Psychological Bulletin
1997, Vol. 122, No. 3, 221–225

0033-2909/97/$3.00

Trials and Tribulations in the Meta-Analysis of Treatment Differences: Comment on Wampold et al. (1997)

Kenneth I. Howard and Merton S. Krause
Northwestern University

Stephen M. Saunders
Marquette University

S. Mark Kopta
University of Evansville

A fair test of the Dodo bird conjecture that different psychotherapies are equally effective would entail separate comparisons of every pair of therapies. A meta-analysis of overall effect size for any particular set of such pairs is only relevant to the Dodo bird conjecture when the mean absolute value of differences is 0. The limitations of the underlying randomized clinical trials and the problem of uncontrolled causal variables make clinically useful treatment differences unlikely to be revealed by such heterogeneous meta-analyses. To enhance implications for practice, the authors recommend an intensified focus on patient–treatment interactions, cost-effectiveness variables, and separate meta-analyses for each pair of treatments.

Wampold et al. (1997) examined studies that directly compared "bona fide" treatments [i.e., treatments that "were based on psychological principles, were offered to the psychotherapy community as viable treatments (e.g., through professional books or manuals)," and "were delivered by trained therapists" (p. 205; with at least a master's degree)] to patients with bona fide clinical problems. The results of their analyses are consistent with those of prior meta-analyses, and proponents of psychotherapy can be reassured by the convergence of their findings. For example, Lipsey and Wilson (1993) examined 156 meta-analyses in which treatments were compared with control conditions. They calculated a mean effect size of .47, which was considerably larger than the mean effect size of many widely used, "validated" medical interventions. Grissom (1996) calculated "probability of superiority estimates" (cf. Howard, Krause, & Vessey, 1994) from prior meta-analyses. His analysis indicated that, in general, therapy was much better than no treatment and better than a placebo and that the median probability of superiority for studies comparing two therapies was only slightly greater than 50–50. So Wampold et al.'s meta-analysis is in a tradition of results indicating that efficacy differences between psychotherapeutic treatments are, on the average, modest to small.

How to Compare Treatments

If we look for sheer differences in outcome among psychotherapies to see whether they are all the same in terms of the patients' mental health status after therapy, we do not care which therapy is better or worse but only how different they are from one another. The measure of sheer difference is the absolute value of the difference because the algebraic signs of the differences are irrelevant, so the sum of these absolute values divided by the number of them is the mean difference. Although this mean has a lower bound of zero, it could only actually take a value of zero if all the absolute values of differences were uniformly zero.[1] The .19 mean absolute value effect size reported by Wampold et al. would thus seem to be the value we want, so there is, on the average, a significant difference (according to their data) in outcome in trials of various pairs of psychotherapies (see, e.g., Lipsey & Wilson, 1993, pp. 1198–1199, for a discussion of "small" effect sizes). But is this really what we want to know if we are interested in differences between therapies?

If we compare applications of psychotherapies by pairing the application of one with the application of another so as to calculate the difference between their outcomes, we are really looking, for practical purposes, to order a set of therapies on a

Kenneth I. Howard and Merton S. Krause, Department of Psychology, Northwestern University; Stephen M. Saunders, Department of Psychology, Marquette University; S. Mark Kopta, Department of Psychology, University of Evansville.

Correspondence concerning this article should be addressed to Kenneth I. Howard, Department of Psychology, Northwestern University, 2029 Sheridan Boulevard, Evanston, Illinois 60208-2710. Electronic mail may be sent via Internet to k-howard@nwu.edu.

