Journal of Consulting and Clinical Psychology
1983, Vol. 51, No. 1, 42-53

0022-006X/83/5101-0042$00.75

Comparative Therapy Outcome Research: Methodological Implications of Meta-Analysis

David A. Shapiro and Diana Shapiro
Medical Research Council / Social Science Research Council
Social and Applied Psychology Unit, University of Sheffield, Sheffield, England

Findings obtained in the course of a meta-analysis of 143 outcome studies, published over a 5-year period, in which 2 or more treatments were compared with a control group, are used to evaluate the quality of such research. The statistical conclusion and internal validity of the research reviewed are generally satisfactory, although construct validity is insufficient to rule out the influence of nonspecific and demand effects. The construct and external validity of the work reviewed are severely limited by its unrepresentativeness of clinical practice. Several studies are marred by nonreproducible accounts of treatments, inadequate description of clients, therapists, and design features, and faulty data presentation. It is concluded that this meta-analysis underlines the urgent need for greater methodological diversity and clinical realism in therapy research.

Recent reviews of comparative psychotherapy-outcome research (Frank, 1979; Luborsky, Singer, & Luborsky, 1975; Smith & Glass, 1977; Smith, Glass, & Miller, 1980) have converged on the conclusion that the widely differing available therapies are modestly, but equally, effective. However, this view has been contested by proponents of behavioral methods (Eysenck, 1978; Kazdin & Wilson, 1978; Rachman & Wilson, 1980). These authors assert that such a conclusion is contradicted by the results of well-conducted studies favoring behavioral methods (Bandura, 1977; Franks & Wilson, 1978; Rachman & Hodgson, 1980) and owes its support to the above-cited reviewers' ill-conceived aggregation of data from unsound research studies.

The reviews by Smith and Glass (1977) and Smith et al. (1980) used meta-analysis, the numerical combination of data from independent studies, to generate collective, summary statistics pertaining to the research question to which they are addressed (Glass, McGaw, & Smith, 1981). The method involves the application of the methodological principles of empirical research to the literature review process, in pursuit of superior objectivity and dependability. Beyond the field of therapy, the method has been applied to such problems as the effects of school class size on attainment (Glass & Smith, 1979), the relationship between social class and achievement (White, 1977), interpersonal expectancy or "experimenter effects" (Rosenthal & Rubin, 1978), self-serving bias in interpersonal influence situations (Arkin, Cooper, & Kolditz, 1980), and sex differences in conformity (Cooper, 1979). Elements of the method include calculation and aggregation across studies of indexes of statistical significance or magnitude of effect, systematic and often exhaustive search procedures, coding of objective and qualitative features of the source studies, and investigation of the correlates of study outcome via disaggregation and regression analysis (Shapiro & Shapiro, 1982a).

A major criticism of meta-analysis concerns its aggregation of data from studies of diverse quality. Eysenck (1978), for example, invokes the computer scientist's dictum, "garbage in—garbage out." On the other hand, recent appraisals of meta-analysis (Cook & Leviton, 1980; Shapiro & Shapiro,

Portions of this study were presented at the annual meeting of the Society for Psychotherapy Research, Aspen, Colorado, in June, 1981.

Diana Shapiro is now with the North Derbyshire District Psychology Service.

We gratefully acknowledge helpful comments on this work from Chris Barker, Gene Glass, Paul Jackson, Mary Lee Smith, Peter Warr, and an anonymous reviewer.

Requests for reprints should be sent to David A. Shapiro, MRC/SSRC Social and Applied Psychology Unit, The University, Sheffield, S10 2TN, England.

1982a; Strube & Hartmann, 1982) argue that it is preferable to retain data from studies of varying quality and then to determine empirically whether these yield differing results, rather than to discard data deemed "inferior" in terms of the reviewer's preferred methodological criteria. Furthermore, this inclusive strategy also yields additional benefits. By coding substantive and methodological features of a large and representative set of studies, we may obtain a composite profile of a research literature. This is particularly useful in appraising the current state of the art in a research domain and guiding suggestions for further research. The present article is devoted to just such a purpose in respect to the comparative therapy-outcome literature, drawing upon our meta-analysis (Shapiro & Shapiro, 1982b).

