The effects of psychological therapies under clinically representative conditions: A meta-analysis

Article *in* Psychological Bulletin · August 2000

DOI: 10.1037//0033-2909.126.4.512 · Source: PubMed

4 authors, including:

Georg Matt
San Diego State University
133 PUBLICATIONS **6,027** CITATIONS
SEE PROFILE

Glenn A Phillips
Biogen Idec
74 PUBLICATIONS **1,345** CITATIONS
SEE PROFILE

Psychological Bulletin
2000, Vol. 126, No. 4, 512–529

0033-2909/00/$5.00 DOI: 10.1037//0033-2909.126.4.512

The Effects of Psychological Therapies Under Clinically Representative Conditions: A Meta-Analysis

William R. Shadish
The University of Memphis

Georg E. Matt
San Diego State University

Ana M. Navarro
University of California, San Diego

Glenn Phillips
The University of Memphis

Recently, concern has arisen that meta-analyses overestimate the effects of psychological therapies and that those therapies may not work under clinically representative conditions. This meta-analysis of 90 studies found that therapies are effective over a range of clinical representativeness. The projected effects of an ideal study of clinically representative therapy are similar to effect sizes in past meta-analyses. Effects increase with larger dose and when outcome measures are specific to treatment. Some clinically representative studies used self-selected treatment clients who were more distressed than available controls, and these quasi-experiments underestimated therapy effects. This study illustrates the joint use of fixed and random effects models, use of pretest effect sizes to study selection bias in quasi-experiments, and use of regression analysis to project results to an ideal study in the spirit of response surface modeling.

Does the treatment outcome literature suggest that psychological therapies work under clinically representative conditions? (Clinical representativeness, a concept we elaborate shortly, occurs when outcome studies use real clients and therapists in actual treatment settings and when treatment is not typically subjected to routine research standardization procedures such as the use of manuals, treatment compliance checks, and special pretherapy training.) This question has come to the fore in recent therapy literature for four reasons. First, after decades of meta-analytic results supporting therapy effectiveness, a credible challenge has been posed to those positive findings—most past therapy research was not conducted under clinically representative conditions, and some evidence suggests therapy might be ineffective under those conditions (see, e.g., Weisz, Donenberg, Han, & Weiss, 1995; Weisz, Weiss, & Donenberg, 1992).

Second, the methodological quality of research on therapy under clinically representative conditions has been criticized. Conducting methodologically strong, clinically representative effectiveness research can be difficult; for example, such studies often use weaker rather than stronger outcome designs (Weisz et al., 1992, 1995). Similarly, the 1995 *Consumer Reports* therapy survey ("Mental Health," 1995; see also Seligman, 1995) was criticized as methodologically inappropriate for drawing outcome conclusions (see, e.g., Nathan, 1998). So, the effects of these therapies remain open to challenge.

Third, third-party payers and governmental bodies want more evidence about real-world effects of psychological therapies (Barlow, 1994; Beutler, 1998; Pallack, 1995). Sometimes they call for therapists themselves to provide data about their patients through evaluations conducted in the therapist's office. The research literature on clinically representative psychological therapies is also germane. Unfortunately, little such research exists, and this lacuna is an obstacle for therapy researchers and practitioners alike.

Finally, the literature on practice guidelines (see, e.g., Nathan, 1998) is based partly on the assumption that therapy under clinically representative conditions is less effective than it could be if therapists used empirically supported psychological therapies that have been found efficacious in controlled research with a delineated population (Chambless & Hollon, 1998; Kendall, 1998). Research about therapies under clinically representative conditions test both whether there is a deficit in existing practice to be remedied and whether "treatments that are found to be efficacious in producing desirable gains in a research clinic [are] transportable to a community service setting" (Kendall, 1998, p. 5).

These interests might be addressed by new, methodologically sound studies of therapies conducted under clinically representative conditions. However, this strategy will not quickly produce many clinically representative studies given (a) the time it takes to do and publish such studies, (b) the difficulty in obtaining funding for them, (c) the reluctance of traditional research outlets to publish them given their frequent lack of certain controls, and (d) the resistance to randomized controls in clinically representative settings. So, a more timely answer requires taking as much advantage of the existing literature as possible.

William R. Shadish and Glenn Phillips, Department of Psychology, The University of Memphis; Georg E. Matt, Department of Psychology, San Diego State University; Ana M. Navarro, Department of Family and Preventive Medicine, University of California, San Diego.

We thank John Weisz, Bahr Weiss, and Vanessa Weersing for extensive correspondence about this program of research.

Correspondence concerning this article should be addressed to William R. Shadish, Department of Psychology, Campus Box 526400, The University of Memphis, Memphis, Tennessee 38152-6400. Electronic mail may be sent to shadish@mail.psyc.memphis.edu.

Fortunately, studies in the current literature already vary along a continuum of clinical representativeness. By assessing these existing studies along that continuum, we were able to (a) study the relationship of clinical representativeness to outcome and (b) generalize by extrapolation from that research to a clinically representative target of interest. The present study contributes to these goals using a quantitative review that builds on a preliminary review of previous meta-analytic work by Shadish et al. (1997).

Shadish et al. (1997) asked authors of previous therapy meta-analyses to send effect sizes for studies that met various criteria for clinical representativeness: The studies (a) were conducted in nonuniversity settings (e.g., community mental health centers, school systems); (b) used patients referred through usual clinical routes, not solicited by the experimenter; (c) used experienced, professional therapists with regular case loads; (d) did not use a treatment manual; (e) did not monitor treatment implementation; (f) used clients who were heterogeneous in personal characteristics (e.g., sex, socioeconomic status); (g) used clients who were heterogeneous in presenting problems; (h) used therapists who were not specially trained in the treatment; and (i) used therapists who were free to use a variety of treatments. Shadish et al. divided studies into three categories: The 54 studies (from 15 meta-analyses) that passed criteria (a) through (c) were in Stage 1; those passing (a) through (e) were in Stage 2; and the one study that passed all criteria was in Stage 3. Shadish et al. discussed problems with these criteria but argued that the criteria were a reasonable starting point for studying clinically representative therapy. Results suggested that effects from more clinically representative studies were about the same as effects reported in the original meta-analyses.

However, Shadish et al. (1997) were vulnerable to a variety of problems that can be remedied only by a new meta-analysis. The present study reports such a meta-analysis with eight specific improvements. First, this study uses a standardized protocol to code both substantive variables and effect sizes from scratch to reduce the possibility that the codings of criteria and effect size may have been inconsistent over the 13 meta-analysts who participated in Shadish et al. or inconsistent with the intended codings in Shadish et al. Second, Shadish et al. used only nine simply defined clinical representativeness codes; the present study expands the number of codes that pertain to clinical representativeness and adds many additional codes that measure other clinical and methodological characteristics of the studies. Third, this study refines the clinical representativeness criteria used by Shadish et al., for example, by allowing some clinically representative treatments to use standardized formats if that is common clinical practice (e.g., relaxation tapes). Fourth, this study uses a graduated scale of clinical representativeness rather than a stage system. Fifth, this study eliminates questionable studies used previously and includes new studies to better estimate the relationship between clinical representativeness and effects. Sixth, this study uses multiple regression to adjust for covariates that are confounded with clinical representativeness. Seventh, this study reports both fixed and random effects analyses to support different inferences about therapy. Finally, this study illustrates two meta-analytic innovations: using pretest effect sizes to explore bias in nonrandomized experiments and creating extrapolations from existing studies to results from a hypothesized ideal study of the effects of clinically representative therapy.

Method

Studies

All studies in this meta-analysis met the definition of therapy suggested by Weisz, Weiss, Alicke, and Klotz (1987): "Any intervention designed to alleviate psychological distress, reduce maladaptive behavior, or enhance adaptive behavior through counseling, structured or unstructured interaction, a training program, or a predetermined treatment plan" (p. 543). We eliminated studies using two criteria. We excluded treatments that used psychotropic medication as part of treatment because this confounds the effect of therapy with the effect of medication. Of course, some study participants were using medications either legal (e.g., a presenting problem of nausea from chemotherapy) or not (e.g., psychological treatment for adults addicted to opiates); in randomized studies (the majority of our studies), such use is not confounded with treatment. We also excluded purely preventive treatments that usually are not considered therapy. Otherwise, we were inclusive rather than exclusive. For example, some of our studies were conducted in hospital or residential settings where the control groups might have gotten some intervention, some studies used bibliotherapy, some involved only parent or teacher training, some used clients with medical problems, and some occurred in schools.

Our sample was 90 studies from three sources. Sample A consisted of 41 studies from Shadish et al. (1997) that met their Stage 1 criteria for clinical representativeness. We excluded 8 of their 54 Stage 1 studies that (a) used clients without a psychological, behavioral, or emotional problem of a kind that therapists are asked to treat, such as unselected nursing students receiving personal growth groups and unselected normal kindergartners receiving psychological education (Amerikaner & Summerlin, 1982; Coleman & Glofka, 1969; Gerler, 1980; Moleski & Tosi, 1976; White & Allen, 1971); (b) administered a psychotropic medication in addition to therapy in the treatment condition because the condition would not yield a pure test of therapy (Maldonado, 1984); or (c) used a comparison condition that might be considered another psychotherapy (Ullrich de Muynck & Ullrich, 1980a, 1980b). We also excluded 5 Stage 1 studies for which Shadish et al. had obtained data from M. L. Smith, Glass, and Miller (1980) but for which references were unavailable so the studies could not be obtained and recorded from scratch.

Sample B contained 40[1] studies randomly sampled from each of the same meta-analyses from which Sample A was drawn (excluding the Sample A studies, of course). The difference between the two samples is simple: Sample A was drawn from those studies that had been nominated as clinically representative by the coauthors of Shadish et al. (1997), whereas Sample B was drawn randomly. As with Sample A, we required that for each Sample B study, (a) no psychotropic medications had been administered along with psychotherapy and (b) the study compared psychotherapy to a control that was not another psychotherapy. Sample B had three functions: (a) It increased the range of clinical representativeness in the present study by expanding the number of less representative studies, (b) it allowed a test of the degree to which the Shadish et al. Stage 1 studies were more clinically representative than other studies, and (c) it assessed the possibility that the 15 meta-analyses from which Shadish et al. drew their Stage 1 studies may have included even more clinically representative studies.

Finally, Sample C included all 9 studies of what Weisz et al. (1995) called *clinic therapy:* clinic-referred children and adolescents getting psychotherapy that is already being conducted as part of the regular service-related program of a service-oriented clinic by practicing clinicians. Clinic therapy is more restrictive than the concept of therapy conducted under clinically representative conditions, and these clinic therapy studies were

[1] Sample B contained 40 studies (rather than the 41 in Sample A) because no study meeting these criteria could be located from Trull, Nietzel, and Main (1988) except for the study already in Sample A.

coded as more rather than less clinically representative according to our criteria. Including them further increased the variability of the present sample on clinical representativeness by expanding the number of more representative studies.

Very few studies in our sample were published in the 1990s, and older studies are disproportionately represented because Shadish et al. (1997) drew their sample from previously published meta-analyses, which in turn drew from previously published psychotherapy outcome studies. Data we present later suggests that more recent studies have nonsignificantly smaller effect sizes and are significantly less clinically representative, so it is unlikely that underrepresentation of more recent studies would mask a negative relationship between clinical representativeness and effect size.

Coding

Codes were both developed for the present study and taken from previous studies (Shadish et al., 1997; Shadish & Ragsdale, 1996). The Appendix presents details of codes, reliabilities, and examples. Effect sizes were computed independently of other coding to avoid bias that might result from knowing outcomes. Except for an interrater reliability study conducted before any other codings (described shortly), one of the four authors was responsible for coding each study. Some problem codings were identified when that coder was unsure of a code, and other problem codings were identified during a systematic review of key clinical representativeness codes by Glenn Phillips; in these cases, disagreements were resolved by William R. Shadish.