[1] Randomly assigning algebraic signs to these absolute values, presumably half pluses and half minuses, and taking the sum of the resulting values to divide by the number of differences yields a mean with no relationship to the mean difference of the absolute values but with an expected value of zero, whatever the mean difference is in the absolute values. This apparently was Wampold et al.'s (1997) procedure for deriving their chief statistic, the mean of the randomly (but equally) signed absolute value differences. But the mean of the randomly (but equally) signed differences can only equal the mean difference of absolute values if the latter is zero (i.e., when each and every difference is zero). Wampold et al.'s significance testing for thick tails in the sample distribution of their randomly assigned differences is not an adequate substitute for testing the mean difference of the absolute values (i.e., .19) because their testing of the thick tails yields an insignificant result, whereas testing the correct statistic apparently does yield a statistically significant result.

common outcome metric. If therapy Time 1 (T1) is *D* outcome units better than T2, T2 is *D* better than T3, and T3 is *D* better than T1, then the mean difference in outcome among the three therapies is *D*, but they cannot be ordered on a one-dimensional outcome metric; that is, the results of the three comparisons are inconsistent (the therapies' outcomes are not transitive). However, if we alter this scenario by having T1 2 *D* better than T3, we get a mean difference in outcome of 1.33 *D* and consistent results that order the three therapies as to outcome: T1 > T2 > T3. If each betters every other therapy in half their comparisons and is bettered in the other half, we have the inconsistent results of our first scenario. If the comparisons yield consistent results analogous to those of our second scenario, we get interpretable standings that order the set of therapies. The point is that we are not interested in awarding prizes contest by contest, comparison by comparison, but according to standings after a round of comparisons, such that each therapy has been paired with every other (several times). We need to scale the therapies on outcome, not to estimate a mean difference between all pairs of therapies.

How are we ever to collect the data we need to scale a set of therapies? Ideally in the randomized experiment tradition, we would conduct an experiment large enough to include all of the therapies at issue and randomly assign patients, therapists, and settings to therapies in large enough numbers to guarantee equality across the therapies of any patient, therapist, and setting variables that had any causal relevance for outcome. Unfortunately, however, we have never been able to do this and probably never will. Instead, we have settled for a practical alternative strategy: We compare pairs of therapies (occasionally more) and aggregate the results of these somewhat disparate comparisons in meta-analyses, therapy pair by therapy pair. Holding to the randomized experiment tradition, we take only the differences between therapies as meaningful (because the main effects of unknown, uncontrolled causal variables that are subject to successful random assignment would affect therapy means but not mean differences between therapies).

Moreover, because different comparisons use different outcome measures, we use a standard-deviation-of-outcomes metric rather than any direct measure of outcome for our aggregation.[2] Because of our restriction to interpreting mean differences in outcomes between therapies, that is, avoiding interpreting mean outcomes per se, we can derive outcome standings for a set of therapies only if the results of the comparisons are ordinally consistent. So meta-analyses based on effect sizes from randomized experiments cannot in general provide what clinicians really want, that is, to know how good each therapy is.

Confounding in Meta-Analyses

The interpretation of meta-analyses faces serious problems. One such problem concerns the potential influence of unknown causal variables, confounds. Attempting to replicate an experiment involving two psychotherapies, where there is an additional unknown causal variable left uncontrolled, allows that unknown causal variable to vary across replications. Most simply, let us assume that a two-level unknown causal variable (e.g., an adequate therapeutic alliance vs. a less than adequate one) predominantly takes its high level for one attempted replicate experiment but its low level for another. As a result of a successful random assignment of the unknown causal variable within experiments, any main effect of that variable (at whatever level) is excluded from the within-experiment, between-therapy differences. Any interaction of the unknown causal variable with the therapies, however, can affect these differences (as a high level of therapeutic alliance may enhance the effects of interpersonal therapy more than it may the effects of in vivo desensitization). Insofar as the effect of such an interaction varies with the level of the unknown causal variable, which is not randomly assigned between experiments, it affects the between-therapy differences differently across attempted replicate experiments (because, e.g., an inadequate therapeutic alliance may detract equally from both interpersonal therapy and in vivo desensitization). Thus, in the presence of such an interactive unknown causal variable, any specific therapy pair's net comparison—by being averaged across replications that differ in mean level of the uncontrolled causal variable—is confounded by the uncontrolled causal variable's interaction effect. So how are we to interpret mean effect sizes constructed from attempted replications when there are unknown, interactively causal variables uncontrolled and varying across these replications (as we know there must be if we are tempted by the heterogeneity of effect sizes to do a meta-analysis in the first place)? At the very least, we must be careful to present our average effect sizes as the best estimates we have so far, not as probably accurate estimates when we still have no idea how accurate they are.

