Journal of Consulting and Clinical Psychology
1991, Vol. 59, No. 6, 883–893

Mediators and Moderators in Meta-Analysis: There's a Reason We Don't Let Dodo Birds Tell Us Which Psychotherapies Should Have Prizes

William R. Shadish, Jr., and Rebecca B. Sweeney
Memphis State University

In primary studies, psychotherapy researchers frequently search for mediator and moderator variables that can help them understand the relationship between treatment and outcome. Yet a review of past psychotherapy meta-analyses revealed that none examined the possible role of mediator variables; and although all of them searched for moderators of study outcome, that search was generally not as complete as it could have been. This article illustrates methods for studying such mediator and moderator variables in meta-analysis, discusses their advantages and disadvantages, and shows how the inclusion of these variables can change interpretation of meta-analytic results. In particular, the perennial interpretation of past psychotherapy meta-analyses that therapeutic orientation makes no difference to outcome—or as the dodo bird put it: "Everyone has won and all must have prizes"—may be wrong. Orientation may make significant difference, but only by virtue of its moderating and mediating effects.

To the best of our knowledge, all meta-analyses ever done have concluded that (on the average) clients receiving psychotherapy do better than clients not receiving psychotherapy. In fact, the computation of average therapy effects over studies is the defining strength of meta-analysis. But this strength leads to a criticism of meta-analysis: Knowledge of average effects says nothing about when, where, why, and how therapy works. The latter questions concern mediators and moderators of therapy outcome. The present article describes methods for addressing such questions in meta-analysis.

Moderators and mediators are third variables that help researchers to understand the relationship between independent and dependent variables (Baron & Kenny, 1986). "A moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relations between an independent or predictor variable and a dependent or criterion variable" (Baron & Kenny, 1986, p. 1174). Moderators cause statistical interactions. Some moderator variables are categorical. Suppose, for example, that behavioral therapies yielded high effect sizes on behavioral presenting problems but low effect sizes on nonbehavioral presenting problems, with the opposite pattern emerging for nonbehavioral therapies. Presenting problem is then a categorical moderator. Other moderators are continuous. An example would be if behavior therapies produced moderate effect sizes no matter how many years' experience a therapist had, but nonbehavioral therapies produced low effect sizes for therapists with few years' experience and high effect sizes for therapists with many years.

Mediators reflect "the generative mechanisms through which the focal independent variable is able to influence the dependent variable of interest" (Baron & Kenny, 1986, p. 1173). The independent variable causes the mediator, which then causes the outcome. For example, suppose that behavioral orientation to therapy (the independent variable) causes the therapist to assess couple communication (a first mediator), with the assessment leading the therapist to change some of those communications (the second mediator), which then leads to increased marital satisfaction (the dependent variable). These mediators of psychotherapy outcome are often called *therapy process*. Not all therapy processes mediate therapy outcomes, because therapy processes may be irrelevant to outcome. Furthermore, some *research* processes, such as reactivity of measurement, also mediate study outcome. If we knew the key processes mediating positive outcome, we could more confidently produce such results.

The preceding discussion oversimplifies more complex and subtle matters. For instance, the same variable can be both a moderator and a mediator in the same model, and mediators can be nonlinear or nonrecursive. Interested readers will find a number of more sophisticated treatments (Aiken & West, 1991; Bollen, 1989; James & Brett, 1984; Smith & Sechrest, 1991; Snow, 1991) that we can only allude to given space constraints.

Traditional Analyses in Meta-Analysis

Traditionally, meta-analysts report an average effect size over studies and then report breakdowns of effect sizes by subgroups. In a classic example, Smith, Glass, and Miller (1980)

This research was supported in part by grants to William R. Shadish, Jr., from the National Institute of Mental Health (R01MH41097) and from the Russell Sage Foundation. Support was also received from a Centers of Excellence grant awarded to the Department of Psychology, Memphis State University, by the State of Tennessee.

We wish to thank two anonymous reviewers for very helpful comments on a draft of this article. We particularly want to thank Steve West for a review that may have been better than our manuscript; we borrowed his ideas shamelessly throughout, regretting only that he could not correct all the remaining problems in this article.

Correspondence concerning this article should be addressed to William R. Shadish, Jr., Center for Applied Psychological Research, Department of Psychology, Memphis State University, Memphis, Tennessee 38152.

reported an average effect size of $d = .85$ over 475 controlled studies of psychotherapy, where

$$d = \frac{X_T - X_C}{s}$$

and where X_T is the mean posttest score for the treatment group, X_C is the mean posttest score for the control group, and s estimates the standard deviation. Then they reported breakdowns of this statistic by such variables as type of therapy, type of outcome, and diagnostic type. They found, for example, that behavioral therapies yielded $d = .98$, verbal therapies yielded $d = .85$, and developmental therapies yielded $d = .42$. Such breakdowns are reported in all 19 psychotherapy meta-analyses we located in recent years in *Psychological Bulletin* (Berman, Miller, & Massman, 1985; Berman & Norton, 1985; Bowers & Clum, 1988; Casey & Berman, 1985; Dush, Hirt, & Schroeder, 1989; Hazelrigg, Cooper, & Borduin, 1987; Matt, 1989; Miller & Berman, 1983; Robinson, Berman, & Neimeyer, 1990; Shapiro & Shapiro, 1982) and *Journal of Consulting and Clinical Psychology* (Benton & Schroeder, 1990; Christensen, Hadzi-Pavlovic, Andrews, & Mattick, 1987; Dew, Bromet, Brent, & Greenhouse, 1987; Dobson, 1989; Hahlweg & Markman, 1988; Nietzel, Russell, Hemmings, & Gretter, 1987; Shoham-Salomon & Rosenthal, 1987; Steinbrueck, Maxwell, & Howard, 1983; Weisz, Weiss, Alicke, & Klotz, 1987).

Analysis of the significance of differences among categories is, in fact, a test of whether the variable is a moderator. Consider why. Imagine that $d = .40$ for behavioral treatment and $d = .20$ for nonbehavioral treatment. Consider how these two effect sizes might be produced in a primary study. If the dependent variable has a pooled standard deviation of 10, the effect size of .20 if nonbehavioral therapy would yield $X_T = 12$ and $X_C = 10$; and an effect size of .40 would result if behavioral therapy $X_T = 14$ and $X_C = 10$. Interactions are a function of the significance of differences among cell deviation scores. Specifically, interaction score = group mean − (row effect + column effect + grand mean), where row effect = row mean − grand mean, and where column effect = column mean − grand mean (Rosnow & Rosenthal, 1989). Computing interaction scores using these formulas, and graphing the results, yields the traditional "crossed lines" interpretation of interactions. The significance of the interaction must still be tested (not done by two meta-analyses cited previously). Hence analyzing differences in effect size between two categories is, in meta-analysis, a test for a moderator variable.