Clinical representativeness criteria. We used 10 criteria for clinical representativeness selected partly on the basis of their use in past research (e.g., Weisz et al., 1992), partly for their consistency with empirical literature about clinical practice (e.g., Norcross & Prochaska, 1982; Perlman, 1985; Prochaska & Norcross, 1983), and partly for their face validity: (a) *clinically representative problems:* mental health or behavioral problems that therapists see, (b) *clinically representative setting:* a setting where clinical services are commonly provided, (c) *clinically representative referrals:* clients initially referred through usual clinical routes, (d) *clinically representative therapists:* practicing clinicians for whom provision of service is a substantial part of the job, (e) *clinically representative structure:* treatment either with a structure used in clinical practice or not structured in a detailed and uniform way, (f) *clinically representative monitoring:* the implementation of treatment was not monitored in a way that could influence therapist behavior, (g) *clinically representative problem heterogeneity:* therapists treated clients (both in and outside the study)[2] who were heterogeneous in presenting problems, (h) *pretherapy training:* therapists did not receive special training immediately before the study in the specific techniques to be used, (i) *therapy freedom:* therapists used multiple techniques in therapy (both in and outside the study), and (j) *flexible number of sessions:* the study did not set limits on number of therapy sessions. A principle-components analysis of these 10 items using a scree test suggested extracting a single factor with coefficient alpha internal consistency reliability of $\alpha = .75$, so we summed them for a total score.

None of these criteria are above reproach. For example, some therapists limit their practice to particular kinds of problems; Shadish et al. (1997) discussed other examples of problems with these criteria. Hence, an underlying rule for all these codings was to judge whether the study feature seemed clinically representative. For example, one study tested the effects of bibliotherapy, so it made no sense to code therapist freedom when there were no therapists. In such cases, we coded the feature as unrepresentative.

Other treatment characteristics. We coded five *treatment orientations;* however, over 50% were behavioral, and other categories had comparatively few cases, so we collapsed the codes into behavioral–nonbehavioral. We coded *number* and *duration* of sessions by using any measure of central tendency that study authors reported or the midpoint when they reported a range. The product of number and duration of sessions was *dose* of therapy. We coded whether therapy was *brief therapy* (fewer than 10 sessions) and whether therapy used a formal structure such as a manual or videotapes.

Dependent variable characteristics. M. L. Smith et al. (1980) showed that some measures produce larger response to treatment than other measures; they called this reactivity. We used four codes for reactivity: (a) *outcome state:* more reactive for affect or cognition, less reactive for behavior or achievement tests; (b) *outcome mode:* more reactive if self-report, less reactive otherwise; (c) *manipulability:* more manipulable variables (e.g., self-report) are more reactive than less manipulable variables (e.g., achievement tests); and (d) M. L. Smith et al.'s five-point *reactivity scale* (see Appendix). After conversion to a common 0–1 scale and factor and item analyses, these items were summed to a total reactivity scale with coefficient alpha reliability of $\alpha = .75$. We intended to add a fifth item to that scale, *specificity* (how closely tied a measure is to what was done in treatment), but psychometric analyses suggested it did not belong with the reactivity items, so specificity was kept separate in subsequent analyses. We also coded the *number of weeks* posttreatment that the dependent variable was taken.

Methodology codes. We coded several useful predictors of effect size in randomized and nonrandomized experiments (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996). *Assignment to condition* was randomized or not. We coded whether the study used *matching, blocking or stratifying.* Control group *activity level* was coded active or passive, and control *similarity* was either internal or external. *Selection process* was either self-selection of participants into conditions or other-selection.

Miscellaneous codes. We coded *year of publication,* as well as *publication status* as published (code = 0) or not (code = 1). We coded whether the study was about *child–adolescent or adult presenting problems.* We coded *number of participants assigned to conditions* at the start of the study and *number remaining when outcome was measured;* from these, we computed *total attrition* (percentage of participants not measured at posttest) and *differential attrition* (treatment attrition percentage minus control attrition percentage).

Interrater Reliability

To assess interrater reliability, William R. Shadish and Georg E. Matt independently coded all of the above codes on $N = 17$ treatment-control comparisons from 11 studies (using only one outcome measure per study). Initial reliability was unacceptably low only for clinically representative structure; this item was rewritten, recoded on all remaining English language studies, and its reliability recomputed. All resulting reliabilities are reported in the Appendix as percentage agreement and kappa for categorical variables (when categories were collapsed during analysis, we used weighted kappa, weighting disagreements among collapsed categories as if they were full agreements) and intraclass correlation (r_i) for continuous variables (Tinsley & Weiss, 1975) with $N = 17$. Percentage agreement ranged from 53% to 100%, with a median of 88%; kappa ranged from .393 to 1.00, with a median of .61; and r_i ranged from .928 to 1.00, with a median of .99. Intraclass correlations were $r_i = .873$ for the clinical representativeness sale and $r_i = .889$ for the reactivity scale. Items in the Appendix with low reliability are prone to be nonsignificant in subsequent analyses, though inspection of the pattern of reliability against the pattern of nonsignificance does not reveal any obvious match.

[2] For this code, and for the therapist freedom code, we also coded an "in-study-only" version that was clinically representative if problems were heterogeneous in the study sample and if the therapist was free to use whatever interventions he or she wished within the study, respectively. Analyses using these codes revealed virtually no difference to results.

Effect Size Calculation

The effect size measure used in this meta-analysis is the standardized mean difference statistic. When possible, that statistic was computed directly as

$$d = \frac{\bar{X}_T - \bar{X}_C}{s_p},$$

where $\bar{X}_T$ is the mean of the treatment group, $\bar{X}_C$ is the mean of the comparison group, and s_p is the pooled standard deviation. Where the latter statistics were not reported, d was estimated using methods described in Shadish, Robinson, and Lu (1999). Results reported only as nonsignificant (17% of the effect sizes) were coded as 0.00, consistent with common practice (e.g., Abramowitz, 1997; T. J. Meyer & Mark, 1995). Hedges and Olkin's (1985) correction for small sample bias was applied to all effect sizes.

We coded 1,324 effect sizes from the 90 studies, with a range of 1 to 168 effect sizes and a mean of 14.71 per study. When outcomes were reported at multiple time points (e.g., posttest and follow-up), we coded the effect size closest to the end of therapy. For 89% of effect sizes, this occurred within 4 weeks of therapy, and in 94%, it occurred within 8 weeks. For the remaining 6% of effect sizes from 9 studies, the first assessment was at 13, 24, 27, 52, 55, 104, 260, 281, or 546 weeks after treatment (some of these figures are midpoints in a range). Nearly all the studies with longer term assessments were from Sample C (Weisz et al., 1995). We also coded 632 pretest standardized mean difference statistics from 53 studies, a mean of 11.92 per study. Pretest d predicts posttest d (Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996) and is useful in understanding pretreatment selection bias in nonrandomized experiments.

Analyses

We aggregated effect sizes to the study level and used both fixed and random effects analyses (Bryk, Raudenbush, & Congdon, 1996; Hedges & Olkin, 1985). Fixed and random effects analyses yield similar results if tests for homogeneity of effect sizes are not rejected; in that case, inference from both models is similar. When homogeneity is rejected, as in the present data, the two models support different inferences. Fixed effects analyses assist inferences about what this particular set of 90 studies says about clinically representative therapy, taking into account uncertainty due only to the particular samples of participants used in each study. They ask, how certain are inferences if these 90 studies had been rerun identically except for using different participants from the same population? Fixed effects models usually yield smaller confidence intervals and more powerful statistical tests but at the cost of restricted generalizability: "inferences apply to *this* collection of studies and say nothing about other studies that may be done later, could have been done earlier, or may have already been done but are not included among the observed studies" (Hedges & Vevea, 1998, p. 487). Random effects analyses assist more general inferences about effects of clinically representative therapy in a universe of studies that differs from these 90 studies in more ways than just sampling of participants. They ask, what might have happened if we analyzed data from 90 new studies from the same population as the original 90 studies, where the new studies not only used different participants but differed in other respects such as using different designs, dose, outcome measures, or levels of clinical representativeness? Increased generalizability is achieved by incorporating an estimate of between-studies variability into error variance estimates and statistical tests. The cost is broader confidence intervals, less powerful statistical tests, and ambiguity about how to define the universe of studies concretely. (Further discussion of strengths and weaknesses of these models can be found in the following readings: Cooper & Hedges, 1994; Hedges & Vevea, 1998; National Research Council, 1992; Overton, 1998; Shadish & Haddock, 1994.)

Results

Some Descriptive Results

Study-level effect sizes ranged from $d = -1.01$ to $d = 2.77$. Scores on the clinical representativeness scale ranged from 1 to 10 with a mean of 6.76. Figure 1 is a scatterplot of the relationship between effect size and clinical representativeness scores. The univariate correlation between these variables was significantly negative (random effects univariate $r = -.29$, $p = .0031$; fixed effects univariate $r = -.35$, $p < .0001$). Below, we present evidence that this correlation was an artifact of confounds with other study features, and it disappears when those confounds are taken into account. The addition of nonlinear (quadratic, cubic, and quartic) transformations of clinical representativeness scores did not significantly improve prediction of effect size. Four studies with effect sizes greater than $d = 1.82$ and one study with an effect size less than $d = -.95$ were statistical outliers using a test by Hoaglin, Mosteller, and Tukey (1983). However, Winsorizing those effect sizes by reducing them to either 1.82 for the positive effect sizes or to $-.95$ for the negative effect size had no substantial effect on the correlation between effect size and clinical representativeness (no subsequent analyses used Winsorization).

In Figure 1, studies from Sample A are represented by squares, Sample B studies by triangles, and Sample C studies by diamonds; randomized studies are represented by solid squares, triangles, or diamonds, and nonrandomized studies by squares, triangles, or diamonds without fill. Inspection of the figure with these distinctions in mind leads to the following qualitative observations. First, randomized studies tend to report larger effect sizes than nonrandomized studies; later, we show that this is due to selection bias in nonrandomized studies.

Second, some Sample B studies scored high on clinical representativeness, suggesting that additional clinically representative studies may exist in the literature from which Sample B was randomly drawn. Such studies probably do not exist in large numbers, but a directed search for them might be warranted given the importance of the question.

Third, clinical representativeness scores were significantly different across the three samples, $F(2, 87) = 33.04$, $p < .001$, with follow-up tests indicating all pairwise differences were significant. Sample C obtained significantly higher clinical representativeness scores ($M = 9.94$) than did Sample A ($M = 7.57$), which in turn was significantly higher than Sample B ($M = 5.21$). These results are expected because Sample C used a definition of clinic therapy that is more restrictive on clinical representativeness than the other samples, and Sample A studies were nominated by the coauthors of Shadish et al. (1997) as being more clinically representative than other studies in the meta-analyses from which they were drawn, the latter in Sample B.

Fourth, the 90 studies range from very clinically unrepresentative to highly representative. For example, Akins, Hollandsworth, and O'Connell (1982) treated introductory psychology and sociology students solicited for dental fear with a 1-hr, researcher-administered intervention given by audio and videotape in a college laboratory. Figure 1 suggests that the meta-analysis literature does contain such studies with questionable clinical representativeness. Other studies are reasonably representative. For example, Lipsky, Kassinove, and Miller (1980) randomly assigned clients who had a diagnosis of neurosis or adjustment disorder of adult-

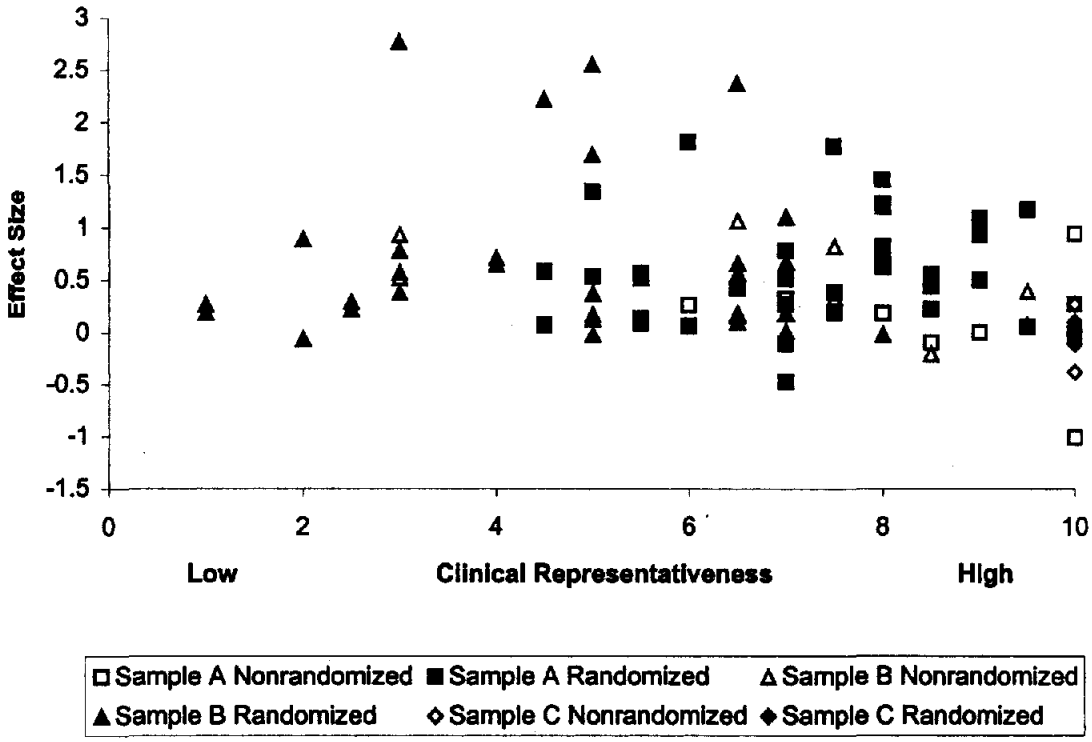

Figure 1. A scatterplot of the relationship between effect size and clinical representativeness scores.

hood and who applied for psychotherapy at an outpatient community mental health center to a control group, to various kinds of rational–emotive therapy (RET), or to a relaxation training and supportive therapy group. Therapy lasted 45 min a week for 12 weeks and was delivered by experienced mental health professionals who had worked in a variety of settings and with a variety of clients and who were not specially trained in RET for the study.

