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)

Bruce E. Wampold, Gregory W. Mondin, Marcia Moody, and Hyun-nie Ahn University of Wisconsin—Madison

On the basis of a meta-analysis of comparisons of bona fide psychotherapies, B. E. Wampold et al. (1997) concluded that the available evidence supported the notion that all psychotherapies are nearly equal in terms of efficacy. K. I. Howard, M. S. Krause, S. M. Saunders, and S. M. Kopta (1997) and P. Crits-Christoph (1997) raised 4 general issues with this conclusion: (a) counterexamples, (b) untested alternative hypotheses, (c) methodological problems, and (d) adequacy of randomized clinical trials. Each of these issues is discussed, and it is asserted that empirically there is no basis to alter the conclusions reached in B. E. Wampold et al.'s (1997) meta-analysis.

The central metaphor used to illustrate the finding that psychotherapies produce generally equivalent outcomes has been that of the Dodo bird, who in *Alice in Wonderland* (Carroll, 1865/ 1962) exclaimed, "everybody has won, and all must have prizes." In spite of the empirical evidence for the conclusion that there are no differences in relative efficacy among bona fide psychotherapies, the Dodo bird verdict cannot achieve much respect, as witnessed by the commentaries (Crits-Christoph, 1997; Howard, Krause, Saunders, & Kopta, 1997) of the most recent meta-analysis of studies that have compared various psychotherapies (Wampold et al., 1997). It is now time to replace the Dodo bird metaphor with one involving the flat Earth, a notion that persists in spite of evidence to the contrary.

Legend has it that, despite data to the contrary, 15th century Europeans believed that the Earth was flat and that Columbus's voyage to the New World was extremely risky because of a possibility that he would fall off the edge of the Earth.¹ His successful voyage provided evidence that the Earth was spherical; any remaining doubt about the general shape of the planet should have been annihilated by Magellan's circumnavigation of the globe. Nevertheless, members of the Flat Earth Society insisted that the world is not spherical. With the context of the metaphor established, we contend that the evidence for the conclusion that there are no differences in efficacy among bona fide psychotherapies is sufficient to give up belief that there are substantial differences.

Smith and Glass (1977) were the first meta-analytic explorers

psychotherapies. As expected, their conclusions were challenged vociferously (e.g., Eysenck, 1978; Rachman & Wilson, 1980; Wilson & Rachman, 1983; see also Glass & Kliegl's, 1983, defense of their meta-analytic findings). However, subsequent meta-analysts returned with conclusions consistent with Smith and Glass's (e.g., Grissom, 1996; Robinson, Berman, & Neimeyer, 1990; and Shapiro & Shapiro, 1982). But even then, increased emphasis on conducting clinical trials comparing psychotherapies shows a persistent belief that true differences must be discovered (Goldfreid & Wolfe, 1996). It was our (Wampold et al., 1997) original hope that by addressing some problems in previous meta-analyses, we could settle the uniform efficacy question with some confidence. Howard et al. (1997) commented that "the results of their [Wampold et al., 1997] analyses are consistent with those of prior meta-analyses, and proponents of psychotherapy can be reassured by the convergence of their findings" (p. 221). Nevertheless, the consistency of results appears to have not yet reassured everyone, as is clear by the Crits-Christoph (1997) and Howard et al. (1997) commentaries. The purpose of this reply is to address the criticisms (Crits-

to report that there were nonexistent to small differences among

Christoph, 1997; Howard et al., 1997) of our (Wampold et al., 1997) meta-analysis. Although space limitations do not allow us to address each of Crits-Christoph's and Howard et al.'s points, we discuss four important issues: (a) counterexamples, (b) untested alternative hypotheses, (c) methodological problems, and (d) adequacy of randomized clinical trials.

Bruce E. Wampold, Gregory W. Mondin, Marcia Moody, and Hyunnie Ahn, Department of Counseling Psychology, School of Education, University of Wisconsin-Madison.

We appreciate the history of science perspective provided by Michael H. Shank, Department of History of Science, University of Wisconsin—Madison.

Correspondence concerning this article should be addressed to Bruce E. Wampold, Department of Counseling Psychology, School of Education, University of Wisconsin — Madison, 321 Education Building, 1000 Bascom Mall, Madison, Wisconsin 53706. Electronic mail may be sent via Internet to wampold@macc.wisc.edu.

