original als Nachdruck veröffentlichen, befasst sich mit dieser immer wieder behaupteten größeren Wirksamkeit der CBT im Vergleich zu anderen Psychotherapieverfahren.

Diese Position wird bis heute vertreten, obwohl eine Arbeitsgruppe um Lester Luborsky bereits 1975 das Ergebnis vergleichender Meta-Analysen bezüglich der Wirksamkeit der verschiedenen Therapieverfahren mit den Worten des Vogels Dodo aus dem Buch „Alice in Wonderland“ zusammenfasste: „Everybody has won, and all must have prizes“[4]. Dieses als „Dodo-Bird-Verdikt“ in die empirische Psychotherapieforschung eingegangene Ergebnis wird auch aus methodischen Gründen immer wieder angezweifelt. So wurden auch immer wieder Studien präsentiert, die eine größere Wirksamkeit von CBT gegenüber humanistischer Therapieverfahren belegen sollen.

In der aktuellen 6. Auflage des Standardwerks von „Bergin and Garfield's Handbook of Psychotherapy and Behavior Change“[5] haben Robert Elliott et al. diese Unterschiede nachgerechnet und dabei auch die therapeutische Orientierung („allegiance“) der Forscher in Rechnung gestellt. Das Ergebnis ihrer Re-Analysen zeigt, dass die Wirksamkeit zwischen den Humanistischen Therapieverfahren und CBT statistisch äquivalent ist.

Auch der nachfolgende Artikel von Bruce Wampold et al. legt am Beispiel von drei Meta-Analysen zur verhaltenstherapeutischen Angstbehandlung dar, dass die dabei festgestellte Überlegenheit von CBT gegenüber anderen Therapieansätzen ein Produkt der Auswertungsmethodik ist. Hier werden die spezifischen Probleme dieser Meta-Analysen nachgewiesen, die eine vermeintliche Überlegenheit von CBT aufzeigen.

Als generelles Fazit lässt sich feststellen: Die Ergebnisse aufwändiger empirischer Vergleichsstudien bzw. vergleichender Meta-Analysen garantieren nicht, dass diese auch valide sind.

Schlüsselwörter zum Artikel: Kognitiv-behaviorale Therapie, Meta-Analyse, Wirksamkeit von Psychotherapie, Ängstlichkeit

1 Zuerst erschienen: Wampold, B. E., Flückiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, Goldberg, S. B., Miller, S. D., Baardseth, T. P., Laska, K. M. & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. *Psychotherapy Research, Vol. 27, (Nos. 1–2),* 14–32.

2 Es sind immer beide Geschlechter gemeint.

3 Wampold, B. E., Imel, Z. E., Flückiger, C. (2018) – *Die Psychotherapie-Debatte – was Psychotherapie wirksam macht.* Bern: Hogrefe.

4 Untertitel des Zeitschriftenartikels: Luborsky, l., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapy: Is it true “Everybody has won and must have prizes”? *Archives of General Psychiatry, 32,* 995–1008.

5 Lambert, M. J. (2013). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th edition). New Jersey: Wiley. Deutsche Fassung: Lambert, M. J. (2013). *Bergin und Garfields Handbuch der Psychotherapie und Verhaltensmodifikation.* Tübingen: Dgvt-Verlag.

Editor's preface to "In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy"

Not only the PCA sees itself currently again increasingly confronted with allegedly empirically substantiated allegations that other psychotherapy methods, in particular CBT (cognitive behavioral therapy), are superior where efficacy is concerned. Hence, the publishers would like to present readers of PERSON with a recent article which provides empirical insight into this discussion.

Bruce Wampold, known as the co-author of the German-language book "Die Psychotherapie-Debatte" (lit.: "The Psychotherapy Debate"), which was published in 2018[6], is a representative of the "contextual meta-model". Contrary to the "medical meta-model", it postulates that the relationship (between patient and therapist providing treatment) is therapeutic in nature.

The following article "In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy", which we are publishing as a reprint of the English original, examines this repeated assertion that CBT is more effective than other psychotherapy methods.

Even today, this position continues to have its advocates, even though a working group led by Lester Luborsky summarized the findings of comparative meta-analyses regarding the efficacy of various therapeutic methods in 1975 with the following words from the Dodo in "Alice in Wonderland": "Everybody has won, and all must have prizes"[7]. Known as the "Dodo bird verdict" in empirical psychotherapy research circles, this finding is also often cast into doubt due to reasons related to methodology. For one, studies were repeatedly presented which claimed to prove the higher efficacy of CBT as compared to humanistic therapy methods.

In the current 6th edition of the definitive work "Bergin and Garfield's Handbook of Psychotherapy and Behavior Change",[8] Robert Elliott et al. recalculated these differences, and also included the therapeutic orientation ("allegiance") of the researchers in the calculations. The findings of their re-analyses show that the efficacy of humanistic therapy methods and CBT is statistically equivalent.

Using three meta-analyses on the treatment of anxiety disorders via behavioral therapy, the following article from Bruce Wampold et al. also demonstrates that CBT's alleged superiority over other therapeutic methods is a product of evaluation methodology. It demonstrates the specific problems of these meta-analyses which claim to show that CBT is superior.

In general, the following conclusion can be made: The findings of sophisticated empirical comparative studies and/or comparative meta-analyses are not a guarantee that these findings are also valid.

6 Wampold, B. E., Imel, Z. E., Flückiger, C. (2018) – *Die Psychotherapie-Debatte – was Psychotherapie wirksam macht (lit.: „The Psychotherapy Debate — What makes psychotherapy effective")*. Bern: Hogrefe.

7 Sub-title of journal article: Luborsky, l., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapy: Is it true "Everybody has won and must have prizes"? *Archives of General Psychiatry, 32,* 995–1008.

8 Lambert, M. J. (2013). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th edition). New Jersey: Wiley.

Person 2019, Vol. 23, No. 1, 58–75

In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy

Bruce E. Wampold [(i, ii)], Christoph Flückiger[(iii)], A. C. Del Re[(iv)], Noah E. Yulish[(i)], Nickolas D. Frost[(i)], Brian T. Pace[(v)], Simon B. Goldberg[(i)], Scott D. Miller[(vi)], Timothy P. Baardseth[(vii)], Kevin M. Laska[(viii)] & Mark J. Hilsenroth[(ix)]

Abstract

Objective: Three recent meta-analyses have made the claim, albeit with some caveats, that cognitive-behavioral treatments (CBT) are superior to other psychotherapies, in general or for specific disorders (e.g., social phobia). Method: The purpose of the present article was to examine four issues in meta-analysis that mitigate claims of CBT superiority: (a) effect size, power, and statistical significance, (b) focusing on disorder-specific symptom measures and ignoring other important indicators of psychological functioning, (c) problems inherent in classifying treatments provided in primary studies into classes of treatments, and (d) the inclusion of problematic trials, which biases the results, and the exclusion of trials that fail to find differences among treatments. Results: When these issues are examined, the effects demonstrating the superiority of CBT are small, nonsignificant for the most part, limited to targeted symptoms, or are due to flawed primary studies. Conclusion: Meta-analytic evidence for the superiority of CBT in the three meta-analysis are nonexistent or weak.

Keywords: cognitive behavioral therapy, meta-analysis, psychotherapy effectiveness, anxiety

Since the origins of psychotherapy, there have been ferocious debates about whether one treatment was better than another. Alfred Adler and Carl Jung parted ways with Sigmund Freud because of their differences about theory and practice—there was a correct way to conduct psychoanalysis and claims were made about the superiority of one approach to another. Over time, the actors changed but the script has remained the same:

> Rivalry among theoretical orientation has a long and undistinguished history in psychotherapy dating back to Freud. In the infancy of the field, therapy systems, like battling siblings, competed for attention and affection in a "dogma eat dogma" environment. … Mutual antipathy and exchange of puerile insults between adherents of rival orientations were much the order of the day. (Norcross & Newman, 1992, p. 3)

(i) Department of Counseling Psychology, University of Wisconsin–Madison, Madison, WI, USA; (ii) Research Institute, Modum Bad Psychiatric Center, Vikersund, Norway; (iii) Department of Psychology, University of Zürich, Zürich, Switzerland; (iv) VA Palo Alto Health Care System, Palo Alto, CA, USA; (v) Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA; (vi) Center for Clinical Excellence, Chicago, IL, USA; (vii) Minneapolis VA Health Care System, Minneapolis, MN, USA; (viii) VA Salt Lake City Health Care System, Salt Lake City, UT, USA & (ix) Derner Institute, Adelphi University, Garden City, NY, USA

Correspondence concerning this article should be addressed to Bruce E. Wampold, Research Institute, Modum Bad Psychiatric Center, N-3370 Vikersund, Norway. E-mail: bwampold@wisc.edu

The presence of more than 400 brands of psychotherapy attests to the effort to develop therapies better than what presently exists, and each therapy has advocates who are prone to defend their territory (Dattilio & Norcross, 2006). Far from being an academic exercise, the process has momentous influence on policy and practice. If one treatment is indeed superior to another, then its adoption should improve the quality of mental health care. On the other hand, if there are no meaningful differences among treatments, limiting the availability of therapies to patients and therapists decreases quality and can be cost ineffective (Laska, Gurman & Wampold, 2014).

For many decades, arguments between rival schools were mainly theoretical or anecdotal in nature. A shift began in the 1950s and 1960s when Hans Eysenck used evidence from studies of psychotherapy to make claims about the superiority of behavioral therapies (Eysenck, 1952, 1961, 1966). Eysenck's claims changed the warrants that were used to argue about superiority, putting evidence at the forefront (Wampold, 2013b). The surge in the number of psychotherapy effectiveness studies led to meta-analysis as a means to synthesize this evidence and draw conclusions (Hunt, 1997; Mann, 1994; Wampold & Imel, 2015).

One of the first applications of meta-analysis involved Smith and Glass's (Smith & Glass, 1977; Smith, Glass, & Miller, 1980) meta-analyses of psychotherapy. Their finding that psychotherapy was remarkably effective contrasted sharply with Eysenck's claims. These meta-analyses further documented that

all treatments—behavioral and otherwise—were essentially equally effective when confounds were identified and controlled. Over the years, meta-analysis after meta-analysis have returned similar results; namely, that psychotherapies, in general, and for any particular disorder, are equally effective, and any differences found tend to be quite small and clinically unimportant (Laska et al., 2014; Wampold & Imel, 2015).