The Dodo Bird and Its Demise

The gist of our argument, then, is that what Wampold et al. (1997) apparently wanted to do should have concerned only the absolute values of differences (not signed differences). Even then, however, what they wanted to do has limited bearing on what we are to make of the comparative efficacy or effectiveness of extant psychotherapies. The mean absolute difference in effect size (in standard-deviation-of-outcome units) of .19 between pairings of applications of psychotherapies could be due to many different comparisons of therapies that have modest effect sizes, to comparisons of one pair of therapies that has received considerable research attention, or to a few comparisons with large effect sizes. The mean absolute difference in effect sizes reported by Wampold et al. does not imply that psychotherapies differ on average .19 in outcome, only that the average experimental comparison results in a .19 effect size.[3] This is useful for setting sample sizes for statistical power in future such experiments where one does not have more specific information regarding the particular therapies to compare.

The Dodo bird conjecture is that "when treatments . . . are compared, the true differences among all such treatments are

[2] This is a strategy necessitated by having to make sense of somewhat disparate comparisons of therapies in some systematic, quantitative way that allows us to weight each comparison's result in a common metric, according to that comparison's sample size and quality.

[3] This is similar to interpreting the grand mean of a stratified sample where the strata have different sample sizes.

zero" (Wampold et al., 1997, p. 204). If it is true, then for every set of replications of the same pair of therapies the mean effect size should not deviate significantly from zero (except for the percentage of such pairs that we would expect to do so by chance). A generalization to all pairs of therapeutic treatments is wanted, so one mean effect size for all comparative experiments is not the relevant statistic (unless the mean effect size is zero). In getting away from comparisons between classes of therapies because the Dodo bird conjecture concerns individual therapies, Wampold et al. went too far, to all comparative experiments rather than to all pairs of therapies, in aggregating for an effect size and an omnibus significance test. Furthermore, it is overly conservative to require the average experimental comparison to yield an effect size significantly different from zero before any pair of psychotherapies can properly be examined for differential efficacy or effectiveness. So long as better mental health status is important, no amount of prior failures to rise above the results of some baseline should obstruct further efforts, and the omnibus significance test used by Wampold et al. represents just such an obstruction. A body of successful replications on the same pair of therapies using the same set of variables and measures and analyzed as a whole, that is, meta-analytically (Schmidt, 1992),[4] is the most legitimate basis for claiming that and how much one therapy is better than another for certain sorts of patients under certain sorts of conditions. Meta-analyses of somewhat more disparate experiments can help lead us to focus on particular pairs, to reliable demonstrations capping a program of experimentation.

Treatment Variables

Wampold et al. (1997) called our attention to a possible source of ideas for new causal variables definitive of psychotherapies when they noted that "it is poignant . . . that the size of the effect between bona fide psychotherapies is at most about half of the effect size produced by treatments with no active psychotherapeutic ingredients (i.e., placebo vs. no treatment)" (p. 210). Does not this (like the immediate improvement inferred in Howard, Kopta, Krause, & Orlinsky, 1986) suggest that we have overlooked some active psychotherapeutic ingredients right under our noses because we have been prejudiced against seeing and looking for active psychotherapeutic ingredients in placebo or putatively no-treatment conditions? This is certainly reason to "persist in attempts to find treatment differences" (Wampold et al., 1997, p. 211). Ought not we look for correlates of outcome within placebo and no-treatment groups, especially ones possibly amenable to therapist influence, and then try to control them in treatment groups?

Some investigators have argued that it is not vital or even relevant to justify psychotherapy by demonstrating that different therapies contain unique active ingredients as medications are purported to. Lambert and Bergin (1994) have emphasized that common factors should not be interpreted as inert factors and have listed 30 common factors existing across therapies. Klein (1996) also described some of these factors: "A strong, knowledgeable, professional ally who therapeutically provides the patient with emotional support, usable coping skills, and success experiences and helps reframe life experiences so as to heighten self-esteem" (p. 82).