¹ Those versed in the history of science will note that we have taken some literary license in this metaphor. Actually, the sphericity of the Earth was recognized by scholars since the time of the Greek philosophers (around 400 BC). Fifteenth-century explorers and scholars accepted the spherical notion, and concerns about falling off the end of the Earth were never raised by Columbus or those who sponsored his voyages. The notion that there was a medieval belief in a flat Earth appeared in the 19th century and was fostered by scholars, emboldened by positivism and at "war" with religious constructions of the world, who sought to show that medieval thought was clouded by religion (see Russell, 1991).

Counterexamples

Crits-Christoph (1997) examined the studies sampled in the meta-analysis and found a trend among studies that cognitivebehavioral treatments were superior to other types of therapy, and thus he maintained that our meta-analysis may have obscured important exceptions to the general equivalence of psychotherapeutic efficacy. There are a number of problems with detecting trends in this fashion.

It is always risky to make generalizations from post hoc examinations of data, as all data (even random data) yield relationships, many of which may be intriguing. However, the speculation that cognitive-behavioral treatments are superior to other treatments has persisted for decades, so this trend must be taken more seriously than just any random pattern in the data. It should be noted that in our meta-analysis, we (Wampold et al., 1997) approached the efficacy problem from the omnibus perspective; that is, we tested whether there were differences among treatments. Had there been differences among treatments generally (i.e., the effects were not homogeneously distributed around zero), then the appropriate hypothesis testing strategy would have been to test post hoc for the sources of differences. Crits-Christoph's (1997) approach was to post hoc examine one particular contrast: cognitive-behavioral versus other treatments (to be referred to as the cognitive contrast).

The data used to support the cognitive contrast, which were displayed in Table 1 of Crits-Christoph (1997, p. 218), were found by identifying studies that (a) contrasted cognitive-behavioral treatments with other treatments or with cognitive-behavioral treatments that had a couples component, (b) had participants who were not undergraduates, and (c) contained "meaning-ful difference[s] between the treatment conditions" (p. 217) on one dependent variable. This search strategy yielded 15 dependent variables, with effect sizes ranging from 0.39 to 1.60.

Our (Wampold et al., 1997) meta-analytic conclusions were based on nearly 3,000 dependent variables, which provides a more robust conclusion than can a trend based on 15 variables. We purposefully aggregated across dependent variables to obtain the best estimate of the effect of the treatment on variables thought to be important to assess psychological functioning by the study researchers. Selecting a single variable that shows a "meaningful difference" represents the prototypic "fishing and error rate" threat to validity, as discussed by Cook and Campbell (1979), in which the alpha level is inflated by selecting post hoc statistically significant variables from a set of predominantly nonsignificant variables. For example, to support the cognitive trend, Crits-Christoph (1997) cited Borkovec and Mathews (1988), who compared coping desensitization to cognitivebehavioral treatments for nonphobic anxiety; an effect size of .660, in favor of the cognitive-behavioral treatment, was calculated for one dependent variable, the Hamilton Rating Scale for Anxiety. However, when the effect size for all 11 dependent variables in the study was determined, using the methods described in the meta-analysis, the effect size was .045 (SE = .333). This indicates that there was absolutely no evidence in this study for the superiority of the cognitive-behavioral treatment, a conclusion reached by Borkovec and Mathews, who stated that "no differences were found between . . . conditions" (p. 877).²

Another "fishing" aspect of the studies cited by Crits-

Christoph (1997) in his Table 1 is the omission of comparisons of cognitive-behavioral treatments with other treatments in which no differences were found; the NIMH (National Institute of Mental Health) Collaborative Study on Treatments of Depression (Elkin et al., 1989) is the most conspicuous absence. To adequately test the cognitive contrast, a researcher must aggregate all studies containing a cognitive-behavioral treatment.