Fifth, nonrandomized studies have significantly higher clinical representativeness scores ($M = 8.15$) than randomized experiments ($M = 6.21$; $t(89) = 3.65$, $p < .001$). It is more difficult to use random assignment under clinically representative conditions, but the presence of substantial numbers of clinically representative randomized studies in our samples suggests it is possible to design externally valid randomized studies, much as other fields such as public health, medicine, and pharmaceuticals have sometimes done.

Sixth, a significant negative correlation ($r = -.269$, $p = .01$) exists between clinical representativeness scores and year of publication (see Figure 2). Early years saw a few studies that were highly clinically representative and nonrandomized. The mid-1960s saw an explosion of outcome studies with a range of clinical representativeness, studies that were increasingly dominated by randomized experiments—although nonrandomized experiments after the 1960s also increased their range of clinical representativeness. This increase in range of clinical representativeness may reflect greater interest in basic than applied research or changing publication standards.

Table 1 presents means and standard deviations for the clinical representativeness items. Item means range from 0–1; high numbers imply higher clinical representativeness. Few studies used flexible number of sessions; half the studies used clinically representative referrals and therapists; about two-thirds used clinically representative settings, structure, monitoring, problem heterogeneity, pretherapy training, and therapist freedom. Nearly all (93%) studies used patients with clinically representative problems.

Primary Analyses

The primary analysis with which to answer the main question of this research is a multiple regression predicting effect size from coded variables. A baseline for subsequent analysis is an intercept-only model that yields an unadjusted overall effect size. The random effects weighted average effect size over all 90 studies was $d = .412$ ($SE = .057$, $p < .001$); model fit statistics yielded a variance component $\tau = .158$, with a heterogeneity $\chi^2(89) = 237.92$, $p < .001$, suggesting that systematic effect size variability remained to be accounted for. The fixed effects analysis yielded $d = .30$ ($SE = .032$, $p < .0001$), with a heterogeneity $\chi^2(89) = 240.91$, $p < .0001$. As the similarity between χ^2 statistics suggests, model fit statistics for fixed and random effects models are very similar, differing slightly because they are computed using different estimation algorithms.[3]

[3] Mean effect sizes for studies reported here are considerably lower than those presented in the early large-scale meta-analyses (e.g., M. L. Smith et al., 1980). This is even true for our Sample B from past meta-analyses. This difference should not be interpreted as indicating that our clinically representative studies show smaller treatment effects or that our Sample B is somehow unrepresentative. Matt (1989; Matt & Navarro, 1997) has pro-

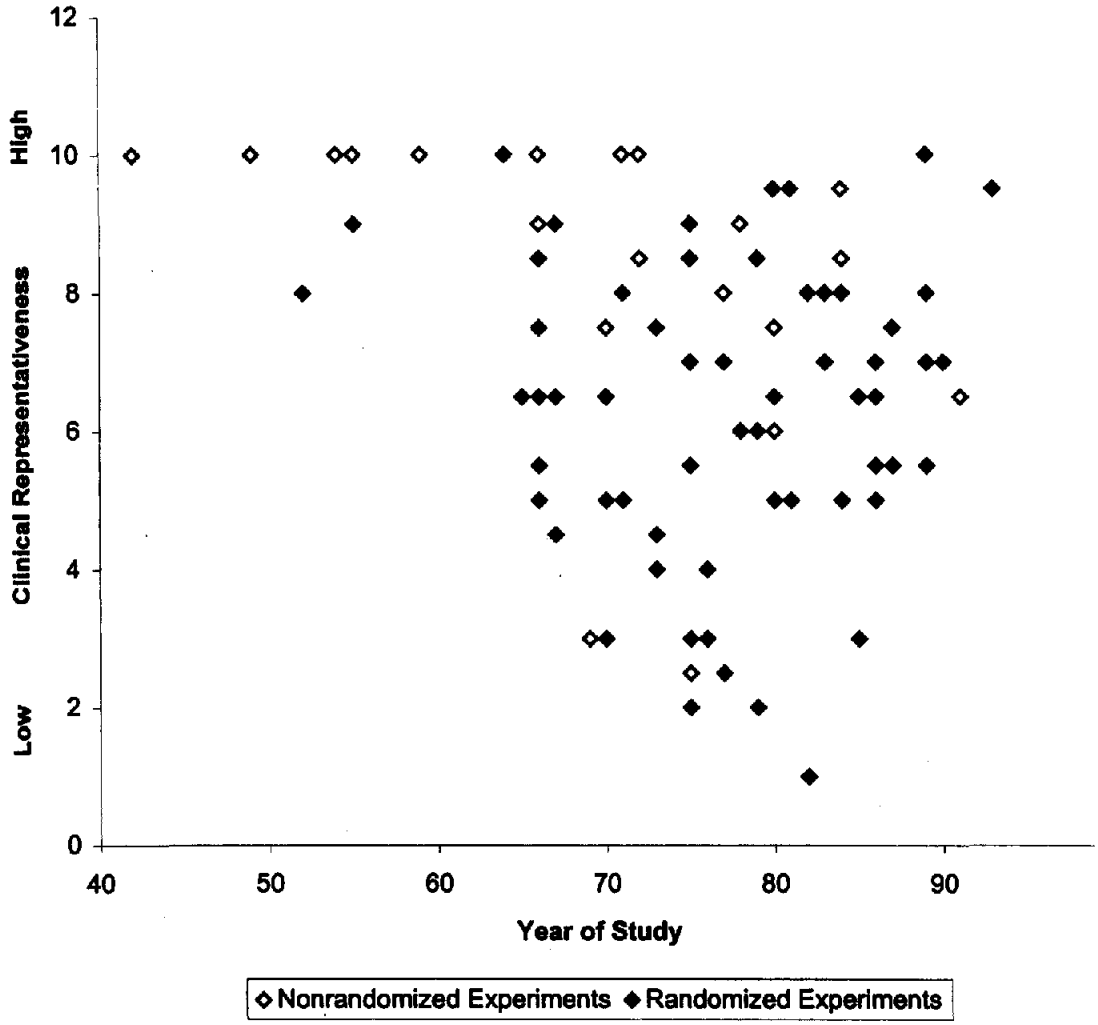

Figure 2. A scatterplot of the relationship between year of publication and clinical representativeness scores.

We then entered all predictors with two variations. First, we used the clinical representativeness scale (CRS) total score rather than items (Table 2). Both fixed and random effects analyses suggested that effect sizes were larger (a) the greater the dose of therapy, (b) when highly specific measures were used to assess outcome, and (c) for behaviorally oriented therapies. Fixed effects analysis also suggested effects were larger (d) when an internal control group was used and (e) when outcome was measured near the end of therapy. Clinical representativeness was unrelated to effect size in both analyses. Model fit statistics for the random effects analysis were $\tau = .102$, $\chi^2(71) = 144.68$, $p < .001$, with all predictors accounting for 36% of the parameter variation in study effect size (Bryk & Raudenbush, 1992, p. 74). In the fixed effects analysis, multiple $R = .63$, again with nearly identical fit statistics, $\chi^2(71) = 145.18$, $p < .0001$; the presence of significant heterogeneity suggests caution in generalizing results of the fixed effects model beyond these 90 studies, and suggests placing greater confidence in the two predictors identified in the random effects model. Though we do not repeat this caution in subsequent analyses, it is implicit because effect sizes were heterogeneous in all of them.

Second, we used the CRS items (Table 3). With all predictors in the random effects equation, results were similar to the analysis with the total score. Effect sizes were larger (a) the greater the dose of therapy and (b) when highly specific measures were used. Model fit statistics were $\tau = .087$, with $\chi^2(62) = 118.21$, $p < .001$, accounting for 45% of parameter variation in study effect size, a bit more than when the total score was used. With all predictors in the fixed effects equation, multiple $R = .72$, with nearly identical fit statistics, $\chi^2(62) = 116.34$, $p < .0001$; and effect sizes were larger (a) the greater the dose of therapy, (b) when highly specific measures were used, (c) when internal control groups were used, (d) with more representative clinical structure, (e) when outcome

vided evidence that estimates of early large-scale meta-analyses most likely overestimated treatment effects, and similarly, Shadish et al. (1997) showed that reanalyzing M. L. Smith et al.'s (1980) data with weighted least squares models also led to smaller effects. Overall therapy mean effect estimates are more realistically in the .40–.60 range, rather than in the .70–.80 range as reported in early meta-analyses.

Table 1
Clinical Representativeness Item and Total Score Statistics

Item	*M*	*SD*
Clinically representative problems	.93	.22
Clinically representative setting	.72	.38
Clinically representative referrals	.59	.49
Clinically representative therapists	.57	.45
Clinically representative structure	.76	.43
Clinically representative monitoring	.77	.42
Clinically representative problem heterogeneity	.67	.46
Clinically representative pretherapy training	.68	.46
Clinically representative therapy freedom	.79	.40
Clinically representative flexible number of sessions	.28	.45
Total score	6.76	2.36

Note. In all cases, a larger mean implies greater clinical representativeness, with a range from 0 to 1. $N = 90$ studies.

was measured near the end of therapy, (f) for participants without clinically representative mental health problems, and (g) when therapy was not limited to a fixed number of sessions. Three of these predictors involved clinical representativeness (d, f, g), with more clinical representativeness increasing effect size for two predictors (d, g) and decreasing it for one predictor (f).

Selection Bias in Nonrandomized Experiments

Psychotherapy researchers have mixed opinions about the role of nonrandomized experiments in estimating the effects of therapy

Table 2
Regression Coefficients and Standard Errors Predicting Effect Size Using the Clinical Representativeness Total Score

	Random effects model		Fixed effects model	
Variable	Regression coefficient	*SE*	Regression coefficient	*SE*
Intercept	0.0553	.657	0.1458	.423
Clinical representativeness	−0.0002	.033	−0.0104	.025
Year of publication	−0.0064	.007	−0.0056	.004
Therapy dose in minutes[a]	0.0001	.000*	0.0001	.000*
Total attrition	−0.1608	.365	−0.0229	.199
Differential attrition	0.1292	.525	0.0359	.343
Reactivity scale	−0.0036	.058	0.0062	.039
Outcome specificity	0.5002	.212*	0.4841	.140*
Matching	0.0103	.126	−0.0165	.086
Internal control group	0.3138	.246	0.3467	.142*
Passive control group	0.1935	.131	0.1630	.089
Self-selection	−0.0117	.250	−0.0677	.146
Did not use structure	−0.0837	.173	−0.0388	.116
Random assignment	0.0960	.160	0.0311	.104
Unpublished work	0.0203	.182	0.0095	.135
Adult presenting problem	−0.0577	.119	−0.1152	.082
Not brief therapy	0.0702	.131	0.0840	.090
Behavioral orientation	0.3360	.165*	0.3030	.110*
Weeks to posttest	−0.0010	.001	−0.0010	.001*

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor.
[a] This variable was measured as total minutes of therapy, so its impact can be substantial despite the small size of the regression coefficient.
* $p < .05$.