Despite this evidence, the quest to identify the "best" treatment approach continues. To this end, a handful of recent meta-analyses have purported to show that cognitive-behavioral treatments (CBT) are superior to other treatments for some specific disorders and more generally (Marcus, O'Connell, Norris, & Sawaqdeh, 2014; Mayo-Wilson et al., 2014; Tolin, 2010, 2014). For social phobia, Mayo-Wilson et al. (2014) stated the following:

> In particular, individual CBT had a greater effect than psychodynamic psychotherapy and other psychological therapies (interpersonal psychotherapy, mindfulness, and supportive therapy). Many of the psychological treatments with large effects were versions of CBT (individual, group, or self-help), suggesting that CBT might be efficacious in a range of formats. Psychodynamic psychotherapy was also effective, although its effects were similar to psychological placebo … Taking these factors into account, NICE [The National Institute for Health and Care Excellence in the UK] recently concluded that individual CBT should be offered as the treatment of choice for social anxiety disorder. (pp. 374, 375)

Marcus et al. (2014) made the following conclusion, with regard to the superiority of CBT generally, albeit with some qualifications:

> Contrary to the Dodo bird hypothesis, there was evidence of treatment differences for primary outcomes at termination … cognitive-behavioral treatments may be incrementally more effective than alternative treatments for primary outcomes. (p. 519)

Tolin (2014), with regard to anxiety disorders, also made a qualified statement about the superiority of CBT:

> 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. 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. (p. 351)

The purpose of the present article is to illustrate how certain conclusions from these meta-analyses may be flawed, based on how meta-analytic procedures are applied. As with any complex analytic procedure, there are opportunities for statistical and procedural errors. In this article, the three meta-analyses that concluded that CBT was superior to other treatments will be used to illustrate that care must be taken before various conclusions can be asserted. First the three meta-analyses are reviewed, then a number of common and specific problems will be identified in the meta-analyses that require diligence and finally solutions and recommendations that address these problems are offered.

The Meta-Analyses

Tolin (2014)

In 2010, Tolin conducted a meta-analysis of CBT versus other psychotherapies by examining studies in which two or more bona fide treatments were directly compared—a commendable feature that will be discussed below. Chief among the findings was that CBT was superior to other therapies for anxiety and depression.

The finding for depression was surprising given it contradicts one of the other meta-analysis to be discussed here (viz., Marcus et al., 2014) as well as other meta-analyses showing that all bona fide treatments for depression to be roughly equally efficacious (Cuijpers et al., 2013; Cuijpers, van Straten, Andersson, & van Oppen, 2008; Driessen et al., 2010; Wampold, Minami, Baskin, & Tierney, 2002). As well, the Tolin (2010) analysis omitted some prominent direct comparisons of CBT with other treatments for depression (e.g., emotion-focused therapy; see Baardseth et al., 2013).

With regard to anxiety, Baardseth et al. (2013) noted that Tolin (2010) only included four studies that directly compared CBT to other bona fide psychotherapies and they were dated (viz., published in 1967, 1972, 1994, and 2001), which makes any conclusion about the superiority of CBT for anxiety tenuous. According to Baardseth et al., the limited number of studies included for anxiety was due to the definition of CBT employed by Tolin, who classified both eye-movement desensitization and reprocessing (EMDR) and present-centered therapy (PCT) as CBT, whereas others have classified these same treatments as *not* CBT, which raises the question, "What is CBT?" (This question will be discussed here, but the reader is referred to Baardseth et al., Tolin, 2014, and Wampold 2013a for a discussion of this as well.) Because the definition of CBT, and the inclusion criteria that operationalize the definition, are ambiguous and change from study to study, Baardseth et al. used ratings of CBT experts to determine which treatments were considered CBT. In contrast to Tolin et al. (2010), the CBT experts classified both EMDR and PCT as not CBT. Based on experts' classification of CBT, the Baardseth et al. analysis found that CBT was not superior to other treatments for anxiety, on both non-disorder-specific and disorder-specific outcome measures.

In response to Baardseth et al.'s finding that CBT was not superior for anxiety disorders, Tolin (2014) reanalyzed the studies

Table I. Corrected effect sizes for CBT versus other therapies for anxiety for intent-to-treat (ITT) and completer samples.

	Combined		ITT		Completers	
Measures	*d*	*p*	*d*	*p*	*d*	*p*
Targeted	0.14	.21	0.14	.20	0.27	.05+
Non-targeted	0.09	.37	0.14	.13	0.10	.49

Source: Tolin (2014, 2015).
Note: Positive effect size indicates CBT produced superior outcomes. Combined samples were ITT, when available, and completers otherwise.

in Baardseth et al. and it is this reanalysis (and its corrigendum, Tolin, 2015) that is one of the three meta-analyses examined in this article. Based on the reanalysis, Tolin (2014) concluded, "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" (p. 357). Unfortunately, for one study (viz., Schnurr et al., 2003), Tolin (2014) miscalculated the effect by using the standard error rather than the standard deviation, which inflated the effect by a factor of 10 in favor of CBT. The corrected analyses appeared in a corrigendum (see Tolin, 2015) and the results of the primary contrast between CBT and other treatments are summarized in Table I. *All effects were small and nonsignificant*. Nevertheless, the following conclusion was still offered: "The basic conclusion that a signal favoring CBT over other psychotherapies is evident" (Tolin, 2015, p. 315).

Tolin (2014, 2015) went on conduct a number of subsequent analyses, searching for a signal detectable from background noise, to use his metaphor. A number of problems related to this meta-analysis will be discussed as issues related to conducting and understanding meta-analyses are presented.

Marcus et al. (2014)

The central purpose of Marcus et al.'s (2014) metaanalysis was to replicate Wampold, Mondin, Moody, Stich, et al. (1997)'s analysis of direct comparisons between bona fide treatments. As will be discussed in some detail below, direct comparisons are the best available means to test relative efficacy. In the 1997 meta-analysis, Wampold et al. retrieved all direct comparisons of bona fide treatments, regardless of theoretical approach, in six major journals and found that the effects of nearly 300 such comparisons were homogenously distributed around zero, indicating that the distribution of effects were as expected if the true difference among treatments was zero. The statistical method used, although criticized (and misunderstood) by some (e.g., Howard, Krause, Saunders & Kopta, 1997), has been investigated and found to be statistically sound (Wampold & Serlin, 2014). Wampold et al. also calculated an upper bound of the difference between treatments, by taking the average of the absolute values of the differences, and found the upper bound to be approximately $d = 0.20$, a small effect and one which overestimates the true difference between treatments (Wampold & Serlin, 2014).[1]

Marcus et al. (2014) retrieved studies from the same six journals used by Wampold, Mondin, Moody, Stich, et al. (1997) but with some important alterations. First, they analyzed only trials published since Wampold, Mondin, Moody, Stich, et al. (viz., 1997–2012). Second, they segregated primary and secondary measures, as opposed to Wampold et al. who aggregated all measures within studies, and analyzed results at termination and follow-up (Wampold and colleagues did both of these analyses in a follow-up report; see Wampold, Mondin, Moody, & Ahn, 1997).

Marcus et al. used Wampold et al.'s method to conduct the omnibus test of differences among treatments. The statistic *W* indexes the degree to which treatments differ and as noted in Table II, *W* was sufficiently large to reject the null hypothesis that all treatments were equally effective for the primary measures, a result different from Wampold et al. No differences among treatments were found, however, on secondary measures. Marcus et al. also calculated an upper bound index using the same method as Wampold et al., which was similar in size to Wampold et al. (viz., 0.29 for primary measures and 0.19 for secondary measures).

In an innovative way, Marcus estimated the upper bound when there were no differences among treatments (i.e., under the null hypothesis of no differences) by using pretest effects, which with random assignment would on average demonstrate no differences. For the pretreatment scores, the estimated upper bound was 0.11 (i.e., when there are no differences, the upper bound will be about 0.11). In Table II, the estimated effect for differences among treatments is presented as the difference between the upper bound à la Wampold et al. and the expected upper bound when there are no true differences à la Marcus et al., yielding an appealing estimate of treatment differences. Calculated this way, the effects were quite small (viz., 0.18 and 0.08 for primary and secondary measures, respectively).[2] The results at follow-up were similar, but generally even smaller (see Marcus et al.).

Marcus et al. (2014) followed up the omnibus test of differences by examining the CBT contrast, which involved comparisons of CBT to other treatments.

1 Wampold and Serlin (2014) discussed the expected value of the mean of the absolute value of the standardized effects, a more appropriate method to describe the effect of produced by differences among treatments.

2 Wampold and Serlin (2014) described an alternative way to examine expected values of effects under the null hypothesis first derived analytically by Geary (1935).

Table II. Omnibus tests of differences and CBT contrast at termination.

Omnibus test						
Outcome	*k*	*W*	*p*	Upper bound	Expected upper bound under null	Estimated effect
Primary	50	107.48	<.01	0.29	0.11	0.18
Secondary	38	43.85	.20	0.19	0.11	0.08
CBT v. other treatments						
	k	*d*	*p*	*Q*	*p*	
Primary	40	0.16	<.01	54.87	.04+	
Secondary	32	0.07	.10	33.81	.33	

Source: Marcus et al. (2014).
Note: The notation used here follows that of Wampold and Serlin (2014) for the omnibus test.

The results at termination are presented in Table II, where it can be seen that the effects were small (viz., 0.16 and 0.07 for primary and secondary measures, respectively). Although the contrast for primary measures was statistically significant, the interpretation of this result is mitigated by significant heterogeneity, as indicated by the Q statistic. Follow-up effects were generally smaller. Marcus et al. also reported effects for CBT versus various other treatments and found that CBT was superior to psychodynamic (PD) therapies ($d = .38$, but based on only three studies, all of which may have significant limitations, see Leichsenring et al., 2015), which resulted in the conclusion, "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*" (emphasis added, p. 528).

Marcus et al. (2014) examined outliers by removing one effect at a time from the analysis. In some instances, removing Clark et al. (2006), who compared CBT and a form of relaxation therapy for social anxiety, changed the conclusions because the effect for this study was extraordinarily large (viz., $d = 1.14$). This trial, as discussed below, is problematic for a variety of reasons.

Given the relatively small effects, significant only for primary measures, affected by an outlier, and which decreased at follow-up, Marcus et al. (2014) seemed to recognize their results necessitated nuance rather than a declaration of superiority when it came to CBT. When examining the nature of studies that showed large effects for CBT, they observed:

> 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). (p. 527)

In the end, Marcus et al. made the following conclusion:

> In support of specific ingredients, at termination some treatments were more effective than others for treating focused symptoms. … 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) [i.e., dodo bird conclusion] almost 40 years ago. (p. 529)

Mayo-Wilson et al. (2014)

Tolin (2014, 2015) and Marcus et al. (2014) meta-analyzed direct comparisons between treatments, a recommended practice for determining relative efficacy (Shadish & Sweeney, 1991; Wampold, Mondin, Moody, Stich, et al., 1997). Unfortunately, for many situations there is an insufficient number of comparisons to estimate the difference between two classes of treatment with much precision (cf., Marcus et al.'s CBT versus PD comparison with only three such comparisons). However, newly developed methods estimate differences between classes of treatments by using indirect comparisons as well as direct comparisons. It may well be that there are few if any direct comparisons of treatments A and B (denoted as AB), but there are many comparisons of A with C (AC) and B with C (BC). Then the effect for AB can be estimated indirectly from the effects for AC and BC using the transitive property, a procedure that has been called network meta-analysis (see Cipriani, Higgins, Geddes, & Salanti, 2013; Lumley, 2002). For example, Cipriani et al. (2009) used network meta-analysis to estimate the comparative efficacy of 12 new-generation antidepressants despite the fact that there were no comparisons between pairs of two particular antidepressants (e.g., Milnacipran versus Mirtazapine). Effects in network meta-analyses are often reported with regard to a reference group, in the case of antidepressants Fluoxetine, based on the fact that it was the first new-generation antidepressant to be marketed in the US and because it was often used as a reference drug in direct comparisons. Network meta-analyses typically are based on Bayesian estimates, although frequentist approaches exist.