Another study used to support the cognitive contrast (viz., Foa, Rothbaum, Riggs, & Murdock, 1991) contained a treatment that was not classified as bona fide in our meta-analysis (Wampold et al., 1997). To make the comparisons of psychotherapies fair, we carefully defined the conditions under which a treatment would be classified as bona fide so that conclusions of the superiority of treatments would not be reached by comparisons with treatments not intended to be therapeutic (e.g., placebo controls or "alternative" therapies). The treatment used by Crits-Christoph (1997) to support the superiority of cognitivebehavioral treatments was supportive counseling for posttraumatic stress in women who had recently (within the previous year) been raped. In this treatment, patients were taught a general problem-solving technique (not tailored to the individual patient), therapists responded indirectly and were unconditionally supportive, and "patients were immediately redirected to focus on current daily problems if discussions of the assault occurred'' (Foa et al., 1991, p. 718). Although a manual existed for this treatment, it did not meet the other criteria used in the meta-analysis (viz., a reference to an established approach, reference to psychological processes, citation to active ingredients); moreover, in the absence of other components, few would accept deflecting women from discussing their recent rape in counseling as therapeutic. Finally, Foa et al. (1991) included supportive counseling "to control for nonspecific therapy effects" (p. 716), clearly making this treatment a placebo control rather than a bona fide treatment.

Any test of the cognitive contrast (and other alternative hypotheses; see the next section) has to carefully define constructs and operationalize them validly for one to make statements about relative effectiveness. One of the reasons we (Wampold et al., 1997) avoided classifying treatments into types of therapy was that there are no generally accepted definitions of therapy types. What is cognitive-behavioral therapy? Apparently, Crits-Christoph's (1997) implicit definition of cognitivebehavioral treatment is expansive because it contains emotionally focused therapy (Goldman & Greenberg, 1992), in which it is assumed that "psychological symptoms are seen as emanating from the deprivation of unmet adult needs," and involves, in part, "identification with previously unacknowledged aspects of experience by enactment of redefined cycle" (p. 964).

When we (Wampold et al., 1997) conducted the meta-analysis, we were surprised by the preponderance of studies that showed extremely small or nonexistent differences between treatments. The point of examining the studies in Crits-Christoph's (1997) cognitive contrast in such detail is that even the most conspicuous trend that could be found pales next to

² Crits-Christoph criticized the meta-analysis because targeted and secondary variables were aggregated. Although we return to this point later, it should be noted that 8 of the 11 variables in the Borkovec and Mathews study targeted anxiety and that the effect size for these targeted variables was small (.041, with SE = .339).

the sea of evidence that exists for the equivalence of outcomes in psychotherapy. When looked at rigorously, the data cannot support a conclusion that therapies, as presently studied, differ in any meaningful way, including the contrast of cognitive-behavioral versus other types.

Alternative Hypotheses

Many of the comments regarding the findings of our metaanalysis were of the following form: "Based on the studies examined, the meta-analysis demonstrated that the population effect size for the comparison of bona fide psychotherapies was zero; but had the data been examined differently, true differences would have appeared." The cognitive contrast is one such alternative hypothesis: Had the data been used to contrast cognitivebehavioral treatments to other treatments, the effect sizes would have been significantly greater than zero.

Note that we (Wampold et al., 1997) tested the primary criticism of previous Dodo bird conclusions, namely, that improving methods would detect the true superiority of some treatments (Stiles, Shapiro, & Elliott, 1986); ad absurdum, alternative hypotheses can always be generated post hoc to challenge any conclusion of any study. Although both commentaries suggested that decisions we made may have biased our results, neither presented any data (except the effects in the cognitive contrast discussed above) to indicate that these decisions will have to conduct either primary studies or meta-analyses to test their own hypotheses.

The alternative hypotheses proposed by Crits-Christoph (1997) and Howard et al. (1997) are discussed below. Where possible, we examined these hypotheses with the meta-analytic data set.

Follow-Up Assessments

Crits-Christoph (1997) commented that our (Wampold et al., 1997) meta-analytic estimate of the effect sizes was attenuated by the inclusion of follow-up assessments because there is a tendency for those symptomatic participants in the less efficacious treatment to obtain other treatment after termination of the study. Although we test this hypothesis in conjunction with another alternative hypothesis (see Severity below), we chose originally to test all points of assessment; but on the recommendation of one reviewer, we modified the data set to include only the follow-up assessment (see Wampold et al., 1997, p. 218, Table 1). The reviewer made this recommendation to eliminate the statistical dependencies inherent in multiple assessments and because long-term outcome is thought to represent the most naturalistic and important assessment. However, as seen later, considering only assessments at termination (along with severity) does not produce the effects predicted by Crits-Christoph.