Table 3
Regression Coefficients and Standard Errors Predicting Effect Size Using the Clinical Representativeness Items

	Random effects model		Fixed effects model	
Variable	Regression coefficient	*SE*	Regression coefficient	*SE*
Intercept	−0.3084	.803	−0.0045	.561
Year of publication	0.0008	.007	−0.0004	.004
Therapy dose in minutes[a]	0.0001	.000*	0.0001	.000*
Total attrition	0.1174	.377	0.1690	.217
Differential attrition	0.3838	.523	0.2824	.359
Reactivity scale	−0.0311	.062	−0.0349	.045
Outcome specificity	0.6316	.216*	0.6430	.152*
Matching	0.0356	.123	0.0232	.088
Internal control group	0.3273	.255	0.4074	.158*
Passive control group	0.1892	.133	0.1777	.096
Self-selection	−0.0529	.257	−0.0476	.160
Did not use structure	−0.1625	.189	−0.1678	.135
Random assignment	−0.0387	.163	−0.0933	.111
Unpublished work	−0.1594	.191	−0.2034	.147
Adult presenting problem	−0.0693	.128	−0.1352	.092
Not brief therapy	0.1199	.137	0.1540	.097
Behavioral orientation	0.2666	.172	0.2354	.123
Weeks to posttest	−0.0007	.001	−0.0009	.001
CR problems	−0.4336	.280	−0.5262	.211*
CR setting	0.0267	.196	−0.0536	.145
CR referrals	0.1285	.162	0.1272	.121
CR therapists	−0.1964	.233	−0.1706	.176
CR structure	0.2901	.159	0.3776	.117*
CR monitoring	0.1514	.154	0.0548	.112
CR problem heterogeneity	−0.0828	.230	−0.0245	.170
CR pretherapy training	−0.1918	.163	−0.1838	.122
CR therapy freedom	0.3567	.223	0.3053	.165
CR flexible number of sessions	−0.1568	.140	−0.1889	.098*

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor. CR = clinically representative.
[a] This variable was measured as total minutes of therapy, so its impact can be substantial despite the small size of the regression coefficient.
* $p < .05$.

(Beutler, 1998). The issue is particularly sharply drawn because studies of therapy under clinically representative conditions are more likely to be nonrandomized, so omitting nonrandomized studies excludes some of the most clinically representative studies. However, effect size estimates from nonrandomized studies may be biased, so such studies may bias overall effect size estimates (Matt & Cook, 1994). To shed light on this conundrum, we compared results from randomized to nonrandomized designs, with particular attention to the biasing effect of selection bias in nonrandomized experiments. One study (Spiegler et al., 1976) used both a randomized and a nonrandomized control, which we categorized separately for this analysis, increasing N from 90 to 91.

First, effects in nonrandomized studies differ from effects in randomized studies. Most studies ($n = 67$, 74%) randomized clients to treatment or control conditions. Randomized studies had an average random effects $d = .52$ ($SE = .067$), significantly higher than the average $d = .20$ ($SE = .095$) from 24 nonrandomized studies ($Q = 7.37$, $df = 1$, $p = .007$). Similarly in the fixed

effects analyses, randomized experiments had an average $d = .43$ ($SE = .042$), significantly higher than the average $d = .14$ ($SE = .047$) from 24 nonrandomized studies ($Q = 21.05$, $df = 1$, $p < .0001$). Second, evidence suggests that this outcome difference between methodologies is due to a self-selection bias that decreases effect sizes from nonrandomized experiments, making therapy look misleadingly ineffective. Specifically, for 12 nonrandomized experiments where clients self-selected into treatment, posttreatment effect size was random effects $d = -.03$ (fixed effects $d = -.02$), but for 12 nonrandomized experiments where clients were put into conditions by another nonrandom mechanism, posttreatment effect size was random effects $d = .46$ (fixed effects $d = .44$); the latter is close to the effect size for randomized experiments. Examples of these other nonrandom mechanisms included clients placed on a no-treatment waiting list because therapists were unavailable (Barron & Leary, 1955), clients assigned to treatment or control in alternating weeks (Coche & Flick, 1975), alternating assignment of clients to treatment or control until treatment slots were filled with all subsequent applicants placed in control (Endicott & Endicott, 1964), controls selected by matching to treatment clients from the same incarcerated population with both groups selected by social workers at the institution (Persons & Pepinsky, 1966), and controls chosen from hospital records of admissions to the same inpatient unit on the same day as treated patients who were identified by treatment staff (Sacks & Berger, 1954). These other mechanisms seem to have two features in common: (a) None involved client self-selection into conditions, and (b) in all cases, both the treated and control samples plausibly came from the same population.

Third, examining pretest standardized mean difference statistics in 53 studies where sufficient pretest data were available reinforced the possibility of bias in nonrandomized experiments. Statistical theory suggests that pretest group differences should be zero and homogenous in the 42 randomized studies, and they are: Their average pretest standardized mean difference statistic was a nonsignificant random effects $d = -.05$ ($SE = .05$, $p = .35$; fixed effects $d = -.05$, $SE = .05$, $p = .35$) and homogenous, $\chi^2(41) = 25.76$, $p = .97$. If nonrandomized experiments are unbiased, their pretest standardized mean difference statistics should be similar, but they are not. For 11 nonrandomized studies, the average pretest standardized mean difference statistic was smaller than zero at random effects $d = -.25$ ($SE = .145$, $p = .09$; fixed effects $d = -.31$ ($SE = .08$, $p = .0001$) and not homogenous, $\chi^2(10) = 33.31$, $p = .0002$. In fact, the difference in standardized mean difference statistics between nonrandomized and randomized experiments at pretest is nearly equal to the difference between them at posttest. Self-selection bias is indicated by the fact that pretest standardized mean difference statistics in 6 nonrandomized studies where clients self-selected into treatment (random effects $d = -.47$, $SE = .16$, $p = .02$) are significantly lower ($Q = 3.99$, $df = 1$, $p = .046$) than in 5 studies where participants were put into conditions by another nonrandom mechanism ($d = .04$, $SE = .19$, $p = .85$). This finding held in a fixed effects analysis where effect sizes for studies where clients self-selected into treatment ($d = -.50$, $SE = .10$, $p = .0007$) are significantly lower ($Q = 10.24$, $df = 1$, $p = .0014$) than in 5 studies where clients were put into conditions by another nonrandom mechanism ($d = .02$, $SE = .13$, $p = .89$).

The nature of this self-selection bias is probably straightforward. People select therapy when they are most distressed, leaving less distressed people to be control group members. On measures related to distress, self-selection results in treatment group clients scoring worse than control group clients at pretest. Even if therapy is effective, it raises treatment group client posttest status to the level of control group clients, resulting in a zero effect size at posttest. Ragsdale (1996) provided additional empirical evidence in support of this interpretation in a sample of marital and family therapy studies.

Would previous regression results (Tables 2 and 3) hold if analyses were limited to randomized experiments? The univariate correlation of effect size with clinical representativeness is smaller and nonsignificant in randomized experiments (random effects $r = -.11$, $p = .33$; fixed effects $r = -.13$, $p = .11$); indeed, eliminating only nonrandomized experiments where clients self-selected into treatment, the univariate correlation between effect size and clinical representativeness in the remaining 79 studies is nonsignificant (random effects $r = -.10$, $p = .32$; fixed effects $r = -.11$, $p = .19$). Running a regression on only randomized experiments yielded generally similar results (Tables 4 and 5; predictors for random assignment, internal control, and self-selection were omitted from this analysis because they are constants in randomized experiments): Effect sizes were larger (a) the greater the dose of therapy, (b) when highly specific measures were used, (c) in studies published longer ago, and (d) when passive controls were used. Model fit statistics for the random effects model were $\tau = .065$, $\chi^2(51) = 83.34$, $p = .003$, accounting for 45% of parameter variation in study effect size. For the fixed effects model, multiple $R = .62$ with similar fit statistics,

Table 4
Regression Results on Randomized Experiments Only Using the Clinical Representativeness Total Score

	Random effects model		Fixed effects model	
Variable	Regression coefficient	*SE*	Regression coefficient	*SE*
Intercept	1.2150	.678	1.3882	.534
Clinical representativeness	0.0561	.035	0.0541	.029
Year of publication	−0.0220	.008*	−0.0238	.007*
Therapy dose in minutes[a]	0.0001	.000*	0.0002	.000*
Total attrition	−0.4515	.486	−0.2688	.377
Differential attrition	−0.6536	.578	−0.9130	.459*
Reactivity scale	−0.0118	.067	−0.0008	.054
Outcome specificity	0.7425	.250*	0.7566	.196*
Matching	−0.0606	.133	−0.0771	.103
Passive control group	0.3293	.150*	0.2989	.118*
Did not use structure	−0.2921	.185	−0.3222	.145
Unpublished work	0.1237	.175	0.1042	.140
Adult presenting problem	−0.0180	.126	−0.0115	.099
Not brief therapy	−0.0276	.152	−0.0396	.120
Behavioral orientation	0.2622	.182	0.1934	.144
Weeks to posttest	0.0001	.019	−0.0039	.015

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor.
[a] This variable was measured as total minutes of therapy, so its impact can be substantial despite the small size of the regression coefficient.
* $p < .05$.

Table 5
Regression Results on Randomized Experiments Only Using the Clinical Representativeness Items

Variable	Random effects model		Fixed effects model	
	Regression coefficient	*SE*	Regression coefficient	*SE*
Intercept	−0.1047	.924	0.0195	.729
Year of publication	−0.0120	.010	−0.0132	.008
Therapy dose in minutes[a]	0.0002	.000	0.0001	.000*
Total attrition	−0.2213	.618	−0.0842	.457
Differential attrition	−0.5587	.654	−0.7763	.502
Reactivity scale	−0.0196	.078	−0.0006	.061
Outcome specificity	0.9064	.293*	0.8907	.222*
Matching	−0.0177	.149	−0.0392	.111
Passive control group	0.3204	.171	0.3112	.133*
Did not use structure	−0.1776	.229	−0.1613	.179
Unpublished work	0.0145	.223	−0.0527	.170
Adult presenting problem	0.0932	.150	0.0960	.114
Not brief therapy	−0.0477	.181	−0.0641	.136
Behavioral orientation	0.2229	.211	0.1723	.165
Weeks to posttest	−0.0098	.023	−0.0154	.017
CR problems	0.0529	.350	0.0697	.274
CR setting	0.4085	.227	0.3920	.176*
CR referrals	−0.0052	.191	0.0159	.149
CR therapists	−0.0594	.261	−0.0380	.202
CR structure	0.1552	.189	0.1893	.147
CR monitoring	0.2591	.182	0.2024	.139
CR problem heterogeneity	−0.0301	.255	0.0199	.197
CR pretherapy training	−0.2532	.191	−0.2854	.152
CR therapy freedom	0.3093	.244	0.2556	.192
CR flexible number of sessions	−0.0406	.196	−0.0755	.146

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor. CR = clinically representative.
[a] This variable was measured as total minutes of therapy, so its impact can be substantial despite the small size of the regression coefficient.
* $p < .05$.

$\chi^2(51) = 84.92$, $p = .002$. Significant fixed effects predictors included the random effects predictors, but effect size was also greater when the percentage of dropouts from the control group was greater than the percentage of dropouts from the treatment group. This might occur if dropouts are less distressed, leaving those clients remaining in the control group more distressed than those remaining in treatment.

Using the clinical representativeness items (Table 5), in the random effects regression, $\chi^2(42) = 72.75$, $p = .0085$, only outcome specificity significantly predicted effect size. For the fixed effects regression, multiple $R = .69$, with similar fit statistics, $\chi^2(42) = 71.63$, $p = .0029$; and effect sizes were higher (a) the greater the dose of therapy, (b) when highly specific measures were used, (c), when passive controls were used, and (d) when therapy occurred in more clinically representative settings.

Generalizing by Extrapolation to the Most Clinically Representative Study

Rubin (1992) proposed reconceptualizing meta-analysis as modeling a multivariate response surface to extrapolate from known data to a target ideal study. Unfortunately, there are no published exemplars of how to do this modeling in meta-analysis, and the requisite techniques (see, e.g., Box & Draper, 1987) have not been adapted to meta-analysis. However, we can comply with the spirit of this proposal by using regression results to extrapolate from the data to an effect size for a hypothetical ideal study of the effects of clinically representative therapy. To do so, we multiply the random effects regression coefficients by codes that represent the ideal study and then add the products to obtain a predicted effect size. Random effects coefficients are used because the inference generalizes from these 90 studies to studies with different clinical and methodological characteristics. We use the full sample coefficients (Table 2 and 3) to take advantage of the maximum amount of information from these 90 studies.