Table III. Estimated effects for psychological interventions based on direct and indirect comparisons.

		versus Waitlist control		versus Group CBT		versus Individual CBT	
Intervention	*k*	SMD	Credible interval	SMD	Credible interval	SMD	Credible interval
Individual CBT	15	1.19	0.81; 1.56	0.27	−0.28; 0.81		
Group CBT	28	0.92	0.51; 1.33			−0.27	−0.81; 0.28
Exposure/social skills	10	0.86	0.29; 1.42	−0.06	−0.74; 0.61	−0.33	−0.99; 0.33
Self-Help with support	16	0.86	0.36; 1.36	−0.05	−0.69; 0.58	−0.32	−0.94; 0.30
Self-Help without support	9	0.75	0.26; 1.25	−0.17	−0.80; 0.47	−0.43	−1.05; 0.19
Psychological placebo	6	0.63	0.36; 0.90	−0.29	−0.72; 0.14	−0.56	−1.00; −0.11
PD	3	0.62	0.31; 0.93	−0.30	−0.80; 0.20	−0.56	−1.00; −0.11
Other psychotherapy	7	0.36	−0.12; 0.84	−0.55	−1.17; 0.06	−0.82	−1.41; −0.24

Source: Mayo-Wilson et al. (2014).
Note: *k* = number of studies in database (not number of direct comparisons); SMD = standardized mean difference. Positive SMDs indicate that the row treatment was superior to the column comparison group and negative SMDs indicate the row treatment was inferior to the column comparison group. Shaded cells indicate effects found to be significantly different from zero.

Mayo-Wilson et al. (2014) conducted a Bayesian network meta-analysis of psychological, self-help, and pharmacological interventions for social anxiety, using wait-list controls as the reference group. Classes of treatments included five drugs (viz., monoamine oxidase inhibitors, benzodiazepines, selective serotonin-reuptake inhibitors and serotonin–norepinephrine reuptake inhibitors (SSRIs and SNRIs, respectively), five psychotherapies (viz., individual CBT, group CBT, PD, exposure and social skills, and other psychological therapies), three types of control groups (viz., waitlist, pill placebo, and psychological placebo), self-help (viz., promotion of exercise, self-help with support, and self-help without support), and combined psychotherapy/drugs. By far, the greatest number of comparisons were between various SSRIs/SNRIs and pill placebo; CBT conditions predominated the psychotherapy conditions.

The pertinent results of Mayo-Wilson et al.'s (2014) meta-analysis are summarized in Table III, which presents the effect sizes for social phobia symptoms (denoted as standardized mean differences, SMD) for the psychological interventions compared to the reference group (waitlist controls) as well as to Group CBT and Individual CBT.[3] Note that in this table, *k* is the number of studies in the data base and *not* the number of direct comparisons between treatments. The credible interval is the Bayesian analog of a confidence interval.

Mayo-Wilson et al. (2014) found that all psychological interventions, with the exception of Other Psychotherapies, were significantly superior to waitlist controls (i.e., credible interval did not include zero), including Psychological Placebos. Notably Group CBT was not significantly superior to any of the other psychological interventions including Psychological Placebo. Individual CBT was superior to PD Therapy, Other Psychotherapy, and Psychological Placebos. Moreover the effect sizes reported for CBT versus other treatments were larger than those of either Tolin (2014, 2015) or Marcus et al. (2014). However, there are issues that call into question the validity of these estimates, as will be discussed in detail below.

Issues in the Interpretation of Meta-Analyses

Nearly 50 years have passed since the publication of Smith and Glass's (1977; Smith et al. 1980) pioneering meta-analyses showed psychotherapy was effective and all treatments were equally effective when confounds were controlled. While widely accepted today, it is easy to forget the controversy surrounding meta-analytic methods when they were first used. Indeed, the entire meta-analytic enterprise was severely criticized (see Wampold, 2013b; Wampold & Imel, 2015). In the critique that follows, it is argued that the method of meta-analysis is not flawed, but rather how the methods were and are being used to reach conclusions is what are of concern, particularly when meta-analyses are used to assess relative efficacy. The three meta-analyses will be used to illustrate the issues; recommendations for correcting some of the problems will then be presented.

Effect Size, Power, Statistical Significance

The power of meta-analysis to detect aggregate effects is generally quite sufficient to detect relatively small effects, if the number of studies and the number of subjects per condition in the studies are reasonably large (Hedges & Pigott, 2001); the omnibus test used by Marcus et al. (2014) to examine relative efficacy is also adequately powered (Wampold & Serlin, 2014). An untoward aspect of relatively high power is that small effects,

3 Mayo-Wilson et al. (2014) also analyzed recovery rates and the results were similar to the social anxiety symptoms. However, secondary measures were not coded or analyzed.

which are clinically unimportant, will be detected (i.e., found statistically significant). In all, 9 out of the 10 effects reported by Tolin (2014, 2015) and Marcus et al. (2014), critical to their case for the superiority of CBT (see Tables I and II), were below 0.20. Despite the relatively high power of meta-analysis, only 2 of the 10 comparisons were statistically significant.

The effects reported by Tolin and Marcus raise critical issues, the first of which is: How large should an effect be to alert the field that something important has been detected? An effect size of 0.20 is generally considered small and clinically unimportant. Given the variability within treatments, due to patient characteristics and therapist effectiveness, and the established contributions of various common factors such as the alliance and empathy, which produce effects (*d* equivalents) in the range of 0.50–0.75 (Norcross, 2011; Wampold & Imel, 2015), it is difficult to argue that a difference between treatments in the range of 0.20 establishes the superiority of a given treatment, even if such differences were statistically significant, which for Marcus et al. (2014) and Tolin (2014, 2015), they were not.

Therapist effects, well established empirically (Baldwin & Imel, 2013), lead to an overestimate of the true difference between treatments, even when therapist effects are quite small (Wampold & Serlin, 2000). This issue, which is generally unaddressed in primary studies, affects the meta-analysis of such effects, resulting in overestimates at the meta-analytic level and in liberal error rates (i.e., falsely rejecting the null hypothesis of no differences; Owen, Drinane, Idigo, & Valentine, 2015). That is, the effects detected in these meta-analyses, which are very small, are actually *inflated*.

A related issue is that the point null (in this case, that the difference between treatments is zero) is most certainly false and a very small effect can be detected with sufficient power (see Meehl, 1967, 1978). But the converse is also problematic: Studies with low power may well fail to detect true differences, even relatively large ones. Relying solely on statistical significance of effects (i.e., ignoring power and effect sizes) will create a paradox, where a small effect, detected in well powered study, is used to justify a claim, whereas a much larger effect, undetected in an underpowered study, is ignored.

To illustrate, consider a test of whether prolonged exposure (PE) exacerbates post-traumatic stress disorder (PTSD) symptoms. In an investigation of this conjecture, Foa, Zoellner, Feeny, Hembree, and Alvarez-Conrad (2002) used the results of a study of female victims of assault, who were randomly assigned to PE or to PE combined with cognitive restructuring (PE/CR). In the PE conditions, prolonged exposures were introduced in session 3 whereas they were introduced in the PE/CT condition in session 4, permitting an examination of deterioration from sessions 3 to 4 in the two conditions—if PE caused deterioration, then patients who received exposure in session 3 (i.e., those in the PE condition) would have shown more ill effects after session 3 than patients who had not received exposure in session 3 (i.e., those in PE/ CT). Effects were assessed on several dimensions, including PTSD symptoms, general anxiety, and depression. Deterioration was determined by a reliable worsening of symptoms. Patients in the PE did indeed demonstrate more deterioration between sessions 3 and 4 than did patients in PE/CT, although in most cases the results were not statistically significant. As well, those who did deteriorate also showed poorer final outcomes, although again, the result was not statistically significant. Foa et al. claimed, "The results of the present study are reassuring about the tolerability of exposure treatment for clients with chronic PTSD" (2002, p. 1026). Although Foa et al. claimed that there was adequate power to detect a medium effect (power in excess of 0.64 with alpha equal to.05), we calculated the effects for all of the tests conducted by Foa et al., converted them to *d*, and found that they ranged from 0.37 to

(mean effect = 0.44). *The problem is evident*: An effect in the neighborhood of 0.20 is used as evidence that CBT is superior to other treatments, but effects in range of 0.44 do not signal an issue for the harm caused by PE. Certainly, Marcus et al. (2014) and Tolin (2014, 2015) cannot be responsible for the claims made by Foa et al. However, scientific results are the product of a community of scientists who are bound to have standards that apply across researchers and instances and remain unchanged over the course of an investigation of a phenomenon (Lakatos, 1970; Lakatos & Musgrave, 1970; Larvor, 1998). The conclusion that an effect of .20 established the superiority of CBT requires the conclusion that PE be deemed harmful (an effect >.40); conversely, if PE is deemed not harmful then it must be concluded that CBT is not superior to other treatments.[4]

A third issue is related to statistical significance as it applies to error rates. A well-known threat to validity is what has been called fishing and error rate threats (Shadish, Cook, & Campbell, 2002). Protection of experiment-wise error rates is critical, as meta-analyses are vulnerable to false rejection of the null (Type I errors) in the same manner as primary studies. Efforts need to be made to control error rates in some way, particularly when one is *looking* for a particular result. Protection of error rates is accomplished by one of two strategies. First, an omnibus test can be conducted—if significant, *post hoc* tests of various types can be conducted. Second, in lieu of an omnibus test planned comparisons can be utilized. Regardless of whether planned comparisons or *post hoc* comparisons

4 Many would say the threshold for harm should be lower that it is for benefits. The risk of failing to claim that one treatment is not more effective when indeed one treatment is truly more effective than another is less than the risk of failing to claim a treatment is harmful when indeed it is harmful (i.e., "First, do no harm").

are examined, corrections to error rates must be made to ensure that the overall error rate is not inflated. Examination of the number of tests conducted in each of the three meta-analyses reveals that each meta-analysis performed an extraordinarily large number of tests.