Smith et al. (1980) then used multiple regression to sort out redundancies among the moderators they tested. Their regression procedures have since been improved in two ways. One is to analyze effect sizes aggregated at the study level rather than individual effect sizes, because multiple effect sizes within studies are dependent, violating important statistical assumptions. The other is using weighted least squares analyses that give more weight to studies with larger sample sizes on the principle that they more accurately estimate population parameters (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Such improved regression analyses of first order moderators are widely available. But only 6 of 19 meta-analyses cited previously used such regressions, mostly not using weighted least squares. Primary researchers long ago rejected the use of multiple *t* tests in favor of more appropriate analyses. Most meta-analysts have yet to catch up.

In this article, we focus on procedures for testing higher order moderator effects, which is the major lacuna in meta-analysis. We use standard regression approaches to testing two-factor interactions using product terms (Cohen & Cohen, 1983) incorporating weighted least squares techniques for meta-analysis (Hedges & Olkin, 1985). Our primary interest is in variables that might moderate the effects of behavioral versus nonbehavioral theoretical orientation in psychotherapy, because the relative efficacy of behavior therapies has been a matter of great debate in the meta-analytic literature. However, the procedures we use generalize to tests of interactions with more levels and more factors. Of 19 meta-analyses previously cited, only 6 investigated higher order interactions; only one used weighted least squares.

In contrast to moderator variables, meta-analytic searches for mediator variables are virtually nonexistent. Baron and Kenny (1986) describe a simple regression strategy that can be implemented without any special analytic knowledge beyond ordinary regression; in fact, most nonrecursive path models can be analyzed using ordinary regression techniques (Bollen, 1989). An attractive alternative is the analysis of path models using structural or simultaneous equation models. Testing such models is now within the grasp of most researchers with the implementation of user-friendly structural equation programs like EQS (Bentler, 1989)—although users will benefit from more extensive statistical knowledge in using such programs. Of the 19 psychotherapy meta-analyses cited previously, none searched for mediators. However, in a meta-analysis about employee decisions to unionize, Premack and Hunter (1988) presented a simple path analysis in which wage level caused extrinsic satisfaction, which caused satisfaction with administration, which caused instrumentality of unionization, which caused a unionization decision. The present article develops this structural equation approach to mediators in meta-analysis and discusses its strengths and weaknesses.

In summary, then, the search for moderators has been relatively simplistic in meta-analysis, and the search for mediators has been largely nonexistent. More is possible, and we will demonstrate some of these possibilities. However, we would stress that our purpose is exploratory and didactic. Many of the procedures we suggest incur significant problems for which only partial or sometimes no answer yet exists. We present these procedures to open debate about the agenda of problems to be addressed in this crucially important area.

Method

The data used in this study are taken from a completed meta-analysis (Shadish et al., 1991; Shadish, in press), but we reanalyze the data in new ways to extend our past findings. Briefly, a total of 163 randomized controlled studies of the effects of marital and family psychotherapies with distressed clients were coded for effect size and potential predictor variables. Of these, 71 studies that compared therapy with a control group at posttest are used in this article; 38 were published articles or book chapters, and 33 were unpublished, almost entirely dissertations. Cohen's (1988) d is the measure of effect size. When sufficient information to compute d was not available, we computed best estimates of effect size using available statistics. Effect sizes reported only as nonsignificant were coded as zero. Effect sizes were corrected for small sample bias (Hedges & Olkin, 1985, p. 81, Equation 10), and multiple effect sizes within studies were aggregated to the study level. Study effect size is weighted by the inverse of its variance, thus giving more weight to studies with larger samples (Hedges & Olkin, 1985).

Results

Mediator Variables in Meta-Analysis

Theoretical orientation to psychotherapy (behavioral versus nonbehavioral) is the independent variable, and effect size is the dependent variable. Behavioral treatments yielded $d = .56$, and nonbehavioral treatments yielded $d = .54$, both of which are significantly different from zero but not from each other ($Q_b = .03$, $df = 1$, $p > .05$; Hedges & Olkin, 1985). Although one might conclude that these orientations make no difference to therapy outcome, hypothesizing that orientation has a direct effect on outcome may be less realistic than hypothesizing that orientation effects are indirect—mediated through choices that researchers with particular orientations make in therapy and research. So we formulated a mediational model, generally shaped by three considerations. First, past authors have hypothesized that treatment is more effective when it is fully implemented (Sechrest, West, Phillips, Redner, & Yeaton, 1979) and that "manualized" treatments are often more effective (Smith & Sechrest, 1991). Hence we included measures related to treatment implementation and standardization as mediators between therapy orientation and outcome. Second, past psychotherapy meta-analyses often report that "reactive" measures (Smith et al., 1980) yield larger effect sizes than other measures. Hence we tested models that included various assessments of reactivity, ending with whether a dependent variable assessed a behavior. Behavioral measures may be more reactive to behavioral treatments by virtue of being more specifically tailored to the interventions. Third, publications tend to yield higher effect sizes than unpublished works. We suspected that reports of behavioral treatments might be more likely to be published because behavioral researchers are overrepresented in university settings where publication pressures are higher.

This model (and variants on it) was tested with generalized least squares estimation in EQS. EQS does not allow direct weighting of meta-analytic data. However, Hedges and Olkin (1985) describe how to create appropriately weighted covariance matrices in standard statistical packages like SPSS Regression (SPSS, Inc., 1990). These matrices can be downloaded as input into EQS.

Models rarely fit on first test. Subsequent specification searches capitalize on chance, so the best fitting model may not replicate on new samples. Commonly, one would deal with this by using both a model development and a cross-validation sample. This strategy is problematic in meta-analysis, because the low number of studies being analyzed may be too small to split into smaller subsamples. Our tentative solution was to randomly split *effect sizes* (not *studies*) into model development and cross-validation subsamples. Studies have multiple effect sizes, which when split usually still leave some effect sizes from a given study in both subsamples. This procedure keeps the overall sample size of studies in each subsample at about its original level. However, the resulting subsamples are clearly dependent, only weakly testing cross-validation. Hence we address this vexing matter further in the Discussion section.

Results were as follows. The final model differed only slightly from the hypothesized model, including the following added paths: Publication status also had indirect effects on outcome through treatment standardization and implementation, and dissertations were less likely to be standardized but more likely to be implemented as intended. Fit statistics for this model in the model development sample were χ^2 (6, $N = 67$) = 4.23, $p = .65$, Bentler-Bonnet normed fit index (NFI) = .99, comparative fit index (CFI) = 1.00, so the model fits extremely well. Path coefficients for this model are presented in Figure 1 (not in parentheses). Adding a direct path from behavioral orientation to outcome did not significantly improve model fit. Fit statistics in the cross-validation sample were χ^2 (6, $N = 70$) = 10.48, $p = .11$, NFI = .98, CFI = .99, again supporting the fit. Path coefficients for this subsample are in parentheses in Figure 1.