An ideal study of the effects of clinically representative therapy would (a) score at the maximum on the clinical representativeness items and (b) use the most accurate methodology for estimating treatment effects: It would match and then randomly assign (so the control group necessarily would be internal and other-selected) and have no attrition, for statistical theory suggests that these characteristics are most likely to yield an unbiased estimate of effect size. The remaining variables do not have ideal levels, so the task is to choose levels that yield inferences of interest. The inference of most direct interest here concerns what these 90 studies would have found if they were fully clinically representative and used the most accurate methods. This inference is obtained by using the average value on the remaining codes: publication date ($M = 1{,}974.91$), dose in minutes ($M = 986.71$), proportion using behavioral therapy ($M = .40$), reactivity scale score ($M = 2.55$), proportion using a specific outcome measure ($M = .55$), weeks after therapy ended that outcome was measured ($M = 15.74$), proportion using passive control group (i.e., either no treatment or wait-list; $M = .64$), proportion not using structure in therapy ($M = .51$), proportion not using brief therapy ($M = .56$), proportion unpublished ($M = .12$), and treating adult (code = 2) versus child (code = 1) presenting problems ($M = 1.52$). These assumptions yield a predicted effect size of $d = .52$. Using the clinical representativeness items (and so, the coefficients in Table 3), the predicted effect size is $d = .41$. For comparison, the random effects weighted average effect size over all 90 studies is $d = .41$, and $d = .55$ for the 41 studies in Sample B (which was drawn randomly from 15 past meta-analyses). That is, an ideal study of the effects of clinically representative therapy would yield effects that are the same as or only slightly smaller than those reported in these past meta-analyses.

One can also project the effects of studies using other assumptions. For example, assume the study was published in 1998, used a dose of 1,250 min of therapy (25 weeks of one 50-min session per week), used nonbehavioral therapy, used measures with moderate reactivity (rating of 2) and high specificity (rating of 1), measured outcome after the last therapy session, had a passive control group (i.e., either no treatment or wait-list), did not use a formal structure in therapy, used longer rather than brief therapy, and treated adult presenting problems. This set of assumptions yields a predicted effect size for an ideal clinically representative study of $d = .54$ for the regression in Table 2. Doubling therapy dose to 50 weeks would increase the d from .54 to .68. Assuming in addition that therapy was behavioral, d would rise from to .68 to 1.02. None of these assumptions are right or wrong. Rather they

show that results from therapy outcome studies vary depending on their substantive and methodological characteristics. As Rubin (1992) suggested:

> In practice, there may be different definitions of the ideal study depending on the purpose of the meta-analyses and the predilections of the meta-analysts. Such differences are natural in all types of scientific endeavors, and the requirement to have to formalize the intended purpose . . . is a positive feature of this perspective. (p. 368)

Discussion

In general, this research supports the effectiveness of psychological therapies that are conducted under clinically representative conditions, and it even suggests that the more such therapy is provided, the better the outcomes. However, accurate interpretation of these results requires clarification of the relationship between clinical representativeness and related constructs (clinical therapy and empirically supported psychological therapies), of the interpretation of nonrandomized experiments, of likely limitations on the dose–effect relationship that we found, and of some more general limitations to the meta-analytic methodology that we used. Finally, we highlight some innovative meta-analytic techniques used in this study that might be more widely applicable in other meta-analyses.

Clinical Representativeness and Related Constructs

What is the relationship between clinically representative therapy and the related constructs of clinic therapy (see, e.g., Weisz et al., 1992, 1995) or empirically supported psychological therapies (Chambless & Hollon, 1998)? Clinic therapy focuses on clinic-referred children and adolescents getting psychotherapy that is already being conducted by practicing clinicians as part of the regular service-related program of a service-oriented clinic. The requirement that therapy already exist as part of the regular service-related program of a service-oriented clinic best differentiates clinic therapy from clinically representative therapy. This requirement excludes studies of therapies transported from the lab to see how they perform under clinically representative conditions, but the latter are of great interest to the present question. So, clinic therapy focuses on a small subset of studies of therapy under clinically representative conditions.

Second, empirical support for the effectiveness of psychological therapies under clinically representative conditions should be discriminated from empirically supported psychological therapies (Chambless & Hollon, 1998). Although our results empirically support the effectiveness of psychological therapies, the term empirically supported psychological therapies has special meaning: clearly specified psychological therapies (i.e., with a treatment manual or its equivalent) that are efficacious in controlled research with a well-delineated population (Chambless & Hollon, 1998). The latter therapies can be conducted in more or less clinically representative conditions and may improve the outcome of clinically representative therapy still further.

Randomized Versus Nonrandomized Trials

The present results suggest that self-selection into treatment in nonrandomized trials can bias effect size estimates (for similar findings in psychotherapy, education, drug abuse prevention, and medicine, see Colditz, Miller, & Mosteller, 1988; Gilbert, McPeek, & Mosteller, 1977; Heinsman & Shadish, 1996; Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996). This self-selection bias makes psychotherapy appear ineffective in nonrandomized trials, and because nonrandomized trials tend to be more clinically representative, this bias also makes it appear as though more clinically representative studies yield smaller effects. Both findings are artifacts of self-selection bias. Unfortunately, few reviews that include nonrandomized trials account for such biases by using covariates such as use of random assignment, presence of self-selection, similarity of the control group, and pretest standardized mean differences. Rather, they typically lump the two kinds of trials together without adjustment, or they have no randomized benchmarks at all. Both practices are worth discouraging. The interpretation of quasi-experiments benefits from careful and detailed efforts to take into account plausible threats to internal validity (Cook & Campbell, 1979; Shadish, Cook, & Campbell, in press). In meta-analytic contexts, the same cautions apply (Matt & Cook, 1994).

As a corollary, conventional wisdom often equates clinically representative studies with effectiveness research or with mental health services research, and the former studies are contrasted with research-oriented, efficacy, or clinical trials studies (e.g., Donenberg, Lyons, & Howard, 1999). In these discussions, the former set of studies is frequently portrayed as emphasizing external over internal validity. The present results belie such an oversimplification. In the classic conceptualization of internal validity (Campbell & Stanley, 1963; Cook & Campbell, 1979), the two crucial methodological features for high internal validity are random assignment and the minimization of attrition. Clearly, many clinically representative studies in this sample met these two criteria—they were both clinically representative and internally valid.

Therapy Dose

The association between increased therapy dose and better outcome has precedent in past research (Bovasso, Eaton, & Armenian, 1999; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Kopta, Howard, Lowry, & Beutler, 1994; Lambert & Cattani-Thompson, 1996). However, three qualifications are crucial to interpreting this result. First, most of the literature fails to manipulate dose experimentally, so dose–response effects may be confounded with other study features. This problem is probably reduced in the present study by including potential confounds in the regressions along with dose, but we can never be certain these confounds are eliminated entirely. Second, it is unlikely that increased dose improves outcome indefinitely; this is crucial given the centrality of dose to the costs of therapy. Kopta et al. (1994) found that outcomes level off over a year; the addition of a quadratic term in the present data to model this leveling-off approached significance in some regressions. Third, a positive dose–effect relationship may hold more strongly for some disorders (see, e.g., Shapiro et al., 1994) or some measures (see, e.g., Kadera, Lambert, & Andrews, 1996), which may explain why some primary studies find dose–effect relationships, but others do not.

Limitations of the Present Study

Five limits to this study suggest caution in its interpretation. First is a construct validity issue—do our codes assess key features of clinical representativeness? The overlap between our criteria and related ones (see, e.g., Weisz et al., 1992) suggests all are tapping a common construct. However, controversies still exist both about whether we have omitted key features or included irrelevant ones and about obvious exceptions to our particular criteria such as the clinician who specializes in treating one kind of problem (see Shadish et al., 1997, for other exceptions). Needed are further conceptual analysis of the convergent and discriminant validity of clinical representativeness and cognate constructs like clinic therapy, and further empirical research on key features of clinically representative practice to inform construct development.

Second, it is difficult to reduce criteria for coding clinical representativeness to a few sentences of prose. Rather, such codings involve difficult clinical and research judgments about which reasonable people may disagree. For example, how long must a description of treatment be to count as a treatment manual? How much selection through informed consent is allowed in an otherwise clinically representative referral? Detailed rules for all these possibilities would yield an impossibly long coding manual, even if the details could be justified for such inherently ambiguous matters.

Third, like most meta-analysts, we inferred codes from study reports, a difficult task because study reports are ambiguous. An alternative is to contact original authors to help with coding. However, the obstacles to that method are substantial. Some studies were published 50–60 years ago, and their authors are unavailable. Even for more recent studies where authors can be located, their memories of study procedures from 10 or more years ago may be untrustworthy, confounding coding biases with age of the study.

Fourth, important interactions can exist in psychotherapy meta-analysis (Shadish & Sweeney, 1991), yet we did not report such tests. Issues of testing and interpreting interactions in multilevel random effects meta-analytic models are unclear, both in regard to how to center multilevel data to prevent colinearity (Kreft, de Leeuw, & Aiken, 1995) and in regard to appropriate follow-up tests. This problem must be addressed if meta-analysts are to use Rubin's (1992) response surface modeling approaches to meta-analysis, where exploring interactions is essential.

Fifth, meta-analytic data are correlational, more like survey research than experimental research, so clinical representativeness is inevitably confounded with other variables. Although we tried to adjust for those confounds, we can never know if they were completely identified and validly measured—that is the nature of selection bias in correlational data. Random effects models help compensate for this uncertainty, but many more primary experiments are needed to cross-validate and extend these results.

Meta-Analytic Methodology

This study illustrates three novel meta-analytic techniques. First is the use of regression analysis to project results to an ideal study, in the spirit of Rubin's (1992) call to conceptualize meta-analysis as a task in response surface modeling. Although the implementation we present is crude compared with Rubin's aspirations, our methods are feasible using existing random effects models. A virtue of this technique is that it forces the meta-analyst to make explicit assumptions about the target of generalization.

Second is the use of pretest effect sizes to explore selection bias in nonrandomized experiments. This method helps to explore the substantive nature of selection bias (e.g., differences in self- versus other-selection) and to relate pretest to posttest biases. Elsewhere, we have used pretest effect size estimates to obtain adjusted posttest treatment effect estimates from both randomized and nonrandomized studies in meta-analyses (Shadish & Heinsman, 1997; Shadish & Ragsdale, 1996).

Third is the joint use of fixed and random effects models. Fixed effects analyses suggested that in these 90 studies, therapy dose and outcome specificity consistently predicted effect size and that other variables predicted less consistently, including the use of internal controls, passive controls, behavioral therapies, year of publication, time to outcome measurement, attrition, and a few clinical representativeness items (the latter items sometimes increased and sometimes decreased outcome). However, rejection of effect size homogeneity suggested that generalizing fixed effect findings beyond these 90 studies is ill advised. Random effects analyses yield more appropriate generalization under effect size heterogeneity, suggesting that only therapy dose and outcome specificity yield robust effects. Generalization beyond these 90 studies is the question of greatest interest, so the random effects analyses warrant more confidence.

Conclusions

The results of this study suggest four main substantive conclusions. First, psychological therapies are robustly effective across conditions that range from research-oriented to clinically representative. Second, previous findings that clinical representativeness leads to lower effect size are probably an artifact of other confounding variables, especially biased self-selection into treatment in many quasi-experiments that happen to be clinically representative. Third, increased dose of therapy is associated with better outcome, though it seems likely that benefits may level off at some point. Fourth, studies tend to show much larger effects if they assess outcome by using measures that are closely tailored to the goals that were focused on in treatment (similar to the "teaching the test" effect from educational research).

References

References marked with an asterisk indicate studies included in the meta-analysis; those studies are identified as part of Sample A, B, or C by the letter in parentheses following the reference.

Abramowitz, J. S. (1997). Effectiveness of psychological and pharmacological treatments for obsessive–compulsive disorder: A quantitative review. *Journal of Consulting and Clinical Psychology, 65,* 44–52.