The problem associated with failing to control error rates can be illustrated by examining the analytic method of network meta-analyses, as illustrated by Cipriani et al. (2009), who used network meta-analyses to compare the relative efficacy of 12 new-generation antidepressants. In their analysis, each pairwise comparison was examined, yielding $(k)(k-1)/2 = 66$ comparisons, of which several comparisons were significant, leading to the conclusion that some antidepressants were more (or less) effective than others. This conclusion was criticized on a number of grounds related to biases inherent in network meta-analyses (Del Re, Spielmans, Flückiger, & Wampold, 2013; Trinquart, Abbé, & Ravaud, 2012; Trinquart, Chatellier, & Ravaud, 2012), but here only the error inflation problem is examined. Del Re et al. (2013) created simulated data sets using the parameters of the Cipriani et al. trials under the null where no treatment differences existed among the antidepressants. They found that under the null hypothesis of no differences, 70% of the time at least one false statistically significant treatment difference would be detected (i.e., 70% of the time, it would be found that one or more drugs would be declared superior to another when there were absolutely no differences between any of the drugs). In about two-thirds of the cases, three or more differences were falsely detected. Clearly, when examining pairwise comparisons, error rates in network meta-analyses are a problem.[5]

One appropriate means for addressing the error rate problem would be to conduct an omnibus test of the null hypothesis that there are no differences among treatments. This is exactly the hypothesis tested by the methods developed by Wampold (Wampold, Mondin, Moody, Stich, et al., 1997 Wampold & Serlin, 2014) and used by Marcus et al. (2014). If the null cannot be rejected, then pairwise comparisons should not be examined. And indeed, in Cipriani et al. (2009) the null hypothesis that all second-generation antidepressants are equally effective could not be rejected (Del Re et al., 2013; Wampold & Serlin, 2014), nullifying the conclusions of the Cipriani network meta-analysis Even when the omnibus null is rejected, adjusted error rates must still be used (i.e., one cannot test all pairwise comparisons at .05 even if the omnibus null is rejected).

We conducted an omnibus test of the psychological treatments for Social Phobia examined by Mayo-Wilson et al. (2014) referenced in Table III, omitting the category Other, as it contained treatments that were designed as "intent-to-fail" treatments (Westen, Novotny, & Thompson-Brenner, 2004), as discussed below. Effect sizes between all 21 pairwise comparisons of the 7 treatments were provided by Mayo-Wilson and the variances of these estimates were derived from the credible intervals. Using these estimates it was found that the null hypothesis of no differences could not be rejected ($W = 26.36$, which when compared to a chi-square distribution with 21 degrees of freedom was not statistically significant, $p = .19$). *That is, there is no evidence that the differences between classes of psychological treatments for social anxiety, including psychological placebos, are other than zero and consequently it does not make sense to examine* post hoc *pairwise differences in the manner of Mayo-Wilson*. In this analysis the upper bound for the differences between classes of treatment was 0.23, a small effect in line with estimates produced by Wampold, Mondin, Moody, Stich, et al. (1997) and Marcus (2014).

Based on the above noted concerns about effect sizes and error rates, three recommendations can be made regarding meta-analyses aimed at estimating the relative effectiveness of psychotherapy approaches:

(1) The scientific community needs to stipulate what is a clinically meaningful effect—and use that standard when making conclusions.
(2) Meta-analysts must preserve error rates within an analysis. To the extent possible, meta-analytic hypotheses should be focused on crucial conjectures or important clinical questions (Rosnow & Rosenthal, 1988). Prior to conducting pairwise tests, where possible, the omnibus null should be tested. In any event, as Matt and Cook (2009) note with regard to meta-analysis, "To reduce capitalizing on chance, researchers must adjust error rates, examine families of hypotheses in multivariate analyses, or stick to a small number of a priori hypotheses" (p. 545).
(3) Rather than testing the point null (effects are zero), a more scientific and valid way to proceed would be adopt a non-inferiority strategy, where one stipulates a priori how large a difference would be meaningful and then test a range null hypothesis (i.e., the true value for the differences is in the stipulated range; see Minami, Serlin, Wampold, Kircher, & Brown, 2008; Serlin & Lapsley, 1985, 1993).[6]

5 Error rates were clearly a problem in Tolin (2014, 2015), who in his search for a signal, conducted in the neighborhood of 40 statistical tests. Although none of the primary contrasts between CBT and other treatments were significant (see Table I), more than 20 tests of various other contrasts were conducted.

6 At present, a range null strategy for meta-analysis has not been devised, although such a test should not be difficult to fashion, a test being investigated by the first author.

Disorder-Specific Symptom Measures

Recall that Marcus et al. (2014) and Tolin (2014, 2015) segregated outcome measures into two categories: primary, defined essentially as disorder-specific symptom measures, and secondary, a class that contains all other measures including symptom measures for disorders other than that being targeted, well-being, quality-of-life, and any other measures of general mental health or distress. Recall also no differences between CBT and other treatments were detected when these secondary variables were examined (see Tables I and II). Rather, only small effects were detected for the primary measures.[7] To be clear, *none of the three meta-analyses claiming CBT superiority reviewed here detected any effects for outcomes other than for the primary targeted symptoms*.

Tolin (2014) argued that psychological treatments should be evaluated exclusively with primary measures as the treatments are intended to be remedial for particular disorders. Examination of outcomes other than the targeted symptoms, according to Tolin (2014), is "rather unique" (p. 353). For the reasons discussed below, the present authors disagree, contending that it is essential to consider, and perhaps emphasize, outcomes other than targeted symptoms when evaluating treatment approaches.

The first issue is that the emphasis in clinical research on psychological treatments for particular disorders, which goes back to the beginning of the empirically supported treatment movement (Chambless & Hollon, 1998; see also Wampold & Imel, 2015), ignores some crucial facts about psychopathology and treatment. Most importantly, comorbidities are typical. Patients with the most prevalent and disabling mental disorders also meet criteria of multiple diagnoses. Between 84% and 97% of patients reporting the symptoms of one disorder qualify for at least one other disorder (Gadermann, Alonso, Vilagut, Zaslavsky, & Kessler, 2012). As one example, motivated by Mayo-Wilson et al.'s (2014) focus on the reduction of symptoms for social anxiety, over 90% of individuals qualifying for a social phobia diagnosis qualified for another diagnosis and the mean number of other diagnoses for people with social phobia was 3.5 (Gadermann et al., 2012). Moreover the disease burden of those with mental disorders is due, to a large extent, to comorbidities and not simply to the additive effects of having more than one disorder. Based on results from the National Comorbidity Survey Replication, it was concluded,

> These results underscore the importance of including information about comorbidity in studies of burden … [and] arguing against a focus on pure disorders in epidemiological studies designed to evaluate the effects of mental disorders on functioning as well as in studies designed to evaluate the effects of treatment in reducing the impairments associated with mental disorders. (Gadermann et al., 2012, p. 84)

A second point that should temper enthusiasm for meta-analyses of disorder-specific measures is related to the problems with the nosology for diagnosis of mental disorders (DSM or alternatives), which are pervasive, as most of us are well aware (see e.g., Greenberg, 2013; Zachar, 2015). Alternative ways to conceptualize, understand, and treat mental disorders are gaining scientific attention, including Research Domain Criteria (Lilienfeld, 2014) and trans-diagnostic treatments (Barlow et al., 2011).

The first two issues discussed here, comorbidity and problems with nosology, raise the issue about higher order factors underlying psychopathology. Based on an extensive longitudinal data from the Dunedin Multidisciplinary Health and Development Study, Caspi et al. (2014) found that a general psychopathology factor explained all psychiatric disorders. Termed the *p factor*, the authors reported,

> … evidence pointing to one general underlying dimension that summarized individuals' propensity to develop any and all forms of common psychopathologies … Higher scores on this dimension were associated with more life impairment, greater familiality, worse developmental histories, and more compromised early-life brain function. (p. 131)

Given that underlying factors might well lead to the development of multiple disorders as well as the prevalence of comorbidity, it is reasonable to suggest that symptoms for one identified disorder may be what in medicine are called *surrogate endpoints*. Surrogate endpoints are outcomes that correlate with clinically important outcomes but are used to substitute for meaningful health outcomes (Psaty, Weiss, & Furberg, 1999). Blood pressure would be a surrogate endpoint in a trial where reduction in blood pressure substitutes for measures of mortality or cardiac morbidity. Use of surrogate outcomes has often obscured the efficacy and risk of medical treatments:

> Surrogate end points sometimes fail to serve as valid predictors of important health outcomes … Drug therapies usually have multiple effects, and resorting to a single surrogate end point that focused exclusively on 1 intermediate effect often precludes the evaluation of other intended or unintended health effects.[8] (Psaty et al., 1999, pp. 786, 787)

If reduction in symptoms of one specific disorder does not also increase quality of life, well-being, interpersonal relations,

7 Interestingly a similar pattern of results (viz., no differences for secondary measures and very small differences for primary measures) were found for dismantling studies (Bell, Marcus, & Goodlad, 2013).

8 Of course, there are differences in disorder-specific symptom measures and many surrogates in medicine. In psychiatric disorders the symptoms are typically distressing to the patient whereas in medicine surrogates can be asymptomatic risk factors, such as hypertension or elevated cholesterol levels in cardiac disease.

and ability to work and function in society, then such symptoms may well be surrogate measures.

The focus on disorder-specific symptoms ignores what is known about psychopathology, epidemiology of mental disorders, the burden of disorders, and clinical reality. To make the claim that CBT is superior to other treatments only on disorder-specific symptoms essentially states that, as Marcus et al. (2014) discussed, attempts to reduce particular symptoms can be successful but have little, if any effect relative to other treatments on relieving the burden of mental disorders. As an example, patients may find a reduction in tics for Tourette Syndrome with a treatment solely focused on the tics, relative to a less focused treatment. However, most patients seek a reduction in the burden of their disorder, which is often captured by non-disorder-specific measures.

Efforts to identify specific treatments for specific disorders underscores the importance of a principle clearly enunciated by Jerome Frank over 50 years ago. To wit, the success of psychotherapy depends on the efforts the patient makes to address particular problematic areas in one's life (Frank & Frank, 1991; Wampold & Imel, 2015). Indeed, the degree to which gains in functioning are attributed to the patient's own efforts leads to sustained benefits (Liberman, 1978; Powers, Smits, Whitley, Bystritsky, & Telch, 2008). Unstructured treatments—that is, therapies without actions the patient believes are associated with directly overcoming particular difficulties—have little power to change the focal problems, an observation that is predicted by common factor theories and is supported by the evidence (Wampold & Imel, 2015). Here, it is important to remember, CBT is not the only approach that focuses on particular problems (e.g., short term dynamic therapies for particular disorders).

Based on the foregoing issues related to reliance on symptom-specific measures, two recommendations can be made regarding meta-analyses aimed at assessing the outcome of various psychotherapies:

(1) Clinical trials of psychological treatments should measure outcomes related to broad categories of symptoms, well-being, life functioning, and quality of life as well as symptoms related to the primary diagnosis.[9]
(2) Meta-analyses should analyze and report effects for the various categories of outcomes as well as primary measures. Meta-analyses, such as Mayo-Wilson et al. (2014) that examine *only* targeted measures can be, and are likely to be, misleading and perhaps clinically unimportant.

Classifying Treatments—What Is CBT?