This analysis suggests a different interpretation of orientation effects. Orientation makes a considerable difference, but this is due to its effects on mediators, not on study outcome itself (the total effects of orientation on outcome are still about zero). Some mediators concern therapy process; others concern methodological choices that researchers make in research. After all, we are trying to understand study outcomes in meta-analysis. Those outcomes are a function of more than just therapy.

Mediational models make far more plausible assumptions about the processes that generated study outcomes than do nonmediational models. In fact, simple univariate tests of both mediators and moderators are almost surely incorrect when taken literally. There is little reason to think that only one variable, such as theoretical orientation, is solely responsible for all variation in outcome. Multiple regression equations are slightly more realistic models in assuming that study outcomes are multiply determined, but they are not as plausible as mediational models. Consider the nonmediational model in Figure 2, which uses the same variables as those in Figure 1. Figure 2 is not a standard regression model, because the latter estimates all correlations among predictors and would have zero chi-square and degrees of freedom. But it resembles a standard regression in that both lack mediator variables. Although Figures 1 and 2 are not directly comparable, the fit of the nonmediational model, χ^2(10, $N = 67$) = 27.49, $p = .002$, NFI = .94, CFI = .96, is apparently not as good as the fit of the mediational model. Nor could the fit be improved significantly by adding, for example, correlations between behavioral orientation and both behavioral dependent variable and publication status, χ^2(8, $N = 67$) = 21.43, $p = .006$. More importantly, the theoretical implications of Figure 2 are less plausible than those in Figure 1. For example, the theoretical orientations of psychotherapy researchers

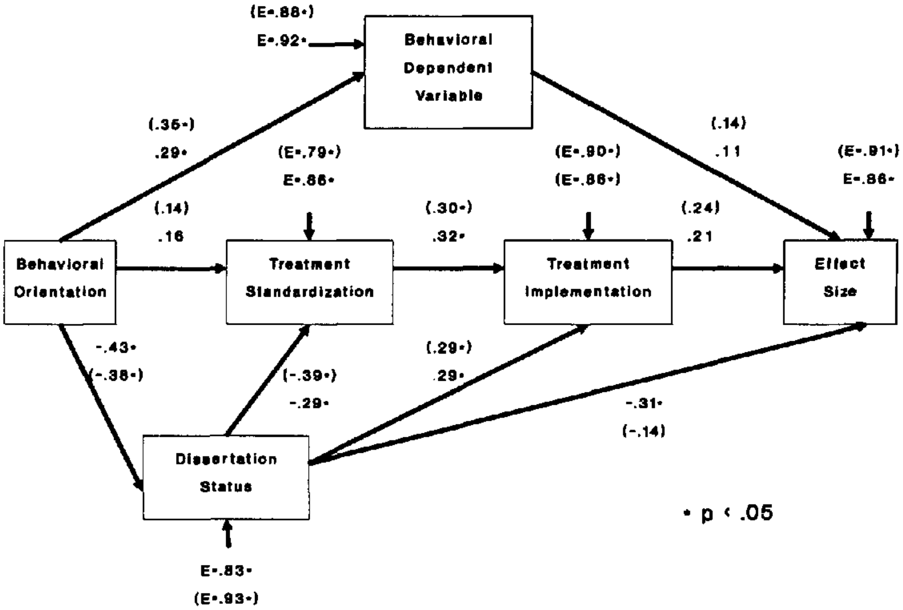

Figure 1. Modeling mediators in meta-analysis: An example. (E = error.)

mostly precede study design, leading to certain choices. For example, behavioral researchers are trained to use behavioral dependent variables specific to treatment (e.g., Barlow & Hersen, 1984, pp. 133–134). Nonmediational models cannot represent this causal process. When one of two models fits the data slightly better and also makes more theoretical sense, it ought to receive serious consideration.

Some readers will object that this approach is a form of causal modeling in correlational data, and significant problems with causal modeling are known (Freedman, 1987). We give this objection extended consideration in the Discussion section.

Moderator Variables in Meta-Analysis

In this section, we illustrate meta-analytic exploration of higher order moderator effects. The independent variable of interest is behavioral versus nonbehavioral orientation, and the dependent variable is effect size. Twenty-eight potential moderator variables studied here included location of treatment in a university, year of publication, proportion of effect sizes reported only as nonsignificant, number of measures reported in the study, treatment dosage (Number of Sessions × Number of Minutes per Session), locus of presenting problem (child, adult, couple, family, extrafamilial), treatment modality (who was

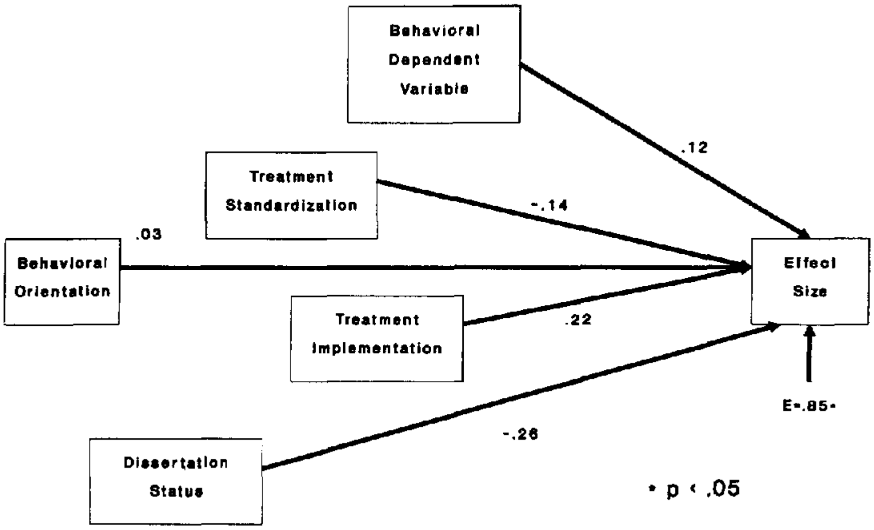

Figure 2. Model without mediating variables.

seen in therapy), gender of therapist, therapist experience, therapist mental health professional degree, experimenter allegiance, present versus historical focus of therapy, use of communication training in therapy, treatment standardization, treatment implementation, reactivity of dependent variable, specificity of dependent variable, manipulability of dependent variable, referral source of client, blindedness of experimenter to treatment condition, use of university-based clients, behavioral versus nonbehavioral dependent variable, self- versus other ratings as dependent variable, experimenter blindedness to dependent variable, differential attrition from conditions, number of therapists, kind of outcome, and study sample size.