Severity

Another alternative hypothesis suggested by Crits-Christoph (1997) is that differences between therapies are only expected for treatments of severe disorders. The equivalence of psychotherapies for other disorders, however, is accepted: "With mild conditions, the nonspecific effects of treatments . . . are likely to be powerful enough in themselves to affect . . . outcomes, leaving little room for the specific factors to play much of a role" (p. 217). To examine the Crits-Christoph hypothesis that true differences exist at termination for severe disorders, which he defined as involving Diagnostic and Statistical Manual of Mental Disorders (4th ed., DSM-IV; American Psychiatric Association, 1994) disorders, we identified the subset of studies in the meta-analytic database that met those criteria. The results indicate that considering severity and point of assessment does not alter the conclusion that there are no differences in efficacy among the psychotherapies compared. There were 50 effects that met the criteria of DSM-IV diagnosis (or equivalent) and assessment at termination, which yielded (a) a Q statistic that indicated homogeneity around zero (Q = 61.85, p = .10) and (b) an upper bound estimate of .23 (see Wampold et al., 1997, for a discussion of these statistics).³

Confounding With Unknown Causal Variables

Howard et al. (1997) argued that all tests of relative efficacy in the meta-analysis studies were attenuated by the confounding of the treatments with unknown causal variables and, therefore, that effect sizes obtained in meta-analyses should be presented as the "best estimates we have so far, not as probably accurate estimates when we still have no idea how accurate they are" (p. 222). We agree with this to the extent that our meta-analysis presented the best estimates so far (of course!), but we would argue that the presence of unknown variables should not be invoked to detract from either the studies or the meta-analysis. Researchers should model the constructs that have been empirically or theoretically identified as important; meta-analysts can only analyze the data generated by the researchers, who of course did not consider unknown causal variables.

Restricted Data Set

Both Crits-Christoph (1997) and Howard (1997) noted that the meta-analytic data set did not include comparisons of all psychotherapies for all disorders and, therefore, that the conclusions should be dramatically restricted. We (Wampold et al., 1997) recognized the limitations of generalizations in our study and clearly stated so:

That there are about 250 types of therapy and 300 disorders (Gold-fried & Wolfe, 1996) clearly indicates that the comparisons reviewed for this meta-analysis were not sampled from a Types of Therapy \times Types of Disorder matrix. . . . Moreover, it should be

³ It might be tempting to see a trend here. When the effects were limited to DSM-IV diagnoses and assessment at termination, it appears that the heterogeneity of the effects is greater (viz., p = .10, which some might say is "approaching" significance) and the upper bound estimate is greater (viz., 23) than was the case in the original meta-analysis. However, this trend was due to one study (Mavissakalian, Michelson, Greenwald, Kornblith, & Greenwald, 1983) that found that at termination paradoxical intention was far superior (i.e., a very large effect) to self-statement training in the treatment of agoraphobia (although this superiority was not present at follow-ups). When this study is removed, the remaining DSM-IV + termination assessment effects yield a homogeneity statistic approximately equal to what is expected under the null hypothesis of homogeneity (Q = 47.95 for 49 effects, p = .47).

recognized that the psychotherapies studied [and included in the meta-analysis] were those bona fide psychotherapies selected by psychotherapy researchers. . . . Consequently, the results of this meta-analysis should not be construed to support the conclusion that all practiced psychotherapies are equally efficacious or as efficacious as the ones we reviewed. (pp. 210-211)

However, given the existing research, we (Wampold et al., 1997) have made the most informed conclusion possible. If the extant research is insufficient to support our conclusion that there are no differences in efficacy, then certainly it is insufficient to support the alternative hypothesis that there are differences (i.e., scientifically, the null hypothesis is retained until there is sufficient evidence to accept the alternative).

Targeted Versus Secondary Variables

Another alternative hypothesis suggested by Crits-Christoph (1997) was that treatment differences would have been found had we (Wampold et al., 1997) focused on outcome measures on problems targeted by the treatment rather than on all outcome variables. Researchers rarely segregated variables into classes such as targeted and secondary; but when they did and made a hypothesis about the results, we accounted for this by calculating effect sizes for the classes. Although an interesting debate could be made for whether meta-analysts should aggregate across all variables or across targeted and secondary variables separately, we want to emphasize that such an analysis would have involved inferring which variables were targeted and which were secondary (see Footnote 2). A hypothesis that there are differences between targeted and secondary variables is reasonable, but so are many other hypotheses about outcome variables. For example, it could be hypothesized that perspective of the outcome assessment (viz., therapist, patient, or observer rated) makes a difference. Meta-analysts cannot be expected to test the universe of possible and reasonable hypotheses; we leave it to others to test hypotheses of their liking.