The current impetus for the unique ingredients position comes, in part, from the demand for evidence required for psychotherapy to strengthen its status as an effective treatment (Task Force on Promotion and Dissemination of Psychological Procedures, 1995). Wampold et al. (1997) suggested letting go of the medication model notion that psychotherapy works because of active ingredients. However, we believe that the active ingredient explanation for psychotherapeutic effectiveness has merit. For example, the dosage model, which specifies the relationship between amount of treatment and patient improvement, uses the session as the critical unit of treatment and is based on the assumption that there is a stochastic relationship between number of sessions and the patient's exposure to a treatment's active ingredients (Howard et al., 1986). The dosage model is the theoretical base for a phase model of treatment outcome, which specifies that subjective remoralization precedes symptomatic remediation (which precedes rehabilitation of functioning; Howard, Lueger, Maling, & Martinovich, 1993), as well as for subsequent research on the relationship between treatment dosage and patient improvement (Barkham et al., 1996; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Howard, Orlinsky, & Lueger, 1995; Kopta, Howard, Lowry, & Beutler, 1994; Tingey, Lambert, Burlingame, & Hansen, 1996). So far, the medication model has lead to interesting findings and insights about how patients' conditions respond to treatment.

Randomized Clinical Trials and Meta-Analysis

The usual meta-analysis is applied to a collection of randomized clinical trials (RCTs) that address the efficacy question, Under controlled circumstances, does Procedure X produce a particular average effect above and beyond the average effect of Procedure Y? (cf. Howard et al., 1996; and Seligman, 1995). Efficacy research is intended to protect the internal validity of findings and to demonstrate that there is a causal relationship between the intervention and the outcome. However, the procedures of RCTs tend to reduce the external validity of findings: The patients are carefully selected on the basis of a single diagnostic category to reduce within-cell variability and are randomly assigned to treatment conditions, and therapists are constrained in their interventions, including the length of treatment, regardless of patient progress, to maximize the "integrity" of the putative therapeutic ingredients. In contrast, the usual psychotherapy patient likely has multiple disorders (Kessler et al., 1994), patients choose both treatment and practitioners that they see as helpful, treatment is typically adjusted according to the response of the patient and is based on the professional judgment

[4] Meta-analyses are useful because by aggregating findings, they correct for a frequent misuse of significance testing. Schmidt pointed out that typical research experiments (with the treatment and control groups having 15 participants each and using a two-tailed test of significance) detect an actual effect size of .50, which is a medium-sized effect in Cohen's (1988) typology of effect sizes, only 26% of the time, resulting in an exaggerated Type II error rate. If a series of such studies were reviewed, it would be concluded that 74% of the studies found no difference between groups. Schmidt showed that, in contrast, a meta-analysis yields the correct conclusion, especially if the number of studies included in the meta-analysis is large.

of the practitioner, and treatment is offered until the patient is better. Moreover, RCTs test for one-way causality, which is inconsistent with clinical practice, wherein patient progress affects the process of therapy as much as the process of therapy determines the outcome. Simply put, RCTs may reflect very little about the reality of psychotherapy practice where patients and clinicians are concerned about whether this treatment, conducted in this manner, is producing the desired effect.

When RCTs fail to reveal expected differences between treatments, a usual recourse is the proposal of treatment–patient interaction effects. For example, it is probably the case that the Dodo bird verdict would hold for comparisons of the general main effects of antibiotic medications (as it would for antidepressants). But some patients are allergic to certain antibiotics, while others are unresponsive to certain antibiotics; so prescriptions have to be tailored to these patient characteristics (as well as, in some cases, to the patient's actual medical condition). Klein (1996) reminded us of the subject confound of "drug unresponsive" patients in treatment groups, whereas Elkin, Gibbons, Shea, and Shaw (1996) implied that there are patients who are much less responsive to cognitive–behavioral therapy than are others. Failures (and even successes) to find significant main effects should be (and most often are) followed by a search for such interactions in post hoc analyses.