Earlier we specified a priori, theoretically based mediational hypotheses to test. By contrast, remarkably few specific statements about potential interactions exist in the literature. Beutler's (1991) schema suggested about 1½ million possible interactions, very few of which have ever been explored empirically. Those that have been explored focus largely on patient–treatment interactions, with inconsistent results (Smith & Sechrest, 1991). Some other interactions are of little relevance to marital and family therapies, such as the superiority of systematic desensitization for phobias versus the superiority of other forms of behavior therapy for obsessive–compulsive disorder. Hence, although our results are quite interesting, our approach is far more exploratory than we would like. Perhaps these analyses will spark development of more specific interactive hypotheses that can be tested in meta-analyses in other areas.

Although Hedges and Olkin (1985) do not describe interaction tests, we adapted their regression techniques by computing product terms to represent interactions (Cohen & Cohen, 1983). To prevent colinearity of product terms with component multipliers, we centered each variable by subtracting the sample mean from each observation (Cronbach, 1987; Marquardt, 1980). The interaction is significant if its associated beta weight is significant. When centering did not reduce colinearity, we entered the main effect terms first, entered the interaction term second, and examined the significance of the increase in the multiple correlation on the second step with a chi-square difference test (Cronbach, 1987).

We computed 28 regression equations, one for each variable previously listed. Each regression used three predictors: behavioral versus nonbehavioral orientation, one of the 28 moderators, and their product term. We assessed the significance of the results both with and without a Bonferroni correction. The latter correction helps prevent capitalization on chance, requiring the interaction beta weight $\alpha = .05/28 = .0018$ or less to be significant. One interaction term reached this level: location of treatment in university setting. But this procedure may be too conservative (see Discussion section), so we also report four additional interactions that were significant at $p < .05$: (a) specificity of measurement, (b) manipulability of measurement, (c) reactivity of measurement, and (d) study sample size. Relevant effect sizes are presented in Table 1.

Table 1
Variables That Moderated the Effects of Theoretical Orientation

Variable	Behavioral	Nonbehavioral
Setting		
University	.73	.36
Nonuniversity	.35	.36
Measurement reactivity		
High	.68	.58
Medium	.48	.39
Low	.07	.48
Measurement specificity		
Treatment specific	.72	.46
General family/marital measure	.50	.44
General measure	.13	.58
Measurement manipulability		
Not very	.15	.76
Moderately	.58	.55
Very	.55	.46
Number of subjects		
Below median	.77	.39
Above median	.45	.48

University setting. The multiple $R = .28$ ($Q_r = 11.13$, $df = 3$, $p < .05$), with the interaction standardized partial $\beta = .28$ ($p = .0012$). We used Hedges and Olkin's (1985) categorical tests to do simple main effect follow-ups, adding a Bonferroni correction ($\alpha = .05/4 = .0125$) to adjust for the number of simple main effects. Two simple main effects were significant. Behavioral studies in university settings (e.g., academic campus, medical school) had significantly higher effect sizes than behavioral studies in nonuniversity settings (e.g., community mental health centers, private practice, school system, prison; $Q_b = 6.27$, $df = 1$, $p < .025$) and than nonbehavioral studies in university settings ($Q_b = 7.13$, $df = 1$, $p < .01$). To explore this finding, we computed Bonferroni-corrected t tests using the other 27 moderators as dependent variables and found only that behavioral treatments in universities used more specific dependent variables than did other studies ($t = 3.69$, $df = 73$, $p < .001$).

Measurement reactivity. Smith et al. (1980) found that more reactive measures yielded higher effect sizes. Using their reactivity scale, $R = .26$ ($Q_r = 17.94$, $df = 3$, $p < .0005$), with the interaction $\beta = -.18$ ($p = .001$). We computed simple main effect follow-ups, but with one modification. Because reactivity had three levels (we collapsed Smith et al.'s original five levels), we followed up significant simple main effects by computing all possible pairs of Bonferroni-adjusted confidence intervals, declaring that members of a pair (say, low vs. high reactivity) were different if their confidence intervals did not overlap. Results were that behavioral studies with less reactive measures had significantly lower effect sizes than behavioral studies with medium reactivity, which had lower effect sizes than behavioral studies with highly reactive measures ($Q_b = 16.90$, $df = 2$, $p < .0005$). Also, behavioral studies using measures with low reactivity yielded lower effect sizes than nonbehavioral studies with low reactivity ($Q_b = 7.99$, $df = 1$, $p < .01$).

Measurement specificity. Specificity had three levels: (a) specific—measures directly constructed from or related to the goals of treatment; (b) general family or marital—not specifically tailored to treatment, but a general family or marital measure; (c) general—measures tangentially related to treatment. $R = .29$ ($Q_r = 23.34$, $df = 3$, $p < .0005$), and the interaction $\beta = .17$ ($p = .0042$). For behavioral studies, specific measures yielded significantly higher effect sizes than general family or marital measures, which yielded significantly higher effect sizes than general measures ($Q_b = 18.70$, $df = 2$, $p < .0005$).

Measurement manipulability. This variable had three levels: (a) not very manipulable—measures not easily controlled by clients or therapists, (b) moderately manipulable—manipulable at a cost to the client (e.g., an observer-rated problem resolution task requiring spouses to comply with treatment recommendations that are inconsistent with their normal behavior), (c) very manipulable—manipulable at no cost for the client or therapist (e.g., self-reports, therapist ratings). The multiple $R = .22$ ($Q_r = 10.01$, $df = 3$, $p < .05$), and the interaction $\beta = -.16$ ($p = .0188$). On follow-up, effect sizes from behavioral studies with not very manipulable outcome measures were significantly lower than those from nonbehavioral studies with such measures ($Q_b = 4.91$, $df = 1$, $p < .05$).

With one exception, then, the findings for reactivity, specificity, and manipulability were consistent: To the extent that they had an effect, measures higher (lower) on these characteristics yielded higher (lower) effect sizes, and this seemed to affect behavioral treatments more than nonbehavioral ones. The exception is the high effect sizes yielded by nonbehavioral studies with not very manipulable measures. Inspection of the measures actually used in the four studies in this cell confirms that they are probably not very manipulable, including achievement test results and school records of truancy and suspensions (D'Elio, 1982), pulmonary function (Herold, 1980), marital reconciliations (Matanovich, 1970), and various recidivism records (McPherson, 1980). We cannot explain this finding, except by chance given the few studies in the cell.