Claiming that CBT is superior to other treatments requires that the essential properties of the approach are known—a task that has proven difficult, if not impossible, to achieve. In their response to Tolin (2010), Baardseth et al. (2013) made the following observation:

> Critical to the proposition that CBT is superior to other treatments is the taxon CBT. What is CBT? What are its essential features? What is the definition of CBT? Although, as Lakatos observed, concepts and taxons can be and are altered as science progresses, they should be done so on a rational basis—in a way that clarifies rather than on an ad hoc basis to protect the hard core of a research program. As Larvor (1998, p. 19), in his commentary on Lakatos, "Nevertheless those meanings (whatever they may be) must remain fixed from one end of the argument to the other." When statements are made about the superiority of CBT, the nature of the taxon CBT either has to be fixed, or altered in a rational way—that is, in a way that clarifies the essential nature of the concept. (p. 402)

A simple reading of the literature shows the definition of CBT, and the classification of particular treatments as CBT, has and continues to vary considerably from one meta-analysis to another. The resulting impact on interpreting results have been discussed at some length (Baardseth et al., 2013; Tolin, 2014; Wampold, 2013a; Wampold et al., 2010) elsewhere and are not rehashed here. What will be discussed are the inconsistencies among meta-analyses, focusing on a particular problem in Mayo-Wilson et al. (2014).

Mayo-Wilson et al. (2014) made a distinction between CBT and other closely related treatments, including exposure, applied relaxation, social skills training (SST), and mindfulness based treatments. Tolin (2010) used a much more inclusive definition. According to Tolin et al. (2010), a treatment was CBT if it contained *any* of the following components: relaxation training (including progressive muscle relaxation, meditation, or breathing retraining), exposure therapy (imaginal or *in vivo* exposure, including flooding and implosive therapy), behavior rehearsal (behavioral training in social skills, habit reversal, or problem solving), cognitive restructuring (including direct strategies to identify and alter maladaptive thought processes), or operant procedures (systematic manipulation of reinforcers or punishers for behavior, including behavioral activation). In short, many treatments deemed *not* CBT by Mayo-Wilson would have been classified as CBT by Tolin et al. (2010). More than a problem about classification, the inability to agree

9 We are well aware that researchers are required to designate the primary outcome for clinical trials by various governmental agencies (see clinicaltrials.gov), in an effort to reduce Type I Errors. We endorse the intent of such efforts—indeed, the purpose of this review is to point out that through various means an omnibus Type I Error has been committed (viz., claiming that CBT is superior to other treatments). Nevertheless, conclusions restricted to a primary measure, which might make sense in medicine (but see the literature on surrogate endpoints), are problematic for patients seeking psychotherapy.

on what constitutes CBT goes to the heart of any conclusions regarding its superiority.

The exclusion of various treatments in the class of CBT by Mayo-Wilson is even more troublesome because many of the CBT treatments examined were actually CBT combined with another treatment, which was not classified as CBT. For example, Herbert et al. (2005) compared group CBT to group CBT augmented by SST and found that patients receiving the latter combined treatment performed significantly better than those in the CBT condition alone. Yet, for Mayo-Wilson et al. both treatments were CBT and the advantage of the SST component was ignored in the meta-analysis. Social skills training in this study provided an advantage to a CBT, yet when a treatment contained only SST it was classified as something other than CBT. It is troublesome when the meta-analysis found that CBT was superior to SST when SST added to CBT outperformed CBT. Similarly, Cottraux et al. (2000) compared a CBT that emphasized social skill training to Supportive Therapy (ST), yet the CBT plus social skills was classified as CBT rather than as SST. Most, if not all of the CBT treatments for social anxiety had exposure elements, although in Mayo-Wilson et al. there was another class for treatments that were based on exposure (viz., Exposure and Social Skill Category; EXP/SST). In yet another example, Alden and Taylor (2011) combined CBT with interpersonal therapy (IPT), but this treatment was classified as CBT, whereas IPT was classified as OTHER. As defined by Mayo-Wilson et al. (2014), CBT perhaps is best characterized as an integrative treatment, and their conclusions should be modified to say that integrative treatments, with a cognitive component, are recommended for social phobia.

Clearly, the definition of CBT is quite expansive. A consequence of this is that two exemplars of CBT may have little in common (see Baardseth et al., 2013 for a more complete discussion of this issue). To illustrate, consider the ingredients intentionally *excluded* in the CBT protocol used by Clark et al. (2006): repeated exposures designed to create habituation, exposure hierarchies, patient assessment of anxiety or thoughts in social situations that are feared, employment of self-instruction (i.e., rational thoughts) in social situations, or SST. The very ingredients excluded by Clark et al. (2006) are included, and often are the essential ingredients, in most of the CBT treatments for social phobia in the Mayo-Wilson et al. meta-analysis, rendering conclusions about the superiority of CBT ambiguous (which CBT is treatment of choice?), if not nonsensical.

The problems with classification in Mayo-Wilson et al. (2014) can be demonstrated empirically. Clearly, the set of CBT treatments is quite diverse, with the various treatments containing many elements not purely *cognitive* in nature and others explicitly excluding behavioral components; several CBT treatments had nothing in common with each other. We sought to examine the heterogeneity of these treatments by examining the outcomes produced by the various CBTs. Accordingly we looked at all direct comparisons of various CBT treatments within the Mayo-Wilson et al. (2014) network meta-analysis.[10] When we examined these nine comparisons, the outcomes produced by CBT were heterogeneous (Using Wampold & Serlin's, 2014 test, $W = 28.54$, $df = 9$, $p = .00+$). That is, there is evidence that some CBT treatments are superior to others, which creates issues about the conclusion of superiority of the general class of CBT for social phobia.

The heterogeneity within the class of treatments called CBT raises questions about what is the essence of CBT. It makes little sense to talk about a unified class of treatments when one or more of the treatments in the class are more effective than another. As well, given that many of the CBT treatments were actually combined treatments, this evidence supports that some of the components added may lead to improvement, at least for targeted measures (see Bell et al., 2013)—that is to say, the components that were isolated in other classes and were at some disadvantage methodologically speaking may be quite important ingredients.

The final point, which is critical to understanding the limitations of network meta-analysis for psychotherapy, is that the treatments within classes (called nodes in network meta-analysis) of a network meta-analysis are assumed to be interchangeable. For example, if two studies of Treatment X are included in the meta-analysis, then the treatment employed in those studies are considered identical. For example, for two studies of fluoxetine, the medication is invariant (both are *N*-methyl-3-phenyl-3-[4-(trifluoromethyl)phenoxy]propan-1-amines) and it is assumed they work exactly the same. Although there may be some study level aspects that change (called effect modifiers in the network meta-analysis literature—see Jansen & Naci, 2013), the treatments themselves are identical. In the Mayo-Wilson et al. network meta-analysis the treatments themselves vary considerably, both in terms of what they contain and the effects they produce. As Cipriani et al. (2013) noted, "Arguments exist for giving priority to direct evidence because it does not rely on the transitivity assumption" (p. 134).

A critical step in the scientific understanding involves classification of objects. Correct conclusions depend on the objects

10 We did not differentiate whether or not the treatments were group administered, as there were no statistically significant differences between the two modalities. Interestingly, Mayo-Wilson differentiated group and individual for CBT but inexplicably did not make that distinction for other treatments; e.g., group PD therapy (see Knijnik, Kapczinski, Chachamovich, Margis, & Eizirik, 2004) was classified together with individual PD. Moreover in our analysis the largest effects were within modality (between two group CBTs and between two individual CBTs).

being categorized on their essential characteristics rather than superfluous ones (in the philosophy of science, the discussion is about "natural kinds"; see Boyer, 1990; Lakatos, 1976; Lakatos & Musgrave, 1970; Larvor, 1998). When we talk about antidepressants, the exact chemical structure of the drug is known; we know that fluoxetine, an SSRI, and alprazolam, a benzodiazepine, are different. We also know that when two patients ingest 20 mg of fluoxetine, they have received the same treatment, although there may be different effects due to metabolic and neurological differences. Psychotherapy is different—it has no physical form and exists as an idea, in a manual guiding treatment, or in the head of the psychotherapist. Psychotherapy only becomes *real* when it unfolds during the course of therapy (see Imel, Steyvers, & Atkins, 2015 for extended discussion of these issues). All psychotherapies, even the most constrained and manualized treatment, unfold differently in each instance, due to characteristics of the therapists (Baldwin & Imel, 2013) and the patient (Boswell et al., 2013; Imel, Baer, Martino, Ball, & Carroll, 2011). In short, care must be taken when talking—and doing research—not to treat psychotherapy as if it were a physical object (i.e., a natural kind). It is not. At best, treatment approaches are fuzzy concepts. When researchers allow categories of treatment (e.g., CBT) to vary from one study to another and from one meta-analysis to another, confusion is generated.

Based on the foregoing issues related to the classification of treatment approaches into categories, three recommendations are made:

(1) There needs to be agreement on what is and what is not CBT, an observation that applies to other treatments as well.
(2) As recommendation #1 is easier said than done, the field needs to identify the ingredients of psychotherapy responsible for change (i.e., what does and does not contribute to change). The ingredients may or may not be what the field has commonly used to classify treatments.
(3) Care must be taken not to reify models of psychotherapy. Within categories such as CBT, many variations exist (e.g., as in Mayo-Wilson et al., 2014) and the manner in which a treatment is delivered depends on the therapist, the patient, and their relationship, as well as external events.

The Studies—Included and Excluded

To this point, it has been argued that the purported superiority of CBT has been based on very small effects, derived from primary measures only, using varying classification strategies. Behind the reports are the individual studies that form the data corpus of the meta-analyses. The conclusions drawn depend on qualities of the primary studies and, importantly, the criteria for inclusion or exclusion based on these qualities. In this section, fundamental problems in the meta-analyses are presented and discussed by examining differences in how studies were either selected or rejected.

Poorly designed included trials. First, consider, a study conducted by Cottraux et al. (2000), which appeared in the Mayo-Wilson et al. (2015) meta-analysis. This study produced the largest effect for a direct comparison between Individual CBT and Other Psychotherapies ($d = 1.15$ for social phobia measures). Recall, in Mayo-Wilson et al., Other Psychotherapies performed more poorly than Psychological Placebos. The reason is clear. The CBT condition (coded, Individual CBT by Mayo-Wilson et al.) was actually a combination of CBT and SST (recall, SST was intentionally not part of CBT in the Clark et al., 2006 protocol for social phobia). The CBT/SST condition had two phases. In the first phase, patients received eight individual sessions of cognitive therapy, which included receipt of a monograph about the treatment, psychoeducation, thought listing and evaluation, modification of maladaptive thoughts and schemas, homework, and preparation for the second phase. In the second phase, patients attended group sessions of 2 hr duration once a week for 6 weeks. The groups, led by two therapists, involved role plays of social situations with feedback, behavioral rehearsals of difficult social skills with feedback, coaching, and modeling, assignments for practice outside of group, and strategies for generalization (n.b., again, most of these components were excluded from the CBT in the Clark et al., 2006, protocol). In all, patients received 20 hr of direct contact over 12 weeks, as well as assignments to complete outside of the treatment time.