The Shapiro and Shapiro Study

Our meta-analysis (Shapiro & Shapiro, 1982b) was designed in response to criticisms of the work of Smith and Glass (1977) advanced by Kazdin and Wilson (1978) and by Rachman and Wilson (1980). We took account also of methodological considerations in comparative-outcome research outlined by Mahoney (1978) and Kazdin (1978). Our aim was to appraise the extent to which an unbiased sample of relatively well-designed, recently published comparative-outcome studies supports the case for differential effectiveness of diverse therapy techniques. Consistent with the emphasis of Mahoney (1978), we confined our attention to studies following a contrast-group design, in which clients are assigned to two or more treatment groups and one or more control groups. This design was followed only by a small minority (less than 10%) of studies concerned with treatment outcomes. Studies were excluded from our meta-analysis if they compared only one active treatment with one or more control groups, if they compared supposedly active treatments without any no-treatment or minimal-treatment groups, or if treatments could only be compared in their data via own-control comparisons, as in multiple-baseline or crossover designs. We confined our search to *Psychological Abstracts*, 1975–79. No reference was made to review articles or bibliographies, in order to avoid the sampling bias such sources might introduce. Our research yielded 143 studies meeting the above design requirements and presenting sufficient data to permit calculation or estimation of effect sizes; only 21 (15%) of these studies were also included by Smith et al. (1980).

Before considering the methodological implications of this study, an outline of its substantive findings is in order. Consistent with previous reviews (Smith & Glass, 1977; Smith et al., 1980) the mean effect size approached 1 standard deviation unit and differences among treatment methods accounted for at most 10% of the variance in effect size. The impact of differences between treatment methods was outweighed by the combined effects of other variables, such as the nature of the target problem under treatment (accounting for some 20% of the variance), aspects of the measurement methods used to assess outcome, and features of the experimental design. The methodological implications of these nontreatment influences upon outcome will be considered below. Multiple regression analysis suggested that differences between treatments were largely independent of these other factors, however. Direct comparisons between pairs of treatments figuring together in the same subsets of the data suggested some consistent differences, with cognitive and certain multimodal behavioral methods yielding favorable results. Dynamic and humanistic methods were sparsely represented in the data and were associated with apparently modest effects.

Issues Addressed

Meta-analysis cannot transcend the limitations of the data upon which it is based. It can but hold a mirror to the scientific community, summarizing the conclusions and the quality of the available evidence concerning the substantive questions at issue. In evaluating the quality of the comparative, controlled-outcome literature encompassed by our meta-analysis, we conceptualize the dependability of conclusions from research in terms of Cook and Campbell's (1979) discussion of validity. These authors have subdivided the concepts of internal and external

validity, familiar from Campbell and Stanley's (1963) treatment, into four types. *Statistical conclusion validity* refers to the validity with which a study permits conclusions about covariation between the assumed independent and dependent variables. Threats to statistical conclusion validity typically arise from unsystematic error rather than systematic bias. *Internal validity* refers to the validity with which statements can be made about whether there is a causal relationship from independent to dependent variables, in the form in which they were manipulated or measured. Threats to internal validity typically involve systematic bias. *Construct validity* of putative causes and effects refers to the validity with which we can make generalizations about higher-order constructs from research operations. Finally, *external validity* refers to the validity with which conclusions can be drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times. In addition, we are concerned with *reporting adequacy*, the extent to which accounts of the procedures and circumstances of each study, and the data obtained, meet minimal requirements of precision and replicability.

The methodological concerns in treatment evaluation raised by Mahoney (1978) and Kazdin (1978) are usefully subsumable within the Cook and Campbell (1979) scheme. For example, Kazdin (1978) urges that the clinical-analogue continuum be analyzed into component dimensions relating to target problem; population; recruitment; therapists, selection, set and setting; treatment; and assessment. Study parameters bearing upon each of these dimensions were coded within our meta-analysis, thus permitting a thorough analysis of the external validity of the research reviewed therein. Since our meta-analytic coding was designed to refine the methods of Smith and Glass (1977) within resource constraints, we do not have data bearing upon every concern voiced by Cook and Campbell (1979), Mahoney (1978), and Kazdin (1978). On the other hand, our data are sufficiently detailed to permit some consideration of all four validity types and to yield important conclusions concerning the current state of the art in outcome research (Frank, 1979; Shapiro, 1980).

Statistical Conclusion Validity

A central issue in statistical conclusion validity concerns statistical power. The greater the sample size, the greater the power of the study to yield a significant effect (Cohen, 1977; Kraemer, 1981). For example, Kraemer (1981) reports that a study with 10 subjects in each of two groups has a 68% probability of detecting an effect size of 1 (group *M*s separated by 1 *SD*) via a two-sample, one-tailed *t* test with a 5% alpha level. Thus the choice of sample size depends on the magnitude of effect the investigator wishes to be able to detect. The 414 treated groups in our meta-analysis contained a mean of 11.98 ($SD = 7.12$) clients; the 143 control groups contained a mean of 12.12 ($SD = 6.64$) clients. Forty two (10%) of the treated groups contained six or fewer clients. Of the treated groups, 263 (64%) contained 10 or more clients and 115 (28%) contained 13 or more clients. If we assume that an effect size of 1 is a reasonable expectation in outcome research, these data suggest that the majority of our published, controlled comparative-outcome studies use adequate, although not impressive, sample sizes. Since the probability of publication is enhanced by the attainment of statistical significance (Shapiro & Shapiro, 1982a), one might expect that published findings with small samples would tend to show larger effects than more powerful studies with larger samples, capable of statistical significance with a smaller effect. Consistent with this expectation, correlations of $-.14$ and $-.21$, respectively, were obtained between treated-group and control-group *N* and effect size.