Methodological Problems

A methodological criticism is Howard et al.'s (1997) discussion of the assignment of the algebraic sign to effect sizes. Their issue apparently is related to differences between the two ways in which we (Wampold et al., 1997) assigned the algebraic signs: "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)" (Howard et al., 1997, p. 221, Footnote 1). This criticism is based on a misunderstanding of the distribution of the effect sizes under the null hypothesis and the resulting homogeneity test. Clearly, the mean of the absolute values of the effect sizes is an upper bound and thus an inflated estimate of the true effect size. In any event, an effect size of approximately .20 (which Howard et al., 1997, accepted) is small compared with other effect sizes produced in psychotherapy studies (see Wampold et al.'s, 1997, p. 210, Table 2).

Howard et al. (1997) suggested that the appropriate way to understand relative effectiveness is to scale therapies unidimensionally vis-à-vis efficacy. They suggested correctly that if all therapies were compared pairwise, such a scaling could be accomplished (e.g., using the Bradley-Terry-Luce model; McGuire & Davison, 1991). However, it is excessively unrealistic to think that researchers will ever conduct systematically pairwise comparisons of even a few treatments (e.g., 10 treatments would produce 45 pairwise comparisons). Moreover, small effect sizes produced by comparisons that we (Wampold et al., 1997) examined would, because of sampling error, produce an unstable scaling (i.e., many intransitive relationships). Howard et al. (1997) made the following conclusion:

Because of our restriction to interpreting mean differences in outcomes between therapies, . . . 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. (p. 222)

We strongly believe that until pairwise comparisons are conducted and treatments are ordered along the efficacy continuum (which realistically will never occur), an estimate of the differences among treatments is exactly what is needed. The alternative is to say that there is no useful way to summarize the outcome data and to leave clinicians to make their own conclusions.

Later, Howard et al. (1997) claimed that the most legitimate strategy is to examine "a body of successful replications on the same pair of therapies using the same set of variables and measures and analyzed as a whole" (p. 223). We do not disagree that replications are important but only make the case that such replications are not currently available to the meta-analyst.

Adequacy of Randomized Clinical Trials

Howard et al. (1997) and Crits-Christoph (1997) differed in their evaluation of the adequacy of randomized clinical trials (RCTs) as a means to establish the relative efficacy of psychotherapies. Howard and his associates (e.g., Howard, Krause, & Orlinsky, 1986; Howard et al., 1997) have argued that there are flaws in clinical trials due to such factors as problems with randomization, attrition, interactions with unknown causal variables, choice of dependent variables, and limited external validity: "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" (Howard et al., 1997, p. 224). It appears that they were criticizing the meta-analysis based on the inadequacy of RCTs. But one cannot have it both ways. Either it should be argued that (a) RCTs are flawed and, therefore, should not be used in psychotherapy research (and consequently meta-analyses are precluded as well) or (b) RCTs are legitimate (albeit with faults) and the meta-analysis of findings of RCTs are legitimate. We are not sure which of these two arguments is endorsed by Howard et al.

Crits-Christoph (1997), however, defended RCTs. Essentially his argument is that "Wampold et al. (1997) conducted a review of comparative studies in which they attempted to draw implications for issues that can only be resolved through comparisons of active treatments with control groups" (p. 219). We strongly disagree; direct comparisons are indeed the very studies one would want to examine to make conclusions about relative efficacy. As we read Crits-Christoph's comment, the point is that he believes that some placebo conditions adequately control for common ingredients in psychotherapy, the effects of psychotherapies that are superior to such placebo groups are due to specific ingredients, and this then validates the therapy. Although we disagree that placebo conditions are adequate to rule out common ingredients (see Wampold, 1997), this issue is independent of the results of the meta-analysis. Two psychotherapies could be shown to be superior to placebo controls, these effects could be due to their unique ingredients, and the two psychotherapies could be equally effective. The results of the meta-analysis indicate that there is not "even weak evidence of differential effectiveness" (Wampold et al., 1997, p. 211). Whether the effects are due to unique or common ingredients, the uniform effectiveness of psychotherapies is corroborated by our and previous meta-analyses.