*Akins, T., Hollandsworth, J. G., & O'Connell, S. J. (1982). Visual and verbal modes of information processing and their relation to the effectiveness of cognitively based anxiety-reduction techniques. *Behaviour Research and Therapy, 20,* 261–268. (B)

*Alper, T. G., & Kranzler, G. D. (1970). A comparison of the effectiveness of behavioral and client-centered approaches for the behavior problems of elementary school children. *Elementary School Guidance and Counseling, 5,* 35–43. (B)

Amerikaner, M., & Summerlin, M. L. (1982). Group counseling with

learning disabled children: Effects of social skills and relaxation training on self-concept and classroom behavior. *Journal of Learning Disabilities, 15,* 340–343.

*Anesko, K. M., & O'Leary, S. G. (1982). The effectiveness of brief parent training for the management of children's homework problems. *Child and Family Behavior Therapy, 4,* 113–126. (B)

*Antons, K. (1972). Soziometrische Kurwirkungen. Eine Test–Retestuntersuchung mit zwei Persönlichkeitsinventaren am Vorsorgeheim Hundseck der Bundesknappschaft Bochum [Sociometric therapy effects: A pretest–posttest investigation with two personality inventories at the preventive care facility Hundseck of the Bundesknappschaft Bochum]. *Zeitschrift für Psychosomatische Medizin und Psychoanalyse, 18,* 369–389. (A)

*Ashcraft, C. W. (1971). The later school achievement of treated and untreated emotionally handicapped children. *Journal of School Psychology, 9,* 338–342. (C)

Barlow, D. H. (1994). Psychological intervention in the era of managed competition. *Clinical Psychology: Science and Practice, 1,* 109–122.

*Barron, F., & Leary, T. F. (1955). Changes in psycho-neurotic patients with and without psychotherapy. *Journal of Consulting and Clinical Psychology, 19,* 239–245. (A)

*Becker, P., Kessler, B., & Fuchsgruber, K. (1975). Ein vergleichendes therapieexperiment zur theorie und effizienz der stotterbehandlung auf operanter grundlage [A comparative therapy experiment regarding the theory and efficacy of stuttering therapy on the basis of operant conditioning]. *Archiv für Psychologie, 127,* 78–92. (B)

Beutler, L. E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting and Clinical Psychology, 66,* 113–120.

*Block, J. (1978). Effects of a rational–emotive mental health program on poorly achieving, disruptive high school students. *Journal of Counseling Psychology, 25,* 61–65. (A)

Bovasso, G. B., Eaton, W. W., & Armenian, H. K. (1999). The long-term outcome of mental health treatment in a population-based study. *Journal of Consulting and Clinical Psychology, 67,* 529–538.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* New York: Wiley.

*Brodaty, H., & Andrews, G. (1983). Brief psychotherapy in family practice: A controlled prospective intervention trial. *British Journal of Psychiatry, 143,* 11–19. (A)

*Brom, D., Kleber, R. J., & Defares, P. B. (1989). Brief psychotherapy for posttraumatic stress disorders. *Journal of Consulting and Clinical Psychology, 57,* 607–612. (B)

*Bruce, J. H. (1981). *The effects of group counseling on selected vocational rehabilitation clients.* Unpublished doctoral dissertation, Florida State University. (B)

*Bruhn, M., Schwab, R., & Tausch, R. (1980). Die Auswirkungen intensiver personenzentrierter Gesprächsgruppen bei Klienten mit seelischen Beeinträchtigungen [The effects of intensive client-centered group therapy on clients with psychological problems]. *Zeitschrift für Klinische Psychologie, 9,* 266–280. (B)

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

*Carrasco, I. (1985). Tratamiento de los problemas de aserción por medio de técnicas cognitivo–conductuales [Treatment of assertiveness problems with cognitive–behavioral techniques]. *Revista Española de Terapia del Comportamiento, 3*(1), 29–50. (B)

*Cattell, R. B., Rickels, K., Weise, C., Gray, B., & Yee, R. (1966). The effects of psychotherapy upon measured anxiety and regression. *American Journal of Psychotherapy, 20,* 261–269. (A)

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

*Clements, B. E. (1966). Transitional adolescents, anxiety, and group counseling. *Personnel and Guidance Journal, 45,* 67–71. (B)

*Coche, E., & Flick, A. (1975). Problem-solving training groups for hospitalized patients. *Journal of Psychology, 91,* 19–29. (A)

Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Information Journal, 22,* 343–352.

Coleman, M., & Glofka, P. T. (1969). Effect of group therapy on self-concept of senior nursing students. *Nursing Research, 18,* 274–275.

*Comas-Diaz, L. (1981). Effects of cognitive and behavioral group treatment on the depressive symptomalogy of Puerto Rican women. *Journal of Consulting and Clinical Psychology, 49,* 627–632. (A)

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

*Costantino, G., Malgady, R. G., & Rogler, L. H. (1986). Cuento therapy: A culturally sensitive modality for Puerto Rican children. *Journal of Consulting and Clinical Psychology, 54,* 639–645. (A)

*Coven, A. B. (1970). *The effects of counseling and verbal reinforcement on the internal–external control of the disabled.* Unpublished doctoral dissertation, University of Arizona. (B)

*Davis, H. J. (1975). *The efficacy of rational–emotive imagery in the treatment of test anxiety.* Unpublished doctoral dissertation, Southern Illinois University, Carbondale. (B)

*De Fries, A., Jenkins, S., & Williams, E. C. (1964). Treatment of disturbed children in foster care. *American Journal of Orthopsychiatry, 34,* 615–624. (C)

*Desrats, R. G. (1975). *The effects of developmental and modeling group counseling on adolescents in child care institutions.* Unpublished doctoral dissertation, Lehigh University. (B)

Donenberg, G. R., Lyons, J. S., & Howard, K. I. (1999). Clinical trials versus mental health services research: Contributions and connections. *Journal of Clinical Psychology, 55,* 1135–1146.

*Endicott, N. A., & Endicott, J. (1964). Prediction of improvement in treated and untreated patients using the Rorschach prognostic rating scale. *Journal of Consulting Psychology, 28,* 342–348. (A)

*Ezzo, F. R. (1980). A comparative outcome study of family therapy and positive parenting with court referred adolescents. *Dissertation Abstracts International, 40,* 6198. (University Microfilms No. 80-13836) (A)

*Fernandez-Rodriguez, C., & Perez-Alvarez, M. (1987). Modificación de conducta y mejora en el cumplimiento del tratamiento en diabéticos tipo II [Behavioral modification and improvement of treatment compliance in Type II diabetics]. *Revista Española de Terapia del Comportamiento, 5,* 233–248. (A)

*Fischer, J., Anderson, J., Arveson, E., & Brown, S. (1978). Alderian family counseling: An evaluation. *International Journal of Family Counseling, 6,* 42–44. (A)

*Florin, I., Rudolf, R., & Meyer-Osterkamp, S. (1973). Eine Untersuchung zum operanten Konditionieren sozialen Verhaltens bei chronisch Schizophrenen [An investigation concerning the operant conditioning of social behavior in chronic schizophrenics]. *Zeitschrift für Klinische Psychologie, 2*(Suppl. 1). (A)

*Foulds, M. L. (1970). Effects of a personal growth group on a measure of self-actualization. *Journal of Humanistic Psychology, 10,* 33–38. (B)

*Garrigan, J. J., & Bambrick, A. (1975). Short term family therapy with emotionally disturbed children. *Journal of Marriage and Family Counseling, 1,* 379–385. (B)

*Garrigan, J. J., & Bambrick, A. (1977). Family therapy for disturbed

children: Some experimental results in special education. *Journal of Marriage and Family Counseling, 3,* 83–93. (B)

Gerler, E. R. (1980). A longitudinal study of multimodal approaches to small group psychological education. *School Counselor, 27,* 184–190.

Gilbert, J. P., McPeek, B., & Mosteller, F. (1977, November 18). Statistics and ethics in surgery and anesthesia. *Science, 198,* 684–689.

*Gilbreath, S. H. (1967). Group counseling with male underachieving college volunteers. *Personnel and Guidance Journal, 45,* 469–476. (B)

*Hammen, C. L., & Glass, D. R. (1975). Depression, activity, and evaluation of reinforcement. *Journal of Abnormal Psychology, 84,* 718–721. (We used Experiment 1 from this report of two experiments.) (B)

*Hannemann, E. (1979). Short-term family therapy with juvenile status offenders and their families. *Dissertation Abstracts International, 40,* 1894B. (University Microfilms No. 79-22867) (A)

*Hardcastle, D. (1977). A mother–child, multiple-family counseling program: Procedures and results. *Family Process, 16,* 67–74. (B)

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods, 1,* 154–169.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York: Wiley.

*Horowitz, M. J., Weiss, D. S., Kaltreider, N., Krupnick, J., Marmar, C., Wilner, N., & DeWitt, K. (1984). Reactions to the death of a parent: Results from patients and field subjects. *Journal of Nervous and Mental Disease, 172,* 383–392. (B)

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). The evaluation of psychotherapy: Efficacy, effectiveness, patient progress. *American Psychologist, 10,* 1059–1064.

*Jacob, T., Magnussen, M. G., & Kemler, W. M. (1972). A follow-up of treatment terminators and remainers with long-term and short-term symptom duration. *Psychotherapy: Theory, Research and Practice, 9,* 139–142. (C)

*Jakibchuk, Z., & Smeriglio, V. L. (1976). The influence of symbolic modeling on the social behavior of preschool children with low levels of social responsiveness. *Child Development, 47,* 838–841. (B)

*Jones, F. D., & Peters, H. N. (1952). An experimental evaluation of group psychotherapy. *Journal of Abnormal and Social Psychology, 47,* 345–353. (A)

Kadera, S. W., Lambert, M. J., & Andrews, A. A. (1996). How much therapy is enough? A session-by-session analysis of the psychotherapy dose–effect relationship. *Journal of Psychotherapy Practice and Research, 5,* 132–151.

*Kassinove, H., Miller, N., & Kalin, M. (1980). Effects of pretreatment with rational emotive bibliotherapy and rational emotive audiotherapy on clients waiting at community mental health centers. *Psychological Reports, 46,* 851–857. (A)

*Katz, A., de Krasinski, M., Philip, E., & Wieser, C. (1975). Change in interactions as a measure of effectiveness in short-term family therapy. *Family Therapy, 2,* 31–56. (A)

Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology, 66,* 3–6.

*Kilman, P. C., Henry, S. E., Scarbro, H., & Laughlin, J. E. (1977). The impact of affective education on elementary school underachievers. *Psychology in the Schools, 16,* 217–233. (A)

*Kinnick, B. C., & Shannon, J. T. (1965). The effect of counseling on peer group acceptance of socially rejected students. *School Counselor, 12,* 162–166. (B)

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

Kreft, I. G. G., deLeeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

*Krop, N., Calhoon, B., & Verrier, R. (1971). Modification of the "self-concept" of emotionally disturbed children by covert reinforcement. *Behavior Therapy, 2,* 201–204. (A)

Lambert, M. J., & Cattani-Thompson, K. (1996). Current findings regarding the effectiveness of counseling: Implications for practice. *Journal of Counseling and Development, 74,* 601–608.

*Larcombe, N. A., & Wilson, P. H. (1984). An evaluation of cognitive-behaviour therapy for depression in patients with multiple sclerosis. *British Journal of Psychiatry, 145,* 366–371. (B)

*Lehrman, L. J., Sirluck, H., Black, B. J., & Glick, S. J. (1949). Success and failure of treatment of children in the child guidance clinics of the Jewish Board of Guardians, New York City. *Jewish Board of Guardians Research Monographs, 1.* (C)

*Levitt, E. E., Beiser, H. R., & Robertson, R. E. (1959). A follow-up evaluation of cases treated at a community child guidance clinic. *American Journal of Orthopsychiatry, 29,* 337–347. (C)

*Lineham, M. M., Goldfried, M., & Goldfried, A. P. (1979). Assertion therapy: Skill training or cognitive restructuring? *Behavior Therapy, 10,* 372–388. (B)

*Lipsky, M. J., Kassinove, H., & Miller, N. J. (1980). Effects of rational emotive therapy, rational role reversal and rational emotive imagery on the emotional adjustment of community mental health center patients. *Journal of Consulting and Clinical Psychology, 48,* 366–374. (A)

*Macia, D., & Mendez, F. X. (1986). Programa de intervención conductual para el cumplimiento de las prescripciones médicas [Behavioral intervention program to enhance compliance with medical prescriptions]. *Revista de Psicologia General y Aplicada, 41,* 369–377. (A)

Maldonado, A. (1984). Terapia de conducta y depresión: Un análisis experimental de las interacciones entre tratamientos cognitivos y conductuales con tratamientos farmacológicos en sujetos depresivos [Behavior therapy and depression: Experimental analysis of interactions between behavioral and cognitive treatments with pharmacological treatments in depressive subjects]. *Revista de Psicologia General y Aplicada, 39,* 517–535.