It is also possible (even likely) that treatment differences are attenuated by selective attrition from treatment groups, for example, by the dropping out of noncompliers. For example, a recent study comparing fluoxetine ($n = 18$) with cognitive therapy ($n = 13$) in the treatment of dysthymia (Dunner et al., 1996) found that "five patients, all who had been randomly assigned to fluoxetine, dropped out prior to the 8-week assessment. Four withdrew consent when randomized (they stated that they were hoping to be assigned to cognitive therapy)" (p. 37). There was also attrition from the cognitive therapy group. It is not surprising, then, that the treatments attained similar results (with final group *n*s of 13 and 11, respectively, and self-selection playing a prominent role through attrition).

Another problem with the legacy of RCTs lies in the dependent variables that investigators have chosen. Human concerns demand that speed of improvement, emotional and financial costs, unpleasant side effects, and so forth should be considered as well as amount of improvement achieved at the termination of treatment (outcome) when comparing psychotherapeutic treatments. Given our concerns about the value of clinical trials research as mentioned above, we believe that future comparative treatment studies should focus on efficiency rather than just effectiveness. Therefore, we recommend future effect size calculations that take into account cost (e.g., rate of improvement) as well as amount of benefit. Even further, we recommend research that is focused on growth curves that reflect patient progress over the course of a treatment (Howard et al., 1996) rather than on main effects of treatment or comparative treatment outcomes at some arbitrary termination point. For research to be clinically applicable, the focus has to be on the individual patient and that patient's response to the treatment so far.

When meta-analyses are based on RCTs or less controlled comparisons, they inherit all of the problems of these kinds of comparative experiments. Meta-analysis does not solve these problems.

Caveat

We want to emphasize that equivalency findings would not represent an indictment of different psychotherapies as valid treatments for psychological disorders. There are many medical interventions that produce equivalent results in the treatment of an illness. In fact, the Food and Drug Administration has guidelines for establishing the equivalence of medical interventions (e.g., the validation of generic drug substitutes). In this spirit, Wampold et al. (1997) stated that "the results of this meta-analysis suggest that the [average] efficacy of the treatments are comparable, not that the treatments are interchangeable" (p. 211). The best of all clinical situations would include a variety of effective treatments, such that treatment selection for a particular patient would be guided solely by the clinical characteristics and treatment responsiveness of that patient.

References

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose–effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64,* 927–935.

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

Dunner, D. L., Schmaling, K. B., Hendrickson, H., Becker, J., Lehman, A., & Bea, C. (1996). Cognitive therapy versus fluoxetine in the treatment of dysthymic disorder. *Depression, 4,* 34–41.

Elkin, I., Gibbons, R. D., Shea, M. T., & Shaw, B. F. (1996). Science is not a trial (but it can sometimes be a tribulation). *Journal of Consulting and Clinical Psychology, 64,* 92–103.

Grissom, R. J. (1996). The magical number .7 ± .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology, 64,* 973–982.

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

Howard, K. I., Krause, M. S., & Vessey, J. T. (1994). Analysis of clinical trial data: The problem of outcome overlap. *Psychotherapy, 31,* 302–307.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy: Causal mediation of outcome. *Journal of Consulting and Clinical Psychology, 61,* 678–685.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51,* 159–164.

Howard, K. I., Orlinsky, D. E., & Lueger, R. J. (1995). The design of clinically relevant outcome research: Some considerations and an example. In M. Aveline & D. A. Shapiro (Eds.), *Research foundations for psychotherapy practice* (pp. 3–47). Sussex, England: Wiley.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of *DSM–III–R* psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry, 51,* 8–19.

Klein, D. F. (1996). Preventing hung juries about therapy studies. *Journal of Consulting and Clinical Psychology, 64,* 81–87.

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

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

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

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

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

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research, 6,* 109–123.

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

Received March 28, 1997
Revision received April 10, 1997
Accepted April 10, 1997 ■