Number of subjects. In this analysis, centering did not remove colinearity, so we tested the significance of the interaction as discussed earlier. The overall $R = .36$ ($Q_r = 18.29$, $df = 3$, $p < .0005$); the chi-square difference test suggested a significant interaction ($Q_{\mathrm{DIFF}} = 9.58$, $df = 1$, $p < .01$). For interpretation, sample size was dichotomized at the median, and simple main effects were computed. Behavioral studies with few subjects had higher effect sizes than both behavioral studies with more subjects ($Q_b = 4.87$, $df = 1$, $p < .05$) and nonbehavioral studies with few subjects ($Q_b = 4.42$, $df = 1$, $p < .05$). One explanation might be that reviewers tend to reject studies with nonsignificant findings (Greenwald, 1975), so studies with small sample sizes must have larger effect sizes to attain significance and be published. This interaction would be consistent with a finding that behavioral studies with few subjects were more often published than nonbehavioral studies with few subjects. A 2 × 2 (behavioral–nonbehavioral, published–dissertation) chi-square revealed a trend in that direction, $\chi^2(1, N = 38) = 2.90$, $p = .09$.

Discussion

Substantive Issues

Two interesting points emerged from these analyses. First, common lore in psychotherapy meta-analysis is that orientation makes no difference to outcome, at least not after adjusting for covariates. In the abstract of this article, we quoted Luborsky, Singer, and Luborsky's (1975) famous dodo bird conclusion to that effect: "Everyone has won and all must have prizes" (p. 995). The analogy is to the dodo bird in *Alice in Wonderland*, who awarded all contestants in the race a prize. Luborsky et al., of course, meant the analogy facetiously, to highlight the fact that all psychotherapies seem to work equally well. But the analogy has an obvious flaw: Dodo birds are not very smart, so it is not clear why we would let them award prizes to begin with. That only happens in Wonderland. The dodo bird conclusion is an artifact of the dodo bird's failure to look for plausible mediators and moderators (as Luborsky et al. acknowledged in ending their article). It is far less plausible to think that orientations to therapy have direct effects on outcome than to think that they have indirect effects through subsequent therapeutic or scientific choices, or to think that they have moderated effects in which outcomes depend both on therapy orientation and on the level of other variables. Our results support this conceptualization. In mediational models, behavioral orientations do affect mediators that themselves then affect outcome. In moderator models, behavior therapies result in either better or worse outcomes than nonbehavioral therapies, depending on the levels of other variables involved. Such a conclusion makes much more theoretical sense than the no-difference finding of the dodo bird.

Second, these analyses highlight the crucial role played by characteristics of outcome variables in primary studies, a finding as old as psychotherapy meta-analysis (Smith et al., 1980). The novelty in our findings is that this effect may influence behavior therapies more than nonbehavior therapies. We can think of at least one reason why this might be the case. For example, behavior therapists tailor treatment to specific target behaviors (Barlow & Hersen, 1984). Suppose therapy is a zero-sum game: Therapeutic outcome is proportional to therapy inputs, and the amount of inputs is finite. Devoting more inputs to specific target behaviors means devoting fewer inputs to other outcomes. So the former change more and the latter less. If nonbehavioral therapies target inputs to outcomes at all levels of specificity, the changes they produce would be more evenly distributed.

We do not want to make too much of these findings, and we particularly do not want to be perceived as advocates of one theoretical orientation. We do want to suggest that meta-analyses have been far too simplistic to support strong inferences about whether things like theoretical orientation make any difference to therapy outcome. We suspect that future meta-analysts who look at the matter will confirm that orientation effects can be significant in models that include mediators and moderators. We are far less confident that they will also confirm our explanations for them, or confirm our specific findings about behavioral orientations, if for no other reason than that the crude behavioral–nonbehavioral dichotomy we used should be replaced by more specific and subtle coding. But the principle, we suspect, will endure.

Problems With Analyzing Moderator Effects

Earlier, we suggested that using Bonferroni corrections may be too stringent in the search for significant interactions in meta-analyses. Finding significant moderator effects in primary research is difficult enough, for a host of reasons. Those reasons are even more problematic in meta-analysis. The reliability of product terms is much lower than the reliability of the component main effect terms (Aiken & West, 1991; Chaplin, in press; Cronbach & Snow, 1977). This problem is greatly exacerbated in meta-analysis where variables of interest are often measured with one, often dichotomous item. Use of dichotomous or polychotomous variables also severely limits the ability to look for nonlinear interactions. Some improvement in this problem could in principle be ameliorated with better meta-analytic measurement, by more attention to continuous variables, and by using latent variable models. These ought to be a high priority for future meta-analytic work. But it seems unlikely that meta-analysts will often be able to approach the quality of measurement in most primary studies, so this already severe problem is likely to remain even worse in meta-analysis.

In addition, power is adversely affected by the small number of studies often used in meta-analysis. The problem is even worse when some cells of a meta-analytic factorial design contain very few studies. For example, we found only one study of a behavioral treatment with an experimenter who was completely blind to treatment. This can be solved only by doing more primary studies. The good news is that power may be less of a problem in meta-analysis than in primary research (Osburn, Callender, Greener, & Ashworth, 1983; Sackett, Harris, & Orr, 1986; Spector & Levine, 1987). The reason is that the unit of analysis in primary research is usually an individual, but the unit in meta-analysis is a study-level effect size that is an aggregate of these primary units. Observations based on such aggregates should generally have smaller standard errors than those based on individual subjects. But even this hypothesis is controversial (Durlak & Lipsey, in press) and warrants a close look by statisticians accustomed to computing power in similar situations such as cluster sampling.

Chaplin (1991) suggests more reasons why interactions will be elusive in meta-analysis:

> The extremely large number of possible moderator effects makes it unlikely that any one of those effects will be large. Moreover, the loss of degrees of freedom that accompanies the addition of moderators to the model, coupled with the fact that moderator effects are residualized from main effects, makes the statistical evaluation of moderator effects a low-power enterprise. (p. 2)

These problems, of course, seem no more or less prevalent in meta-analysis than in primary research.

We did not explore another strategy for studying interactions that may partly remedy these problems—aggregation of effect sizes generated from interaction terms in primary studies. If, for example, more than one study reported an interaction term between treatment type and university setting, effect sizes from these terms could be computed and then averaged. Such integration should yield more powerful analyses than those generated in this article. The viability of this alternative is largely unexplored, but the number of studies available for integrating such interaction terms is probably quite small relative to the number of studies to which the present strategy can be applied. In addition, such integrations will be more difficult than aggregating main effects unless good measurement of the strength of treatment and the strength of the moderator variable is reported in primary studies (Cooper & Arkin, 1981), which is often not the case.

Despite the fact that we used the traditional $\alpha = .05$ level of significance in this article, that level may not always be best in searching for interactions (Smith & Sechrest, 1991). We might carefully consider the relative costs of detecting versus failing to detect these effects. Particularly if we view meta-analysis as a tool for generating hypotheses that can be tested in new primary studies, perhaps we can accept a less stringent alpha level—particularly for locating otherwise elusive interactions (Snow, 1991).