The comparison condition in the Cottraux et al. (2000) trial, labeled Supportive Therapy (ST), involved one 30-min session every 2 weeks for the 12 weeks of the trial, resulting in a total direct contact of 3 hr. During the 30 min sessions the therapist was prohibited from giving any advice, homework, any action that might expose the patient to avoided situations, psychoanalytic interpretations, or cognitive restructuring, although the therapist was allowed to listen empathically, reformulate, clarify, summarize, and show "positive consideration" (p. 138). Cottraux et al. stated that ST was "practiced '*as usual*" (p. 138, emphasis added) in France—a statement that is in direct contradiction with several of the present authors' experience with non-CBT therapists practicing in France and other therapists who treat people with social phobia. Nevertheless, ST was classified as a "first-line" Other Psychotherapy in the Mayo-Wilson et al. analysis. Is it any surprise in this trial that CBT outperformed Other Psychotherapies (or why Other Psychotherapies were inferior even to Psychological Placebos!)?

The issue in network meta-analysis is that the extraordinarily large and problematic effect derived from Cottraux et al. (2000) is used as an indirect path that increases the effect of CBT over every other treatment. In the Cottraux study CBT outperformed Other Psychotherapy by an effect of 1.15. To

illustrate, suppose a PD researcher compared PD therapy to IPT, which was also classified as Other Psychotherapy, with a legitimate implementation of IPT (e.g., same dose, a legitimate set of procedures, a focus on the patients' problems, conducted by therapists who believed IPT would be effective etc.) and found that PD = IPT. Network meta-analysis would then estimate, through the transitive property, that CBT was superior to PD by an effect of 1.15! That is, CBT would clearly be found to be superior to PD even though CBT was never compared to PD (in this example) but only compared to a bogus treatment called ST, which was also classified as Other Psychotherapy. Cottroaux's biased comparison privileged CBT over all other types of therapy even though the comparison was made to only one type of (bogus) therapy.

A second study, which appeared in both the Mayo-Wilson et al. (2014) and Marcus et al. meta-analyses, to consider is Clark et al. (2006), which compared cognitive therapy (CT), classified as Individual CBT in both meta-analyses, with Exposure/Applied Relaxation (EX/AP), and a waitlist control (WL). In this study, CT was superior to EX/AP with a large effect ($d = 0.87$) and CT was superior to the waitlist with an extraordinarily large effect ($d = 1.85$). However, as was the case with Cottraux et al. (2000), this trial was problematic. David Clark, the lead author, both developed the CT and supervised the therapists. The therapists who administered both treatments had an allegiance to CT (viz., based on their CT publications). In this study, there are indicators of both researcher and therapist allegiance, well-known problems in psychotherapy trials (Munder, Brütsch, Leonhart, Gerger, & Barth, 2013; Munder, Flückiger, Gerger, Wampold, & Barth, 2012; Munder, Gerger, Trelle, & Barth, 2011). The major problem with Clark et al. (2006), however, is the nature of the EX/AR condition. Clark et al. reviewed the literature on CT versus exposure treatments, and concluded, "Existing comparisons between exposure and other established CBT programs have failed to show convincing differences" (p. 569). However, instead of using one of the existing and well researched exposure treatments that were cited, Clark et al. used a peculiar combination of various methods. The stated reason for this decision was to avoid a purported problem with dropout in exposure treatment: "We attempted to minimize EXP dropouts by combining the treatment with Öst's (1987) well-known applied relaxation (AR) training program" (p. 569). As evidence of the low dropout rate, Clark et al. (2006) cited his own 1994 trial (Clark et al., 1994) of CT versus EXP/AR, which allegedly resulted in low dropout.[11] What Clark et al. (2006) did not mention when referencing this earlier work was that *the combined EXP/AR treatment used in the 2006 trial was found to be particularly ineffective in the 1994 trial*. Thus, not only was the EXP/AR an intent-to-fail treatment, it was a *proven-to-fail* treatment (see Wampold, Imel, & Miller, 2009 for another critique of the EXP/AR protocol). There is a particular reason why the EXP/AR combination used in 1994 and 2006 by Clark et al. failed to be as effective as Clark's CT conditions.[12] CT has a number of components, including experiential exercises designed to demonstrate the adverse effects of self-focused attention and safety behaviors, systematic training of externally focused attention, techniques for restructuring distorted self-imagery using video feedback in particularly structured situations, surveys to collect data on others peoples' beliefs, and carefully planned exposures to feared social situations, with instructions not to use habitual safety behaviors. The EXP/AR consisted of two components, graduated exposures and relaxation training. Patients were encouraged not to avoid situations they would normally avoid and in-session exposure focused on *in vivo* exercises rather than role plays within the therapy. Importantly, the relaxation protocol used by Clark was adapted from Öst (1987), however, with some critical adaptations:

> In the original AR protocol (Öst, 1987), exposure is not introduced until after the relaxation techniques h ave been fully mastered. We deviated from this practice by using exposure exercises throughout treatment. However, as advocated by Öst (1987), patients were instructed to refrain from using their newly acquired relaxation techniques in phobic situations until they had completed all the steps in the relaxation training program (around Session 10). (Clark et al., 2006, p. 571)

This turns Öst's (1987) treatment inside out: In Clark's incarnation, patients are exposed to anxiety producing situations before they have learned any skills for coping with the situation and indeed were instructed *not* to use the skills they might have learned. Furthermore, according to Öst's protocol, "After 8–10 sessions and weeks of homework practice the patient is ready to start applying the relaxation skill in natural situations to cope with anxiety" (p. 401). It is contradictory to behavioral principles to expose patients to the feared stimulus before they have learned coping skills and further to instructs them not to use skills they are learning to cope with the anxiety. Clearly, one can increase avoidance by increasing the frequency of the conditioned stimulus (social situations) paired with conditioned response (fear) without any strategies for reducing the fear.

The modified Öst protocol used in these two trials (viz., Clark et al., 1994, 2006) has never been tested or used in any other

11 Dropout rate is reported ambiguously in Clark et al. (1994) but is likely low because only patients who started treatment and attended one or two sessions were classified as drop-outs—that is, if they attended three of more sessions of a 15 session treatment they were classified as not having dropped out of treatment.

12 It is also important to note that one of the therapists in the 2006 trial was also a therapist and co-author of the 1994 trial, so certainly this therapist was well aware of the problems with EXP/AR.

context. The impact of the Clark et al. (2006) trial on the findings and conclusions is critical. Indeed, several of the effects in Marcus et al. (2014) disappear when this trial is omitted.[13]

A third trial, which produced the second largest effect for CBT versus other therapies, in Tolin's (2014) meta-analysis, as corrected in 2015, was a trial conducted by Shear, Houck Greeno, and Masters (2001). Although titled "Emotion-Focused Psychotherapy for Patients with Panic Disorder," the treatment actually offered bears no resemblance to Emotion-Focused Therapy developed and disseminated by Leslie Greenberg and colleagues (Greenberg, 2010). In fact, it can be said, the treatment offered in Shear et al. (2001) is unlike any other particular treatment designed for panic or any other disorder. Instead, the Emotion-Focused therapy tested in the study was best characterized by what it was not: "Emotion-focused psychotherapy was not a psychoanalytic psychotherapy in that the therapist did not utilize transference and did not formulate or provide psychodynamic interpretations." Still, Shear et al. believed it "bears resemblance to … usual-care psychotherapy" (p. 1994).[14] Further problems occurred because of non-random assignment, as the original trial could not enroll sufficient patients in the Emotion-Focused arm and accepted patients in this arm who refused to discontinue medication or who refused to be randomly assigned to condition.

Clearly, the Emotion-Focused Therapy in this trial was not designed as a treatment that had any rationale for its success other than the authors believed that therapists in practice did something similar with panic patients.

To this point, we have examined three trials that demonstrated strong evidence for superiority of CBT (i.e., large effects), each involving a deficient comparison treatment. In the first trial, 20 hr of CBT with homework was compared to 3 hr of ST without any structure. In the second, CBT was compared to treatment where patients were exposed to fearful stimuli before they had learned anxiety coping skills as well as being told not to use the skills in anxiety-provoking situations. In the third trial, CBT was compared to an emotion-focused treatment that purposefully was different from PD treatments, bore no resemblance to any known affect-focused treatment, and was not designed to treat panic disorder. A fourth trial, which was conducted by David Clark and was contained in the Mayo-Wilson et al. (2014) meta-analysis, was an unpublished trial. This trial produced an extraordinarily large effect in favor of Individual CBT in comparison to waitlist controls (viz., $d = 1.63$). Unfortunately, this trial could not be evaluated because the author of the trial would not provide the report of the trial to this article's authors (Clark, personal communication, February 20, 2015).

There appears to be a bias in what has been published and included in meta-analyses claiming superiority of CBT. Turning the table, would any of the following trials be published in mainline clinical journals or included in meta-analyses? (a) A trial with 3 hr of CBT versus 20 hr of PD therapy, (b) comparison of Greenberg's Emotion-Focused Therapy versus a CBT condition involving what the researchers believed CBT therapists did in practice, (c) a comparison of a focused PD therapy for a particular avoidant anxiety condition versus CBT where patients were exposed to fearful situations before they learned about misattributions and were told not to apply what they learned about their cognitions in fearful situations, or (d) an unpublished study purporting to show one method is superior to another, the results of which favored the treatment developed by the author and included in a meta-analysis co-authored by the developer.

Excluded studies. As all of the authors of this article have conducted meta-analyses, we know the anxiety associated with the possibility of omitting a study that would have met inclusion criteria. And occasionally an interested reader finds a prominent study omitted. Indeed, Marcus et al. (2014) missed just such a study that met all of their inclusion criteria, including that it was published in one of the six journals they reviewed (viz., *Journal of Consulting and Clinical Psychology*). The omitted trial compared CBT to Process-Experiential Therapy (now called Emotion-Focused Therapy) for depression, which interestingly found no differences between the two treatments, with the exception of an advantage for Process-Experiential Therapy with regard self-report of interpersonal problems (Watson, Gordon, Stermac, Kalogerakos, & Steckley, 2003).

The omission of another study is more difficult to explain. In 2008, Borge et al. reported the results of a comparison of CBT to IPT for social anxiety in a residential treatment context. No significant differences between the two treatments were found. This trial could not have been unintentionally omitted from the Mayo-Wilson et al. (2014) meta-analysis because David Clark was a coauthor of both the CBT/IPT trial (as well as a supervisor of the CBT therapists in the trial) and the Mayo-Wilson et al. meta-analysis.

The omission of Borge et al. (2008) should be put into the context of the inclusion criteria for the Mayo-Wilson meta-analysis. To be included, a treatment had to be a *first-line* treatment, as discussed by Mayo-Wilson et al.:

13 After several requests, David Clark failed to provide the EXP/AR manual, so it is not possible to know exactly what treatment components were contained in EXP/AR and how they were sequenced. Of particular interest is whether the therapists in this condition were proscribed from various actions that therapists would universally believe to be therapeutic.

14 The authors stated, "Strategies and techniques for interventions were outlined in a detailed treatment manual" (pp. 1993–1994), but according to the lead author the manual was not archived and could not be provided to the authors of this article.