*Manos, N., & Vasilopoulou, E. (1984). Evaluation of psychoanalytic psychotherapy outcome. *Acta Psychiatrica Scandinavia, 70,* 28–35. (B)

*Martinez-Sánchez, J. J. (1986). *Eficacia terapeútica y niveles de cambio: La eficacia diferencial y el estudio de distintos parámetros de personalidad y parámetros situacionales* [Therapy efficacy and change levels: Differential efficacy and study of different personality and situational parameters]. Unpublished master's thesis, University of Valencia, Valencia, Spain. (B)

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.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Matt, G. E., & Navarro, A. M. (1997). What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. *Clinical Psychology Review, 17,* 1–32.

*Mavissakalian, M., & Michelson, L. (1983). Self-directed *in vivo* exposure practice in behavioral and pharmacological treatments of agoraphobia. *Behavior Therapy, 14,* 506–519. (A)

Mental health: Does therapy help? (1995, September). *Consumer Reports, 60,* 734–739.

*Meyer, A. E. (Ed.). (1981). The Hamburg short psychotherapy comparison experiment. *Psychotherapy and Psychosomatics, 35,* 81–208. (A)

*Meyer, J. B., Strowig, W., & Hosford, R. E. (1970). Behavioral rein-

forcement counseling with rural high school youth. *Journal of Counseling Psychology, 17,* 127–132. (B)

Meyer, T. J., & Mark, M. M. (1995). Effects of psychosocial interventions with adult cancer patients: A meta-analysis of randomized experiments. *Health Psychology, 14,* 101–108.

*Mitchell, K. R., & Ingham, R. J. (1970). The effects of general anxiety on group desensitization of test anxiety. *Behaviour Research and Therapy, 8,* 69–78. (B)

Moleski, R., & Tosi, D. J. (1976). Comparative psychotherapy: Rational emotive therapy versus systematic desensitization in the treatment of stuttering. *Journal of Consulting and Clinical Psychology, 44,* 309–311.

*Morrow, G. R. (1986). Effect of the cognitive hierarchy in the systematic desensitization treatment of anticipatory nausea in cancer patients: A component comparison with relaxation only, counseling, and no treatment. *Cognitive Therapy and Research, 10,* 421–466. (A)

Nathan, P. E. (1998). Practice guidelines: Not yet ideal. *American Psychologist, 53,* 290–299.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy Press.

*Naun, R. J. (1971). Comparison of group counseling approaches with Puerto Rican boys in an inner city high school. *Dissertation Abstracts International, 32,* 742–743A. (B)

Norcross, J. C., & Prochaska, J. O. (1982). A national survey of clinical psychologists: Characteristics and activities. *Clinical Psychologist, 35,* 5–8.

*Obler, M., & Terwilliger, R. F. (1969). Pilot study on the effectiveness of systematic desensitization with neurologically impaired children with phobic disorders. *Journal of Consulting and Clinical Psychology, 34,* 314–318. (B)

*Omizo, M. M., & Michael, W. B. (1982). Biofeedback-induced relaxation training and impulsivity, attention to task, and locus of control among hyperactive boys. *Journal of Learning Disabilities, 15,* 414–416. (A)

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Pallack, M. S. (1995). Managed care and outcomes-based standards in the health care revolution. In S. C. Hayes, V. M. Follette, R. J. Dawes, & K. E. Grady (Eds.), *Scientific standards of psychological practice: Issues and recommendations* (pp. 73–77). Reno, NV: Context Press.

*Parson, B. V., & Alexander, J. F. (1973). Short-term family intervention: A therapy outcome study. *Journal of Consulting and Clinical Psychology, 41,* 195–201. (A)

*Paul, G. L. (1966). *Insight vs. desensitization in psychotherapy: An experiment in anxiety reduction.* Stanford, CA: Stanford University Press. (B)

Perlman, B. (1985). A national survey of APA-affiliated masters-level clinicians: Description and comparison. *Professional Psychology: Research and Practice, 16,* 553–564.

*Persons, R. W., & Pepinsky, H. B. (1966). Convergence in psychotherapy with delinquent boys. *Journal of Counseling Psychology, 13,* 329–334. (A)

*Piper, W. E., Azim, H. F., McCallum, M., & Joyce, A. S. (1990). Patient suitability and outcome in short-term individual psychotherapy. *Journal of Consulting and Clinical Psychology, 58,* 475–481. (A)

*Poser, E. G. (1966). The effects of therapists' training on group therapeutic outcome. *Journal of Consulting Psychology, 30,* 283–289. (A)

Prochaska, J. O., & Norcross, J. C. (1983). Contemporary psychotherapists: A national survey of characteristics, practices, orientations, and attitudes. *Psychotherapy: Theory, Research and Practice, 20,* 161–173.

Ragsdale, K. K. (1996). *Selection bias in nonrandomized experiments: An investigation using meta-analysis.* Unpublished doctoral dissertation, The University of Memphis.

*Reiter, G. F., & Kilmann, P. R. (1975). Mothers as family change agents. *Journal of Counseling Psychology, 22,* 61–65. (B)

*Rosser, R., Denford, J., Heslop, A., Kinston, W., Macklin, D., Minty, K., Moynihan, C., Muir, B., Rein, L., & Guz, A. (1983). Breathlessness and psychiatric morbidity in chronic bronchitis and emphysema: A study of psychotherapeutic management. *Psychological Medicine, 13,* 93–110. (B)

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics, 17,* 363–374.

*Sacks, J. M., & Berger, S. (1954). Group therapy techniques with hospitalized chronic schizophrenic patients. *Journal of Consulting Psychology, 18,* 297–302. (A)

*Schandl, V., & Löschenkohl, E. (1980). Kind im Krankenhaus: Evaluierung eines Interventionsprogrammes bei Verhaltensstörungen [Children in the hospital: Evaluating an intervention program for behavior disorders]. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 29,* 252–258. (A)

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

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

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do you get the same answer? In W. J. Bukoski (Ed.), *Meta-analysis of drug abuse prevention programs* (pp. 147–164). Washington, DC: Superintendent of Documents.

Shadish, W. R., Matt, G., Novaro, A., Siegle, G., Crits-Christoph, P., Hazelrigg, M., Jorm, A., Lyons, L. S., Nietzel, M. T., Prout, H. T., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works under clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Shadish, W. R., Robinson, L., & Lu, C. (1999). *ES: A computer program and manual for effect size calculation.* Minneapolis, MN: Assessment Systems.

Shadish, W. R., & Sweeney, R. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology, 59,* 883–893.

Shapiro, D. A., Barkham, M., Rees, A., Hardy, G. E., Reynolds, S., & Startup, M. (1994). Effects of treatment duration and severity of depression on the effectiveness of cognitive-behavioral and psychodynamic-interpersonal psychotherapy. *Journal of Consulting and Clinical Psychology, 62,* 522–534.

*Shefler, G., & Dasberg, H. (1989, June). *A randomized controlled outcome and follow-up study of James Mann's time-limited psychotherapy in a Jerusalem community mental health center.* Paper presented at the annual convention of the Society for Psychotherapy Research, Toronto, Ontario, Canada. (A)

*Shepherd, M., Oppenheim, A. N., & Mitchell, S. (1966). Childhood behaviour disorders and the child-guidance clinic: An epidemiological study. *Journal of Child Psychology and Psychiatry, 7,* 39–52. (C)

*Sloan, R. B., Staples, F. R., Cristol, A. H., Yorkston, N. J., & Whipple, K. (1975). *Psychotherapy versus behavior therapy.* Cambridge, MA: Harvard University Press. (A)

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

*Smith, R. E., & Nye, S. L. (1973). A comparison of implosive therapy and

systematic desensitization in the treatment of test anxiety. *Journal of Consulting and Clinical Psychology, 41,* 37–42. (B)

*Smyrnios, K. X., & Kirkby, R. J. (1993). Long-term comparison of brief versus unlimited psychodynamic treatments of children and their parents. *Journal of Consulting and Clinical Psychology, 61,* 1020–1027. (C)

*Spiegler, M. D., Cooley, E. J., Marshall, G. J., Prince, H. T., II, Puckett, S. P., & Skenazy, J. A. (1976). A self-control versus a counterconditioning paradigm for systematic desensitization: An experimental comparison. *Journal of Counseling Psychology, 23,* 83–86. (B)

*Steier, F. (1983). Family interaction and properties of self-organizing systems: A study of family therapy with addict families. *Dissertation Abstracts International, 44,* 863A. (University Microfilms No. 83-16093) (A)

*Stotsky, B. A., Daston, P. G., & Vardack, C. N. (1955). An evaluation of the counseling of chronic schizophrenics. *Journal of Counseling Psychology, 2,* 248–255. (A)

*Stover, L., & Guerney, B. (1967). The efficacy of training procedures for mothers in filial therapy. *Psychotherapy: Theory, Research, and Practice, 4,* 110–115. (A)

*Szapocznik, J., Santisteban, D., Rio, A., Perez-Vidal, A., Santisteban, D., & Kurtines, W. M. (1989). Family effectiveness training: An intervention to prevent drug abuse and problem behaviors in hispanic adolescents. *Hispanic Journal of Behavioral Sciences, 11,* 4–27. (A)

*Tausch, A.-M., Kettiner, U., Steinbach, I., & Tonnies, S. E. (1973). Effekte kindzentrierter Einzel- und Gruppengesrache mit unterprivilegierten Kindergarten- und Grundschulkindern [Effects of child-centered individual and group therapy with underprivileged preschool and elementary school children]. *Psychologie in Erziehung und Unterricht, 20,* 77–88. (B)

*Thompson, L. W., Gallagher, D., & Breckenridge, J. S. (1987). Comparative effectiveness of psychotherapies for depressed elders. *Journal of Consulting and Clinical Psychology, 55,* 385–390. (A)

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358–376.

*Truax, C. B., Wargo, D. G., & Silber, L. D. (1966). Effects of group psychotherapy with high accurate empathy and nonpossessive warmth upon female institutionalized delinquents. *Journal of Abnormal Psychology, 71,* 267–274. (A)

Trull, T. J., Nietzel, M. T., & Main, A. (1988). The use of meta-analysis to assess the clinical significance of behavior therapy for agoraphobia. *Behavior Therapy, 19,* 527–538.

Ullrich de Muynck, R., & Ullrich, R. (1980a). Erster Effizienznachweis des Assertiveness-Training-Programm (ATP) [First demonstration of the efficacy of the Assertiveness Training Program (ATP)]. In K. Grawe (Ed.), *Soziale Kompetenz* (Vol. 2, pp. 1–20). Munich, Germany: Pfeiffer.

Ullrich de Muynck, R., & Ullrich, R. (1980b). Spezifische Effekte des ATP im Arbeits- und Leistungsbereich [Specific effects of the ATP in the work and performance domain]. In K. Grawe (Ed.), *Soziale Kompetenz* (Vol. 2, pp. 21–32). Munich, Germany: Pfeiffer.

*Walker, J. M. (1985). A study of the effectiveness of social learning family therapy for reducing aggressive behavior in boys. *Dissertation Abstracts International, 45,* 3088B. (University Microfilms No. 84-29327) (B)

*Ward, H. C. (1966). *Effects of non-directive group counseling upon selective cognitive functioning and interpersonal relationships of junior high students.* Unpublished doctoral dissertation. East Texas State University. (B)

*Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52,* 666–678. (A)

*Weissman, M. M., Prusoff, B. A., DiMascio, A., Neu, C., Goklaney, M., & Klerman, G. L. (1979). The efficacy of drugs and psychotherapy in the treatment of acute depressive episodes. *American Journal of Psychiatry, 136,* 555–558. (A)

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 688–701.

*Weisz, J. R., & Weiss, B. (1989). Assessing the effects of clinic-based psychotherapy with children and adolescents. *Journal of Consulting and Clinical Psychology, 57,* 741–746. (C)

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.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

White, K., & Allen, R. (1971). Art counseling in an educational setting: Self-concept change among pre-adolescent boys. *Journal of School Psychology, 9,* 218–225.