Mediational Models and Causal Modeling

Previously we noted that our approach to mediational models in meta-analysis is a form of causal modeling. Because criticisms of such models are legion, one must wonder how causal modeling can be justified in meta-analysis. We suggest two conditions to be met to justify that use. First, the researcher must be interested in causal inferences. Second, stronger experimental or quasi-experimental methods must be impossible to implement or insufficient by themselves. These conditions hold in meta-analysis.

On the first point, the language of meta-analysis often invokes causation. For example, Benton and Schroeder (1990) conclude that "the results of the current meta-analysis indicate that social skills training leads to significant improvements in the social behavior of schizophrenics. . . . Similarly, training appears to have a positive impact on schizophrenics' perceptions of themselves" (p. 744). Berman et al. (1985) suggest that experimenter "allegiances may affect the outcome of a study" (p. 458). Whether these authors had causal hypotheses in mind is not the point; the ordinary discourse of meta-analysis invokes terms such as *impact* or *affect* that are intimately and logically tied to causation. Even if the interest is only in generating causal hypotheses to be tested later in experiments, this is still an interest in causation.

The second condition also holds in meta-analysis. Just as we distinguish between experimental, quasi-experimental, and nonexperimental methods in primary studies, we can distinguish between experimental, quasi-experimental, and nonexperimental meta-analysis. In such terms, meta-analysis is mostly nonexperimental, occasionally quasi-experimental, and probably never experimental. Consider why. In primary experiments, we usually facilitate causal inference by assigning subjects randomly to levels of the independent variables. Lacking this, threats to internal validity like selection bias thwart causal inference. A truly experimental meta-analysis would have to follow the same logic: Studies would have to be assigned to levels of the independent variables of interest at random. Because this does not hold, inferences about relationships between any given study characteristic and effect size are confounded. For example, studies are not assigned randomly to behavioral orientation. Rather, the choice to use behavior therapy is confounded in unknown ways with other choices such as using a behavioral dependent variable, conducting the study in a university, or standardizing treatments. Hence meta-analysis is probably never experimental (see Shadish, in press, for a possible exception).

Meta-analysis is sometimes quasi-experimental. The essence of quasi-experimentation is the use of a design feature to minimize a threat to causal inference. Meta-analysts occasionally do this. For example, Berman et al. (1985) found that researcher allegiance may have affected study outcome, so allegiance would have to precede outcome. But Berman et al. realized that

> [A] possible problem with interpreting these results is that the allegiance of the investigators was determined by the introduction to the published article, and investigators might have written this material after they had seen their results. Therefore, the findings may have influenced how the introduction was written rather than the researcher's allegiance affecting the outcome of the study. (p. 455)

To address this threat, they introduced a new design feature "in which we designated investigators as having an allegiance only if they had indicated this preference in a work published prior to the study included in the review" (p. 455). Inasmuch as it is implausible to suggest that results of a study at Time 2 caused allegiance in a study at Time 1—because causation does not work backwards in time—this feature minimizes the particular threat to validity they identified.

Shadish et al. (1991) suggest another quasi-experimental design feature for meta-analysis, the use of within-study treatment–treatment comparisons that directly compare two treatments with each other in the same study. Such comparisons are often excluded from meta-analysis in favor of examining only treatment–control comparisons, partly because the latter are easier to analyze. But within-study treatment–treatment comparisons can often rule out some confounds between treatment and other study characteristics. For example, publication status is constant in a study and so can less easily confound inferences about which treatment is more effective; similarly, measurement characteristics are constant in such comparisons, because the effect size is computed on a single variable, again minimizing a confound with treatment. Shadish et al. (1991) and Shadish (in press) elaborate this rationale and suggest meta-analytic adaptations of cohort designs and nonequivalent dependent variable designs. Such design solutions to causal inference problems in meta-analysis need more attention.

Nonetheless, such quasi-experimental design features are rare in meta-analysis and are not uniformly applicable to all situations calling for causal inferences; so most meta-analytic data is nonexperimental (correlational) in nature (Louis, Fineberg, & Mosteller, 1985). This renders suspect any strong causal inference in meta-analysis. If so, the meta-analyst's job is to construct the most plausible models possible of the processes that generated study outcomes. Simple univariate or regression analyses are unlikely to yield well-founded causal hypotheses, because they are almost certainly misspecified; and incorrectly specified models yield estimates of effects that may be wrong in both magnitude and sign (Bollen, 1989). Hence models like Figure 1 are not only justified, they are essential to plausible causal inference in meta-analysis.

Having said all this, we must turn to the many limitations of these mediational models and statistical analyses. First and foremost, in meta-analysis we are trying to draw causal inferences mostly from correlational data. Elaborate path models that fit the data *may still be wrong.* In fact, given the difficulties of correctly specifying the model, any given model is almost surely wrong. The hope—and at the current stage it is just a hope—is that because these models are more plausible than the patently implausible univariate or regression models, the resultant causal hypotheses are incrementally less likely to be far astray.

Similarly, for a given set of variables, many different specifications of the model may fit the data equally well (Stelzl, 1986)—although not all such models will be plausible. In Figure 1, one could replace the causal relationship between behavioral orientation and dissertation status with the assumption that the two variables are correlated, and the fit statistics and path coefficients would be identical. Hence one must examine the plausibility of the direction of causality that is posited and determine whether causal or correlational relations are worth hypothesizing. In Figure 1, it is unlikely that effect size causes any variables in the model or that selection of a behavioral dependent variable causes one to investigate a behavioral treatment. Making paths causal rather than correlational is more speculative, but defensible for developing interesting causal hypotheses for future work. Similarly, not all plausible models will fit the data equally well. The analyst may be able to specify competing theories and show that one accounts for the data better than its competitors. Similarly, we can use nested models to test the worth of adding or subtracting particular plausible paths from models (Bentler & Bonnett, 1980), as when the addition of a direct path in Figure 1 between behavioral orientation and outcome did not improve the fit of the model.

In addition, different model specifications of the same variables can change the magnitude, direction, and significance of path coefficients. Hence it is critically important to conduct sensitivity analyses to assess the stability of path coefficients and test statistics under diverse model specifications. In developing the model in Figure 1, we analyzed scores of different models, some involving different path specifications among the same variables and others involving addition of different variables to the model. In 15 models that included a causal path from behavioral orientation to publication status, the coefficient was always significant and ranged from −.37 to −.46, remarkably stable over model specifications. Similarly, over many different specifications, theoretical orientation never had a significant direct effect on outcome.