Conclusions

Crits-Christoph (1997) disagreed "most strongly" with the implications of the meta-analysis because of the limitations. Howard et al. (1997) went further and claimed that the results of our meta-analysis represent "an obstruction." We posed a null hypothesis and found that the data were consistent with this null hypothesis. The critics of our findings have suggested that several alternative hypotheses are true. We challenge those critics to carefully state their hypotheses and collect the data to show that the null hypothesis can be rejected in favor of one of these alternatives.

To return to our flat Earth metaphor, precise measurements of the Earth have documented that the Earth is not perfectly spherical, but it is much closer to spherical than flat. It may be that continued research in psychotherapy will show that some treatments are slightly more effective than others, although the model of uniform effectiveness would fit these data better than a model that indicates that treatments vary in their effectiveness. We would cherish the day that a treatment is developed that is dramatically more effective than the ones we use today. But until that day comes, the existing data suggest that whatever differences in treatment efficacy exist, they appear to be extremely small, at best. Although uniform efficacy may not be a popular finding for some, this empirical result should guide, rather than obstruct, research and practice.

References

- American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.
- Borkovec, T. D., & Mathews, A. M. (1988). Treatment of nonphobic anxiety disorders: A comparison of nondirective, cognitive, and coping desensitization therapy. *Journal of Consulting and Clinical Psychol*ogy, 56, 877-884.
- Carroll, L. (1962). Alice's adventures in wonderland. Harmondsworth, Middlesex, England: Penguin Books. (Original work published 1865)
- Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally.
- Crits-Christoph, P. (1997). Limitations of the "Dodo bird" verdict and the role of clinical trials in psychotherapy research: Comment on Wampold et al. (1997). *Psychological Bulletin*, 122, 216–220.

Elkin, I., Shea, T., Watkins, J. T., Imber, S. D., Sotsky, S. M., & Collins,

J. F. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. Archives of General Psychiatry, 46, 971–982.

- Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.
- Foa, E. B., Rothbaum, B. O., Riggs, D. S., & Murdock, T. B. (1991). Treatment of posttraumatic stress disorder in rape victims: A comparison between cognitive-behavioral procedures and counseling. *Jour*nal of Consulting and Clinical Psychology, 59, 715–723.
- Glass, G. V., & Kliegl, R. M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology*, 51, 28–41.
- Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained alliance. American Psychologist, 51, 1007-1016.
- Goldman, A., & Greenberg, L (1992). Comparison of integrated systemic and emotionally focused approaches to couples therapy. *Journal* of Consulting and Clinical Psychology, 60, 962-969.
- Grissom, R. J. (1996). The magical number $.7 \pm .2$: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology*, 64, 92–103
- Howard, K. I., Krause, M. S., & Orlinsky, D. E. (1986). The attrition dilemma: Toward a new strategy for psychotherapy research. *Journal* of Consulting and Clinical Psychology, 54, 106-110.
- 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.
- Mavissakalian, M., Michelson, L., Greenwald, D., Kornblith, S., & Greenwald, M. (1983). Cognitive-behavioral treatment of agoraphobia: Paradoxical intention vs self-statement training. *Behaviour Re*search and Therapy, 21, 75-86.
- McGuire, D. P., & Davison, M. L. (1991). Testing group differences in paired comparisons data. *Psychological Bulletin*, 110, 171-182.
- Rachman, S. J., & Wilson, G. T. (1980). The effects of psychological therapy (2nd ed.). New York: Pergamon Press.
- Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin*, 108, 30–49.
- Russell, J. B. (1991). Inventing the flat Earth: Columbus and modern historians. New York: Praeger.
- Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, 92, 581-604.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.
- Stiles, W. B., Shapiro, D. A., & Elliott, R. (1986). "Are all psychotherapies equivalent?" American Psychologist, 41, 165-180.
- Wampold, B. E. (1997). Methodological problems in identifying efficacious psychotherapies. Psychotherapy Research, 7, 21-43.
- 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.
- Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology*, 51, 54-64.

Received May 9, 1997

Revision received May 19, 1997

Accepted May 20, 1997 ■