*Whitman, T. L. (1969). Modification of chronic smoking behavior: A comparison of three approaches. *Behavior Research and Therapy, 7,* 257–263. (B)

*Winston, A., Pollack, J., McCullough, L., Flegenheimer, W., Kestenbaum, R., & Trujillo, M. (1991). Brief psychotherapy of personality disorders. *Journal of Nervous and Mental Disease, 179,* 188–193. (B)

*Witmer, H. L., & Keller, J. (1942). Outgrowing childhood problems: A study in the value of child guidance treatment. *Smith College Studies in Social Work, 13,* 74–90. (C)

*Wolk, R. L., & Goldfarb, A. I. (1967). The response to group psychotherapy of aged recent admissions compared with long-term mental hospital patients. *American Journal of Psychiatry, 123,* 1251–1257. (A)

Appendix

Coding Criteria

Clinical Representativeness Codes

Clinically representative problems (percentage agreement = 88%)[a]

1 = *Clinically representative:* Participants had mental health or behavioral problems. Examples: traditional clinically distressed patients, patients whose clinical symptoms are secondary to a primary medical problem such as multiple sclerosis, patients being treated for resulting depression, and classroom behavior problems.

0.5 = *Partially clinically representative:* Participants had problems that were not mental health or behavior problems. Examples: students who are identified as underachievers, patients treated for noncompliance with a medical regime such as taking hypertensive medications.

0 = *Not clinically representative:* Participants were without identified problems. Examples: unselected nursing students wanting personal growth, unselected grammar school children who are taught about interpersonal relationships.

Clinically representative setting (percentage agreement = 88%; κ = .605)

1 = *Clinically representative:* a setting in which clinical services are commonly provided and that would be considered primarily a service-delivery site. Examples: outpatient mental health clinics, community mental health centers, general hospitals, Veterans Administration, private practice, prisons, school systems, university-affiliated service-delivery clinics whose primary function is clinical.

0.5 = *Partially clinically representative:* a mixed clinical–research setting in which both clinical and research functions routinely occur. Examples: university-affiliated medical schools or free-standing clinics that regularly conduct research, psychological services centers associated with academic psychology departments.

0 = *Not clinically representative:* a setting in which the research function clearly dominates any clinical function. Examples: a research laboratory on an academic campus, a university-affiliated clinic devoted entirely to research.

Clinically representative referrals (percentage agreement = 71%; κ = .400)

1 = *Clinically representative:* Patients in the treatment condition were initially referred through usual clinical routes. Examples: health professional referral, self-referral, referral by family/friend, treatment required.

0 = *Not clinically representative:* researcher-solicited patients; patients in the treatment condition initially solicited by experimenter without going through any usual clinical referral route. Examples: researcher media advertisement, introductory psychology students receiving course credit for research.

Clinically representative therapists (percentage agreement = 88%; κ = .800)

1 = *Clinically representative:* Therapists are practicing mental health clinicians (the provision of such services is a substantial part of their job duties) and professionals (they earn their living in this job). Examples: clinic therapist, counselor in a school setting, recovered alcoholics paid to counsel other alcoholics.

0.5 = *Partially clinically representative:* professional clinical researchers who are qualified to provide clinical services but their primary duties are in research and the provision of clinical services occurs infrequently. Example: clinical psychologist employed in academic psychology department and seeing clients rarely.

0 = *Not clinically representative:* nonclinicians or clinicians in training; therapists not practicing clinical professionals or in training. Examples: medical general practitioner, graduate students, psychiatry residents.

Clinically representative structure (percentage agreement = 76%; κ = .406)

1 = *Clinically representative:* Treatment was not structured in a detailed and uniform way, or its structure was representative of some clinical practice. Examples: psychodynamic therapy without a detailed manual, or therapy structured by relaxation tapes, systematic desensitization, or biofeedback.

0 = *Not clinically representative:* Therapy was structured in a way that is not representative of clinical practice. Examples: psychodynamic therapy where therapists were instructed to follow a detailed manual or protocol or that used a structure such as video or audiotapes prepared specifically for this study and not typically used in clinical practice.

Clinically representative monitoring (percentage agreement = 82%; κ = .610)

1 = *Clinically representative:* Study did not monitor the implementation of treatment in a way that could influence therapist behavior in the study. Examples: therapy without supervision or monitoring, videotaping and coding of therapist behavior for later use as a dependent variable without feedback to therapist, therapists could consult with an expert to see if they were doing therapy correctly but did not have regular sessions with that expert.

0 = *Not clinically representative:* Study monitored treatment implementation regularly for its integrity and adherence to a treatment plan or model. Examples: observing therapist behavior and providing immediate feedback to therapists, supervision given in a way to affect therapist behavior.

Clinically representative problem heterogeneity (percentage agreement = 100%; κ = 1.00)

1 = *Clinically representative:* Therapists treated clients who were heterogeneous in focal presenting problems across all clients treated both in and out of study. Examples: different clients having different problems, the same clients having multiple presenting problems.

0 = *Not clinically representative:* Therapists treated only clients who were homogeneous in presenting problem both in and outside of the study. Examples: graduate students who were not described as doing therapy outside the study, a professional therapist working solely to relieve patient pain in a pain clinic.

Clinically representative pretherapy training (percentage agreement = 76%; κ = .393)

1 = *Clinically representative:* Therapists did not receive intensive training before the study in the treatment. Example: therapists providing usual care, experienced rational–emotive therapists not trained specially for the study.

0 = *Not clinically representative:* Therapists received intensive training before the study in the treatment. Example: graduate students trained in treatment before the study.

(*Appendix continues*)

Clinically representative therapy freedom (percentage agreement = 100%; κ = 1.00)

1 = *Clinically representative:* Therapists used multiple techniques in all the therapy they did. Examples: counselors in a school, therapists in a mental health center.
0 = *Not clinically representative:* Therapists relied on a specific structured technique or a narrow set of substantially similar techniques in their work, both in the study and outside the study. Examples: therapists in a pain clinic who routinely administer the same treatment to all clients, therapists who work full time in a study in which therapy is narrowly constrained.

Clinically representative flexible number of sessions (percentage agreement = 59%)[a]

1 = *Clinically representative:* The research did not place limits on the number of therapy sessions. Examples: unlimited psychodynamic therapy, therapy that could continue after posttest.
0 = *Not clinically representative:* Therapy was limited to a fixed number of sessions. Example: a structured hypnosis treatment with fixed sessions.

Other Treatment Characteristics

Treatment orientation (percentage agreement = 94%; weighted κ = .881)

1 = *Behavioral/psychoeducational/problem solving.*
2 = *Systemic.*
3 = *Humanistic/experiential.*
4 = *Psychodynamic/psychoanalytic.*
5 = *Eclectic.* Examples: A treatment specifically labeled eclectic, or a combination of two more of the above orientations.
6 = *Didactic.* Examples: bibliotherapy without therapist, classroom instruction.
7 = *Other.*

Number of sessions (r_i = .994)
Duration of sessions in minutes (r_i = .928)
Brief therapy (percentage agreement = 100%; κ = 1.00)

1 = *Short duration* (fewer than 10 sessions).
2 = *Long duration* (10 or more sessions).

Use of structure (percentage agreement = 82%; κ = .485)

1 = *Did not use a formal structure.* Examples: no structure such as a manual, video or audiotapes, or detailed instructions to therapists about how to conduct therapy.
0 = *Did use a formal structure.*

Dependent Variable Characteristics

Outcome state (percentage agreement = 76%; weighted κ = .493)

1 = *Primarily a measure of behavior.* Examples: behavioral observations, self-report of behavior.
2 = *Primarily nonbehavioral.* Examples: measures of thoughts, affect, physiology.
3 = *Primarily achievement test.* Examples: Scholastic Aptitude Test score.
4 = *Other.*

Outcome Mode (percentage agreement = 88%; weighted κ = .757)

1 = *Self-report of client.* Examples: Beck Depression Inventory, Minnesota Multiphasic Personality Inventory (MMPI).
2 = *Therapist rating.* Example: improvement rating by therapist.
3 = *Rating by other.* Examples: one family member rating another, trained observer, physiological measures taken by others.
4 = *Other.*

Manipulability (percentage agreement = 76%; κ = .590)

1 = *Not very manipulable:* measures not easily controlled by clients or therapists. Examples: most physiological measures, grade point average.
2 = *Moderately manipulable:* manipulable a cost to the respondent. Example: an observer-rated problem-resolution task requiring spouses to comply with treatment recommendations that are inconsistent with their normal behavior.
3 = *Very manipulable:* manipulable at no cost to the respondent. Examples: self-report of satisfaction, therapist rating of outcome.

Reactivity (from M. L. Smith, Glass, & Miller, 1980; percentage agreement = 53%; κ = .470)

1 = *Physiological measures.* Examples: palmar sweat index, pulse, galvanic skin response, grade point average.
2 = *Blinded ratings and decisions.* Examples: blind projective test ratings, blind ratings of symptoms, blind discharge from hospital.
3 = *Standardized measures of traits having minimal connection with treatment or therapist.* Examples: MMPI, Rotter Internal–External Control Scale.
4 = *Experimenter-constructed inventories (nonblind), ratings of symptoms (nonblind), any client self-report to experimenter, blind administration of Behavioral Approach Tests.*
5 = *Therapist rating of improvement or symptoms, projective tests (nonblind), behavior in the presence of therapist or nonblind evaluator (e.g., Behavioral Approach Test), instruments that have a direct and obvious relationship with treatment (e.g., where desensitization hierarchy items were taken directly from measuring instrument).*

Specificity (percentage agreement = 83%; κ = .726)

1 = *Specific:* measures directly constructed from or related to the goals of treatment. Examples: target behaviors that are focus of therapy, count of number of quarrels as dependent variable for communication training to reduce quarrels.
2 = *Not specifically tailored to treatment, but a general therapy measure.* Example: a general rating of distress.
3 = *General:* Measure tangentially related to therapy. Example: IQ test.

Number of weeks after treatment terminated that this measure was taken, with zero being immediately after therapy (r_i = .974)

Methodology Codes

Assignment to condition (percentage agreement = 100%; κ = 1.00)

1 = *Clients were randomly assigned to treatment and comparison conditions for this effect size.*
2 = *Clients were not randomly assigned.*
3 = *Clients were haphazardly assigned.* Example: alternating order.

Matching/blocking/stratifying (percentage agreement = 100%; weighted κ = 1.00)

1 = *Matching/blocking/stratifying carried out on reliable variable.* Example: gender.
2 = *Matching/blocking/stratifying carried out on fallible variable.* Example: IQ.
3 = *Matching/blocking/stratifying carried out on both reliable and fallible variables.*

4 = *Matching/blocking/stratifying not reported.*

Control group activity level (percentage agreement = 100%; $\kappa = 1.00$)

1 = *Passive control.* Example: no treatment or wait list.
2 = *Active control.* Example: placebo.

Control group similarity to treatment group (percentage agreement = 88%)[a]

1 = *Internal:* Another group from the same pool of participants. Examples: Students from the same school, all randomized groups.
2 = *External:* A group from a patently different pool of clients. Example: students from different schools.

Selection process (percentage agreement = 88%)[a]

1 = *Self-selection:* Participants actively chose to which condition they were assigned. Example: clients who respond to a media add for a treatment.
2 = *Other-selection:* Someone else selected units into condition. Examples: all randomized studies, nonrandomized studies in which the experimenter gave treatment to participants who could make certain appointment times.

Miscellaneous Codes

Year of publication ($r_i = 1.00$)
Publication status (percentage agreement = 100%; weighted $\kappa = 1.00$)

1 = *Journal.*
2 = *Book or book chapter.*
3 = *Dissertation or master's thesis.*
4 = *Convention paper or other speech.*
5 = *Unpublished manuscript.*

Child–adolescent or adult presenting problems (percentage agreement = 94%; $\kappa = .866$)

1 = *Child–adolescent presenting problems.* Example: family therapy with a child presenting problem.
2 = *Adult presenting problems.*

Number of participants initially assigned ($r_i = .985$)
Number of participants remaining at outcome ($r_i = .995$)

Coding manuals with more detail are available from William R. Shadish. In this appendix, $N = 17$ for all reliability coefficients, and the symbol r_i represents the intraclass correlation (Tinsley & Weiss, 1975).

[a] No other reliability coefficient could be computed because one set of ratings had no variability.

Received January 14, 1999
Revision received January 10, 2000
Accepted January 11, 2000 ■