Of course, these models still leave much unexplained—why dissertations have lower effect sizes, for example. It may be that (a) authors of journal articles drop nonsignificant findings before publishing them, whereas dissertations report complete results; (b) reviewer bias against null findings creates a publication bias and resulting file drawer problem; (c) dissertations include more measures than most other studies, but some of these measures are tangential and show lower effects; or (d) to save journal space, nonsignificant findings are reported only as nonsignificant in publications (and so coded zero in meta-analysis), but are reported in detail in dissertations. Some such hypotheses could also be tested; for example, adding number of measures as a mediator between dissertation status and effect size was never significant in any model we tested.

All of these criticisms are as true of univariate or regression models in meta-analysis as they are of our mediational models, because the criticisms stem from the correlational nature of meta-analytic data, not from the analytic technique. But other problems are particular to the analysis. First, although the statistics underlying programs like EQS are based on large sample theory (but see Tanaka, Panter, Winborne, & Huba, 1990), even the present meta-analysis, which is large relative to many others, has a rather small sample of studies. Hence the overall chi-square test that the model fits may not be rejected due to insufficient power, and the likelihood of finding significant path coefficients is lowered. Fit indices that are independent of sample size help remedy these problems (Bentler, 1990; Bollen, 1990), as may the previous observation that meta-analytic data may have more statistical power than primary data. The extent to which this mitigates sample size problems is currently unknown. Additionally, low power is partially addressed by finding some plausible models that can be rejected. In the present data, for example, some researchers might consider the model in Figure 2 to be plausible, but it did not fit the data. The null model used as the baseline in goodness of fit indices (Bentler & Bonnett, 1980) also failed to fit the data. Hence power is more likely to be an issue when trying to distinguish between models with similar but not identical levels of fit. Of course, sample size issues will also affect some other statistical approaches to meta-analysis. For example, when fixed effects regressor models do not fit, random regressor models are sometimes suggested. These, too, often require large sample theory.

Another problem is that meta-analytic predictors are often categorical and often nonnormally distributed, which can adversely affect maximum likelihood and generalized least squares estimates. Statistical inference in such conditions will be adversely affected, although substantive interpretation will not be affected (Tanaka et al., 1990). Muthen (1988) has developed alternative estimators for dichotomous and polychotomous data, but they require very large sample sizes. Arbitrary distribution theory methods in EQS could be useful, but they also require very large sample sizes. They also require analyzing raw data, precluding use of appropriately weighted covariance matrices as input. The latter trade-off may be undesirable, because weighted least squares analyses can yield quite different results and interpretations than ordinary least squares methods (Shadish, in press), and the former are theoretically preferable. LISREL (Joreskog & Sorbom, 1988) and PRELIS (Joreskog & Sorbom, 1986) suffer from a similar problem, so here is another area where more work is needed.

Another problem is that of modeling dependencies in the data set caused by, for example, multiple studies being done on the same subjects or multiple studies being done by the same investigator. Recent work by Weng (1990) suggests that some relatively simple models for analyzing such data may be feasible and ought to be explored further for their applicability to the present problem.

Structural modeling programs like EQS and LISREL are not the only statistical approach to estimating mediational models. Becker (in press) presents a quite different approach that extends the generalized least squares analysis proposed by Raudenbush, Becker, and Kalaian (1988) for meta-analytic data, with different advantages and disadvantages compared with our approach (Shadish, in press). In particular, it allows modeling within-study covariances that are not modeled in our approach. Becker also used a different method of constructing covariances among variables, again with its own advantages and disadvantages relative to the one used here (Shadish, in press). Finally, she also suggests developing hierarchical analysis models (Raudenbush & Bryk, 1985) for this problem. All these possibilities remain to be explored and contrasted.

A final problem stems from the exploratory nature of meta-analytic model development, which undoubtedly capitalizes on chance—although this is often just as true of model search procedures in standard regression analyses. If a meta-analyst simply tested one model in a confirmatory analysis, this problem would not occur. Very careful previous specification of models might increase the likelihood that such an initial model will fit. But even in the best of cases, initial models rarely fit with no modification at all, so further specification search almost always occurs. Hence cross-validation is essential. But obtaining a cross-validation sample by splitting *studies* into two random groups is very problematic, because the number of studies in most meta-analysis is already quite small. By using *effect sizes* rather than studies to split the sample, as we did, one can retain a nearly complete sample size of studies in both sets, because studies almost always report more than one effect size. But our cross-validation method yields dependent samples, because the same studies are represented in both samples. If so, cross-validation is artifactually high. This, too, needs further exploration, especially as to the trade-offs between our cross-validation technique and the alternative of splitting studies into groups. As an alternative, EQS provides bootstrap and jackknife procedures that can also be used to evaluate model stability on a single sample; however, this again requires raw data so would preclude using weighted least squares analyses. Fienberg (1980, p. 108) describes another cross-validation technique that may be useful in meta-analysis.

Two improvements to the mediational models we used are worth exploring. One is the use of latent variable models (although this may increase sample size requirements). Unreliability of measurement can severely and unpredictably bias path coefficients (Bollen, 1989). Latent variable models could help remedy this problem. Although the technology for implementing such models is straightforward, a major practical limitation is the generally poor measurement techniques in most meta-analyses—typically assessing a construct with only one item, a practice that would be considered unacceptable in many primary research areas. Shadish (in press) explores such models in meta-analyses and further discusses relevant issues. The second improvement would be to combine mediational models with moderator models. This can be done two ways. One is the use of product terms as outlined by Kenny and Judd (1984). The math and programming for this option are complex in the latent variable case, but are relatively simple in the observed variable case. The other is to use the multigroup features of EQS or LISREL to test for different influences in different groups, functionally a test of a moderator variable. This option is easy to implement to test simple interactions.

In the end, then, the trade-off between the analyses we use and more traditional univariate and regression models is this: Traditional models may raise fewer controversies about statistical issues, but they are almost certainly wrong as representations of the processes that generated study outcomes. Hence they almost certainly yield invalid inferences about relationships between variables. Our approach may yield more plausible inferences, but at the expense of incurring statistical problems that as yet we know little about.

Conclusion

Gordon Paul (1967) once asked the question that is probably the most oft-quoted question in psychotherapy research: "*What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" (p. 111). Paul aimed his remarks at primary psychotherapy researchers, but more reliable answers to this question should be available from syntheses of multiple studies. Hence we have outlined some methods that meta-analysts can use to address Paul's question. The methods have many problems. Our claim is not to have solved the problems, but to have provided some directions to pursue and some stimulation for others to take up the task.

References

Aiken, L. S., & West, S. G. (1991). *Testing and interpreting interactions in multiple regression.* Newbury Park, CA: Sage.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Becker, B. J. (in press). Models of science achievement: Forces affecting male and female performance in school science. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Bentler, P. M. (1989). *EQS: A structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Benton, M. K., & Schroeder, H. E. (1990). Social skills training with schizophrenics: A meta-analytic evaluation. *Journal of Consulting and Clinical Psychology, 58,* 741-747.