> We limited the network meta-analysis to interventions that people with social anxiety disorder and clinicians might regard as first-line treatments because network analysis assumes that treatment effects are transferable across studies … Clinically, people choosing a first-line intervention have a different set of treatment options compared with people choosing second-line interventions; there would be a high risk that the assumption of exchangeability would be violated by the inclusion of clinically heterogeneous populations … We identified eligible interventions by reviewing published and unpublished studies and through consultation with clinicians and experts (including people with social anxiety disorder, pharmacists, psychologists, and psychiatrists). We included interventions rather than excluded them if some experts thought they could be used as a first-line treatment.[15] (pp. 369– 370)

The most problematic aspect of the inclusion criteria is that the determination of what is a first-line treatment is contorted and seemingly broad. Consider that the Mayo-Wilson meta-analysis included as first-line treatments (a) one that consisted of six 30 min sessions with a therapist who was only supportive, (b) a treatment (EXP/AR) invented by Clark et al. only as a comparison in clinical trials, never disseminated, the manual for which is not available, and which had previously been found to be ineffective for an anxiety disorder, and (c) conditions with virtual reality, and mindfulness only groups. Yet inexplicably the treatments (CBT and IPT) offered in the Borge et al. (2008) trial, albeit modified for residential treatment, were not even considered for this meta-analysis (viz., Borge did not appear in the list of excluded studies, see Appendix 6 of the Mayo-Wilson et al. Supplemental Materials).

Conclusions and recommendations for included/excluded studies. In this section, we have discussed the problems with including trials with treatments that are, as Westen (Westen et al., 2004) said, *intent-to-fail* and even some that are *proven-to-fail*. There are other trials that could be discussed as well (e.g., Durham et al., 1994, a notoriously poorly conducted trial) and some of these flawed trials are dated (Durham et al., 1994; Shear et al., 2001) and appear in multiple meta-analyses. Then there are trials that are inexplicably omitted (Borge et al., 2008; Watson et al., 2003)—and interestingly these omitted trials demonstrated no differences between CBT and comparison treatments. Conflict of interest and spin in meta-analyses of psychological treatment has been documented (Lieb, Osten-Sacken, Stoffers-Winterling, Reiss, & Barth, 2016).

15 Actually, the group of clinicians and experts were the National Collaborating Centre for Mental Health (NICE) Guideline Development Group for the guideline Social Anxiety Disorder, the chair of whom was David Clark. Clark was not involved directly in decisions about inclusion or exclusion of his research (Mayo-Wilson, personal communication, 9 September 2015).

Given the impact that inclusion criteria can have on the results of meta-analyses, the following recommendations are made:

(1) The problem of included and excluded studies could be addressed by creating open-access databases of studies, with effect sizes. Baldwin, Del, and Re (2016) have developed prototypes for open access to effects for family therapy for delinquency and alliance-outcome psychotherapy studies. The user can select studies based on criteria and conduct meta-analyses with imbedded software. The community of scientists can modify the databases to ensure that all qualifying studies are included and deficient studies can be identified, reducing the allegiance effect of meta-analysts.
(2) The allegiance of researchers of primary studies should be coded and analyzed.
(3) All manuals used in clinical trials should be available to meta-analysts who wish to know what treatment actions are prescribed and proscribed in treatments.
(4) Results from unpublished studies should not be included in meta-analyses unless treatment protocols and data from those trials are made available to researchers for review.

Discussion

The review of the three meta-analyses (viz., Marcus et al., 2014; Mayo-Wilson et al., 2014; Tolin, 2014, as corrected in 2015) claiming the superiority of CBT have illustrated some issues inherent in meta-analytic attempts to examine relative efficacy and establish the superiority of CBT to other treatments. Various concerns have been examined, including (a) effect size, power, and statistical significance, (b) focusing on disorder-specific symptom measures and ignoring other important indicators of psychological functioning, (c) problems inherent in classifying treatments provided in primary studies into classes of treatments, leaving the question "What is CBT?" unanswerable, and (d) the inclusion of problematic trials, which bias the results, and the exclusion of trails that fail to find differences among treatments. Due to space, a thorough discussion of other important issues in these meta-analyses was not possible, including an analyses of allegiance, whether the meta-analysts used completers or intent-to-treat samples, and dropout rates.

Science is conservative in that the null hypothesis should not be rejected unless there is strong evidence in favor of the alternative. This canon, which has its origins with Sir Ronald Fisher, is not simply a tradition being passed along, an anachronism, if you will. Rejecting the null hypothesis of no treatment differences in favor of one particular treatment has consequences for science, policy, and practice. We know that many patients drop out of trials and many of those who remain do

not benefit from the treatment—to claim that one treatment is superior to another will limit patients' access to other treatments that are equally effective and have a reasonable likelihood of being effective. As seen, the purported superiority of CBT has resulted in conclusions that research on other treatments, such as PD treatments, should be abandoned. If such admonishments were headed, then our scientific endeavors would become constrained to some narrow corridors, prohibiting scientists from discovering anything lying outside that corridor. When CBT is declared superior to other treatments, without carefully specifying what CBT is, the opportunity to discover what factors are actually creating the benefits of psychotherapy is precluded. When two CBT treatments, without any elements in common, produce adequate benefits, little is learned about what makes CBT an effective treatment. Finally, falsely declaring a treatment as superior leads policy makers to believe that they are acting in the best interests of patients and the mental health field when they mandate that only these treatments can be used. Such dissemination attempts, however well meaning, are costly and do not improve the quality of mental health services (Laska et al., 2014).

The purpose of this review was not to criticize CBT as a treatment. The evolution of CBT has dramatically affected how psychotherapy is delivered and has led to a revolution from long-term and relatively unstructured treatments (think classical psychoanalysis) to those that are focused on patients' problems, utilize psychoeducation and skill development, and have emphasized that psychotherapy should result in demonstrable and measurable outcomes. And to be clear, CBT is not the only treatment that exaggerates its benefits (e.g., claims that PD treatment produces superior outcomes in the long term; cf., Kivlighan et al., 2015 and Shedler, 2010). Moreover, making dubious claims about CBT harms its reputation among many researchers but more importantly among clinicians.

There is an important issue that has not been explicitly discussed. Are the conclusions of these three meta-analyses biased? There is certainly a case that could be made for that (see Lieb et al., 2016). Some operations (e.g., inclusion/exclusion of studies, determination of what are first-line treatments, refusal to provide treatment manuals, involvement of individuals who have vested interests in treatments) could be interpreted as evidence for bias. That said, no claim of intentional bias is being made and it would not be appropriate to do so, in our opinion. Science involves a community of researchers who, through dialogic processes, illuminate what is known and what is artifact. Meta-analysis is one of many tools used to better understand psychotherapy, but, like any analytic method, care must be taken to use it appropriately. The purpose of this article was to demonstrate problems in the conclusions made, based on the evidence only.

What makes psychotherapy work? As Kazdin notes (2007, 2009), this is the "most pressing question" (Kazdin, 2009, p. 418), but one with few answers: "Central … is the thesis that, with isolated exceptions, we do not know why or how therapies achieve therapeutic change, the requisite research to answer the question is rarely done, and fresh approaches are needed in conceptualization and research design" (p. 489). Claims of superiority of treatments of one type or another have not provided the evidence that is needed and, in our opinion, obscures important questions. Clearly, our agenda must change if we are to progress.

References

Alden, L. E., & Taylor, C. T. (2011). Relational treatment strategies increase social approach behaviors in patients with generalized social anxiety disorder. *Journal of Anxiety Disorders*, *25*(3), 309–318. doi:10.1016/j.janxdis.2010.10.003

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*(3), 395–405. doi:10.1016/j.cpr.2013.01.004

Baldwin, S. A., & Del Re, A. C. (2016). Open access meta-analysis for psychotherapy research. *Journal of Counseling Psychology*, *63*(3), 249–260. doi:10.1037/cou0000091

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Finding and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). New York, NY: Wiley.

Barlow, D. H., Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Allen, L. B., & Ehrenreich-May, J. (2011). *Unified protocol for transdiagnostic treatment of emotional disorders: Therapist guide*. New York, NY: Oxford University Press.

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*(4), 722– 736. doi:10.1037/a0033004

Borge, F.-M., Hoffart, A., Sexton, H., Clark, D. M., Markowitz, J. C., & McManus, F. (2008). Residential cognitive therapy versus residential interpersonal therapy for social phobia: A randomized clinical trial. *Journal of Anxiety Disorders*, *22*(6), 991–1010. doi:10.1016/j.janxdis.2007.10.002

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., … Barlow, D. H. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*, *81*(3), 443–454. doi:10. 1037/a0031437

Boyer, P. (1990). *Tradition as truth and communication: A cognitive description of traditional discourse*. New York, NY: Cambridge University Press.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., … Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119– 137. doi:10.1177/2167702613497473

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

Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P. T., Churchill, R., … Barbui, C. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: A multiple-treatments

meta-analysis. *The Lancet*, *373*(9665), 746–758. doi:10.1016/s0140–6736(09)60046–5

Cipriani, A., Higgins, J. P. T., Geddes, J. R., & Salanti, G. (2013). Conceptual and technical challenges in network meta-analysis. *Annals of Internal Medicine*, *159*(2), 130–137. doi:10.7326/ 0003–4819–159–2-201307160–00008

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* (3), 568–578. doi:10.1037/0022–006X.74. 3. 568

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

Cottraux, J., Note, I., Albuisson, E., Yao, S. N., Note, B., Mollard, E., … Coudert, A. J. (2000). Cognitive behavior therapy versus supportive therapy in social phobia: A randomized controlled trial. *Psychotherapy and Psychosomatics*, *69*(3), 137–146. doi:10.1159/000012382

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *The Canadian Journal of Psychiatry/La Revue canadienne de psychiatrie*, *58*(7), 376–385.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, *76*(6), 909–922. doi:10.1037/a0013075

Dattilio, F. M. & Norcross, J. C. (2006). Psychotherapy integration end (*sic*) the emergence of instinctual territoriality. *Archives of Psychiatry and Psychotherapy*, *8*(1), 5–16.

Del Re, A. C., Spielmans, G. I., Flückiger, C., & Wampold, B. E. (2013). Efficacy of new generation antidepressants: Differences seem illusory. *PLoS ONE*, *8*(6), e63509.

Driessen, E., Cuijpers, P., de Maat, S. C. M., Abbass, A. A., de Jonghe, F., & Dekker, J. J. M. (2010). The efficacy of short-term psychodynamic psychotherapy for depression: A meta-analysis. *Clinical Psychology Review*, *30*(1), 25–36. doi:10. 1016/j.cpr.2009.08.010

Durham, R. C., Murphy, T., Allan, T., Richard, K., Treliving, L. R., & Fenton, G. W. (1994). Cognitive therapy, analytic psychotherapy and anxiety management training for generalised anxiety disorder. *The British Journal of Psychiatry*, *165*, 315–323. Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*, 319–324.

Eysenck, H. J. (1961). The effects of psychotherapy. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology*. New York, NY: Basic Books.