Berman, J. S., Miller, R. C., & Massman, P. J. (1985). Cognitive therapy versus systematic desensitization: Is one treatment superior? *Psychological Bulletin, 97,* 451-461.

Berman, J. S., & Norton, N. C. (1985). Does professional training make a therapist more effective? *Psychological Bulletin, 98,* 401-407.

Beutler, L. E. (1991). Have all won and must all have prizes? Revisiting Luborsky et al.'s verdict. *Journal of Consulting and Clinical Psychology, 59,* 226-232.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256-260.

Bowers, T. G., & Clum, G. A. (1988). Relative contribution of specific and nonspecific treatment effects: Meta-analysis of placebo-controlled behavior therapy research. *Psychological Bulletin, 103,* 315-323.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388-400.

Chaplin, W. F. (1991, January). Introduction: Differential assessment of persons. *The Score,* pp. 1-2.

Chaplin, W. F. (in press). Personality, interactive relations, and applied psychology. In S. R. Briggs, R. Hogan, & W. H. Jones (Eds.), *Handbook of personality psychology.* Orlando, FL: Academic Press.

Christensen, H., Hadzi-Pavlovic, D., Andrews, G., & Mattick, R. (1987). Behavior therapy and tricyclic medication in the treatment of obsessive-compulsive disorder: A quantitative review. *Journal of Consulting and Clinical Psychology, 55,* 701-711.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H. M., & Arkin, R. M. (1981). On quantitative reviewing. *Journal of Personality, 49,* 225-236.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analysis recently proposed. *Psychological Bulletin, 102,* 414-417.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington Press.

D'Elio, A. R. (1982). An investigation of the effectiveness of intervention strategies on juvenile anti-social behaviors. *Dissertation Abstracts International, 43,* 1466A. (University Microfilms No. 82-23594)

Dew, M. A., Bromet, E. J., Brent, D., & Greenhouse, J. B. (1987). A quantitative literature review of the effectiveness of suicide prevention centers. *Journal of Consulting and Clinical Psychology, 55,* 239-244.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57,* 414-419.

Durlak, J. A., & Lipsey, M. W. (in press). A practitioner's guide to meta-analysis. *American Journal of Community Psychology.*

Dush, D. M., Hirt, M. L., & Schroeder, H. E. (1989). Self-statement modification in the treatment of child behavior disorders: A meta-analysis. *Psychological Bulletin, 106,* 97-106.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge, MA: MIT Press.

Freedman, D. A. (1987). A rejoinder on models, metaphors, and fables. *Journal of Educational Statistics, 12,* 206-223.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1-20.

Hahlweg, K., & Markman, H. J. (1988). Effectiveness of behavioral marital therapy: Empirical status of behavioral techniques in preventing and alleviating marital distress. *Journal of Consulting and Clinical Psychology, 56,* 440-447.

Hazelrigg, M. D., Cooper, H. M., & Borduin, C. M. (1987). Evaluating the effectiveness of family therapies: An integrative review and analysis. *Psychological Bulletin, 101,* 428-442.

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

Herold, P. L. (1980). The effects of psychosocial intervention with children who have asthma on children's locus of control and self-esteem scores, and measures of physical status. *Dissertation Abstracts International, 40,* 5075B. (University Microfilms No. 80-07900)

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69,* 307-321.

Joreskog, K. G., & Sorbom, D. (1986). *PRELIS: A program for multivariate data screening and data summarization.* Mooresville IN: Scientific Software.

Joreskog, K. G., & Sorbom, D. (1988). *LISREL 7: A guide to the program and applications.* Chicago, IL: SPSS, Inc.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96,* 201–210.

Louis, T. A., Fineberg, H. V., & Mosteller, F. (1985). Findings for public health from meta-analyses. *Annual Review of Public Health, 6,* 1–20.

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

Marquardt, D. W. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association, 75,* 87–91.

Matanovich, J. P. (1970). The effects of short-term counseling upon positive perceptions of mate in marital counseling. *Dissertation Abstracts International, 31,* 2688A. (University Microfilms No. 70-24405)

Matt, G. E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin, 105,* 106–115.

McPherson, S. J. (1980). Family counseling for youthful offenders in the juvenile court setting: A therapy outcome study. *Dissertation Abstracts International, 42,* 382B. (University Microfilms No. 81-09550)

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavioral therapies: A quantitative review of the research evidence. *Psychological Bulletin, 94,* 39–53.

Muthen, B. O. (1988). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model* (2nd ed.). Mooresville, IN: Scientific Software.

Nietzel, M. T., Russell, R. L., Hemmings, K. A., & Gretter, M. L. (1987). Clinical significance of psychotherapy for unipolar depression: A meta-analytic approach to social comparison. *Journal of Consulting and Clinical Psychology, 55,* 156–161.

Osburn, H. G., Callender, J. C., Greener, J. M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology, 68,* 115–122.

Paul, G. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31,* 109–118.

Premack, S. L., & Hunter, J. E. (1988). Individual unionization decisions. *Psychological Bulletin, 103,* 223–234.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103,* 111–120.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

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.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71,* 302–310.

Sechrest, L., West, S. G., Phillips, M., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest and Associates (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Newbury Park, CA: Sage.

Shadish, W. R. (in press). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. M. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Shadish, W. R., Montgomery, L. M., Wilson, P., Wilson, M. R., Bright, I., & Okwumabua, T. M. (1991). *The effects of family and marital psychotherapies: A meta-analysis.* Manuscript submitted for publication.

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

Shoham-Salomon, V., & Rosenthal, R. (1987). Paradoxical interventions: A meta-analysis. *Journal of Consulting and Clinical Psychology, 55,* 22–28.

Smith, B., & Sechrest, L. (1991). Treatment of Aptitude × Treatment Interactions. *Journal of Consulting and Clinical Psychology, 59,* 233–244.

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

Snow, R. E. (1991). Aptitude–treatment interactions as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59,* 205–216.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72,* 3–9.

SPSS, Inc. (1990). *SPSS reference guide.* Chicago: Author.

Steinbrueck, S. M., Maxwell, S. E., & Howard, G. S. (1983). A meta-analysis of psychotherapy and drug therapy in the treatment of unipolar depression with adults. *Journal of Consulting and Clinical Psychology, 51,* 856–863.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21,* 309–331.

Tanaka, J. S., Panter, A. T., Winborne, W. C., & Huba, C. J. (1990). Theory testing in personality and social psychology with structural equation models: A primer in 20 questions. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 217–242). Newbury Park, CA: Sage.

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542–549.

Weng, J. L. (1990). *Aspects of covariance structure analysis with dependent observations.* Unpublished doctoral dissertation, University of California, Los Angeles.

Received February 6, 1991
Revision received May 13, 1991
Accepted May 29, 1991 ■