Eysenck, H. J. (1966). *The effects of psychotherapy*. New York, NY: International Science Press.

Foa, E. B., Zoellner, L. A., Feeny, N. C., Hembree, E. A., & Alvarez-Conrad, J. (2002). Does imaginal exposure exacerbate PTSD symptoms? *Journal of Consulting and Clinical Psychology*, *70*(4), 1022–1028. doi:10.1037/0022–006X.70.4.1022

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

Gadermann, A. M., Alonso, J., Vilagut, G., Zaslavsky, A. M., & Kessler, R. C. (2012). Comorbidity and disease burden in the National Comorbidity Survey Replication (NCS-R). *Depression and Anxiety*, *29*(9), 797–806. doi:10.1002/da.21924

Geary, R. C. (1935). The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika*, *27*, 310–332.

Greenberg, G. (2013). *The book of woe: The DSM and the unmaking of psychiatry*. New York, NY: Penguin.

Greenberg, L. S. (2010). *Emotion-focused therapy*. Washington, DC: American Psychological Association.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. doi:10. 1037/1082–989X.6. 3. 203

Herbert, J. D., Gaudiano, B. A., Rheingold, A. A., Myers, V. H., Dalrymple, K., & Nolan, E. M. (2005). Social skills training augments the effectiveness of cognitive behavioral group therapy for social anxiety disorder. *Behavior Therapy*, *36*(2), 125–138. doi:10.1016/S0005–7894(05)80061–9

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.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York, NY: Russell Sage Foundation.

Imel, Z. E., Baer, J. S., Martino, S., Ball, S. A., & Carroll, K. M. (2011). Mutual influence in therapist competence and adherence to motivational enhancement therapy. *Drug and Alcohol Dependence*, *115*(3), 229–236. doi:10.1016/j.drugalcdep. 2010.11.010

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient– provider interactions. *Psychotherapy*, *52*(1), 19–30. doi:10. 1037/a00368841

Jansen, J. P., & Naci, H. (2013). Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Medicine*, *11*, 159. doi:10.1186/1741–7015–11–159

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*, 1–27. doi:10.1146/annurev.clinpsy.3.022806.091432

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research*, *19*(4–5), 418–428. doi:10.1080/10503300802448899

Kivlighan, D. M., Goldberg, S. B., Abbas, M., Pace, B. T., Yulish, N. E., Thomas, J. G., … Wampold, B. E. (2015). The enduring effects of psychodynamic treatments vis-à-vis alternative treatments: A multilevel longitudinal meta-analysis. *Clinical Psychology Review*, *40*, 1–14. doi:10.1016/j.cpr.2015. 05. 003

Knijnik, D. Z., Kapczinski, F., Chachamovich, E., Margis, R., & Eizirik, C. L. (2004). *Psicoterapia psicodinâmica em grupo para fobia social generalizada* [Psychodynamic group treatment for generalized social phobia]. *Revista Brasileira de Psiquiatria*, *26* (2), 77–81. doi:10.1590/S1516–44462004000200003

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge: Cambridge University Press.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge: Cambridge University Press.

Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.

Larvor, B. (1998). *Lakatos: An introduction*. London: Routledge. Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy*, *51*(4), 467–481. doi:10.1037/a0034332

Leichsenring, F., Luyten, P., Hilsenroth, M. J., Abbass, A. A., Barber, J. P., Keefe, F. J., … Steinert, C. (2015). Psychodynamic therapy meets evidence-based medicine: A systematic review using updated criteria. *Lancet Psychiatry*, *2*, 448–660.

Liberman, B. L. (1978). The role of mastery in psychotherapy: Maintenance of improvement and prescriptive change. In J. D. Frank, R. Hoehn-Saric, S. D. Imber, B. L. Liberman, & A. R. Stone (Eds.), *Effective ingredients of successful psychotherapy* (pp. 35–72). Baltimore, MD: Johns Hopkins University Press.

Lieb, K., Osten-Sacken, J., Stoffers-Winterling, J., Reiss, N., & Barth, J. (2016). Conflicts of interest and spin in reviews of psychological therapies: A systematic review. *BMJ Open*, *6*(4), e010606. doi:10.1136/bmjopen-2015–010606

Lilienfeld, S. O. (2014). The research domain criteria (RDoC): An analysis of methodological and conceptual challenges. *Behaviour Research and Therapy*, *62*, 129–139. doi:10.1016/j.brat.2014.07.019

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, *21*(16), 2313–2324.

Mann, C. C. (1994). Can meta-analysis make policy? *Science*, *266*, 960–962.

Marcus, D. K., O'Connell, D., Norris, A. L., & Sawaqdeh, A. (2014). Is the dodo bird endangered in the 21st century? A meta-analysis of treatment comparison studies. *Clinical Psychology Review*, *34*(7), 519–530. doi:10.1016/j.cpr.2014.08.001

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537–560). New York, NY: Russell Sage Foundation.

Mayo-Wilson, E., Dias, S., Mavranezouli, I., Kew, K., Clark, D. M., & Pilling, S. (2014). Psychological and pharmacological interventions for social anxiety disorder in adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, *1*, 368–376. doi:10.1016/S2215–0366(14)70329–3

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. doi:10.1037/0022–006x.46.4.806

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity*, *42*, 513–525.

Munder, T., Brütsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review*, *33*(4), 501–511. doi:10.1016/j.cpr.2013. 02. 002

Munder, T., Flückiger, C., Gerger, H., Wampold, B. E., & Barth, J. (2012). Is the allegiance effect an epiphenomenon of true efficacy differences between treatments? A meta-analysis. *Journal of Counseling Psychology*, *59*(4), 631–637. doi:10.1037/a0029571

Munder, T., Gerger, H., Trelle, S., & Barth, J. (2011). Testing the allegiance bias hypothesis: A meta-analysis. *Psychotherapy Research*, *21*(6), 670–684. doi:10.1080/10503307.2011.602752

Norcross, J. C. (2011). *Psychotherapy relationships that work: Evidence-based responsiveness*. New York, NY: Oxford University Press.

Norcross, J. C., & Newman, C. F. (1992). Psychotherapy integration: Setting the context. In J. C. Norcross & M. R. Goldfried (Eds.), *Handbook of psychotherapy integration* (pp. 3–45). New York, NY: Basic Books.

Öst, L. G. (1987). Applied relaxation: Description of a coping technique and review of controlled studies. *Behaviour Research and Therapy*, *25*(5), 397–409. doi:10.1016/0005–7967(87)90017–9

Owen, J., Drinane, J. M., Idigo, K. C., & Valentine, J. C. (2015). Psychotherapist effects in meta-analyses: How accurate are treatment effects? *Psychotherapy*, *52*(3), 321–328. doi:10.1037/pst0000014

Powers, M. B., Smits, J. A. J., Whitley, D., Bystritsky, A., & Telch, M. J. (2008). The effect of attributional processes concerning medication taking on return of fear. *Journal of Consulting and Clinical Psychology*, *76*(3), 478–490.

Psaty, B. M., Weiss, N. S., & Furberg, C. D. (1999). Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. *JAMA*, *282* (8), 786–790. doi:10.1001/jama.282.8.786

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology*, *35*, 203–208.

Schnurr, P. P., Friedman, M. J., Foy, D. W., Shea, M. T., Hsieh, F. Y., Lavori, P. W., … Bernardy, N. C. (2003). Randomized trial of trauma-focused group therapy for posttraumatic stress disorder: Results from a Department of Veterans Affairs cooperative study. *Archives of General Psychiatry*, *60*(5), 481–489. doi:10.1001/archpsyc.60.5.481

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. doi:10.1037/0003–066x.40.1.73

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren, C. Lewis, G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. (pp. 199–228). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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.

Shear, M. K., Houck, P., Greeno, C., & Masters, S. (2001). Emotion-focused psychotherapy for patients with panic disorder. *American Journal of Psychiatry*, *158*(12), 1993–1998.

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist*, *65*(2), 98–109. doi:10.1037/ a0018378

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

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: The Johns Hopkins University Press.

Tolin, D. F. (2010). Is cognitive–behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review*, *30*(6), 710–720. doi:10.1016/j.cpr.2010.05.003

Tolin, D. F. (2014). Beating a dead dodo bird: Looking at signal vs. noise in cognitive-behavioral therapy for anxiety disorders. *Clinical Psychology: Science and Practice*, *21*(4), 351–362. doi:10.1111/cpsp.12080

Tolin, D. F. (2015). Corrigendum to "Beating a dead dodo bird: Looking at signal vs. noise in cognitive-behavioral therapy for anxiety disorders". *Clinical Psychology: Science and Practice*, *22*, 315–316.

Trinquart, L., Abbé, A., & Ravaud, P. (2012). Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. *PLoS ONE*, *7*(4), 1–8. doi:10.1371/ journal.pone.0035219

Trinquart, L., Chatellier, G., & Ravaud, P. (2012). Adjustment for reporting bias in network meta-analysis of antidepressant trials. *BMC Medical Research Methodology*, *12*(1), 150.

Wampold, B. E. (2013a). Corrigendum to "Cognitive-behavioral therapy versus other therapies: Redux". *Clinical Psychology Review*, *33*, 1253. doi:10.1016/j.cpr.2013.08.001

Wampold, B. E. (2013b). The good, the bad, and the ugly: A 50-year perspective on the outcome problem. *Psychotherapy*, *50*(1), 16–24. doi:10.1037/a0030570

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The research evidence for what works in psychotherapy* (2nd ed.). New York, NY: Routledge.

Wampold, B. E., Imel, Z. E., Laska, K. M., Benish, S., Miller, S. D., Flückiger, C., … Budge, S. (2010). Determining what works in the treatment of PTSD. *Clinical Psychology Review*, *30*(8), 923–933. doi:10.1016/j.cpr.2010.06.005

Wampold, B. E., Imel, Z. E., & Miller, S. D. (2009). Barriers to the dissemination of empirically supported treatments: Matching messages to the evidence. *The Behavior Therapist*, *32* (7), 144–155.

Wampold, B. E., Minami, T., Baskin, T. W., & Tierney, S. C. (2002). A meta-(re)analysis of the effects of cognitive therapy versus "other therapies" for depression. *Journal of Affective Disorders*, *68*, 159–165.

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: Empiricially, "All must have prizes". *Psychological Bulletin*, *122*, 203–215.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, *5*, 425–433.

Wampold, B. E., & Serlin, R. C. (2014). Meta-analytic methods to test relative efficacy. *Quality and Quantity*, *48*, 755–765. doi:10. 1007/s11135–012–9800–6

Watson, J. C., Gordon, L. B., Stermac, L., Kalogerakos, F., & Steckley, P. (2003). Comparing the effectiveness of processexperiential with cognitive-behavioral psychotherapy in the treatment of depression. *Journal of Consulting and Clinical Psychology*, *71*, 773–781.

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.

Zachar, P. (2015). Psychiatric disorders: Natural kinds made by the world or practical kinds made by us? *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *14* (3), 288–290. doi:10.1002/wps.20240