Uncontrolled variation in the way a treatment is implemented presents a threat to statistical conclusion validity. In our meta-analysis, we examined the precision with which treatment methods were described, assessed in terms of the extent to which the source paper permitted reproduction of the method by other workers. Three levels of *reproducibility* were defined. Of the groups included in our analysis, 159 (38%) received treatments that were well described in the text or in documents available from the source-study author or in other cited work. These treatments were either fully manualized or used

well-defined and standard procedures that any recently trained clinical psychologist might be expected to replicate on the basis of the information supplied. At the opposite extreme, 47 (11%) of the groups were treated by methods that could not be reproduced with any precision on the basis of available information. Between these extremes lay the 208 (50%) groups receiving treatments given outline descriptions or adapted from methods that were well described elsewhere. Thus, almost two-thirds of our data were based on treatments that were not well specified. Even among the 38% that were coded as highly reproducible, full manuals were available in only a minority of cases. Thus, our standards in coding this variable were not exacting. Of course, it must be acknowledged that some treatments are intrinsically more readily specifiable than others. Nonetheless, the majority of the literature reviewed is deficient in this respect. However, the impact of this variable upon outcome was minimal, $r(412) = .03$, $p > .10$.

A related requirement is the monitoring of the actual carrying out of experimental procedures in order to check that the independent variables are manipulated as intended. So rarely was this done in the studies under review that it did not occur in the sample of the studies on which the coding system was developed, and hence was not coded in the meta-analysis proper. This virtually total neglect by researchers of the need to monitor treatments is a source of grave concern.

> One need not lack integrity or competence to exhibit variability in the administration of even the most simple experiment, and these variations may figure prominently in the interpretation of one's data (Mahoney, 1978, p. 665).

Internal Validity

A major safeguard against bias and consequent lack of internal validity is the random assignment of clients to treatment conditions. Only 10% of the client groups within our meta-analysis were not randomly assigned. Although somewhat controversial (Mahoney, 1978, p. 664), procedures in which systematic constraints were imposed on otherwise random assignment (either to assure matched group means or to match individual cases), as adopted in 32% of the groups under study, were considered to result in at least as much control over extraneous sources of variation as achieved by unconstrained randomization. It is of some interest to note that the 57% of groups using unconstrained randomization yielded an overall effect size typical of the study as a whole, whereas the poorest designs (failing to randomize) yielded a somewhat reduced effect size, $M = .76$, although there was no significant effect of the assignment variable upon effect size, $F(3, 384) = 1.54$, $p > .10$.

A related design feature is client attrition; if high, this can introduce bias in the comparisons between groups. With a mean of 10.68% of clients lost from the treated group ($SD = 13.49$) and a mean of 9.19% of clients lost from the control group ($SD = 13.56$), attrition was commendably low in most cases. Similarly, 75% of the effect sizes were obtained from groups showing no significant differences on any measure at pretreatment testing, and the adequacy of pretreatment equivalence was, if anything, positively correlated with obtained effect size, $r(1401) = .07$, $p < .05$. In these respects, therefore, we may conclude that most studies are well designed and executed and that the overall results are, if anything, attenuated by the presence of inferior studies.

Construct Validity

Representativeness of Treatments

Cook and Campbell (1979) consider the adequacy with which the independent variable is represented within a study as an important aspect of construct validity. Several aspects of treatment as the independent variable are germane to the issue of representativeness. Behavioral and cognitive therapies were overwhelmingly predominant, with only 5% of the groups receiving dynamic or humanistic methods. As reported in detail by Shapiro and Shapiro (1982b), these latter methods yielded relatively modest effect sizes, implying that the overall mean effect size may be inflated relative to clinical practice by the disproportionate preponderance of these methods. Any inference concerning the relative efficacy of behavioral and nonbehav-

ioral methods is, however, subject to the important cautionary limitations detailed by Shapiro and Shapiro (1982b).

Only 166 (41%) of the treated groups were seen individually; the majority, 210 (52%), were treated entirely in groups. Individual therapy was associated with larger effect sizes, $r(405) = .15$, $p < .01$. Thus it appears that the preponderance of group treatments may have attenuated the overall effect size obtained in the meta-analysis. Therapists typically had some 3 years' experience, $M = 2.91$; this average level represents advanced graduate students, rather than fully qualified and established professionals. However, the correlation between therapist experience and effect size was apparently negative, $r(269) = .14$, $p < .05$, although this association was probably due to the tendency of researchers to employ more experienced therapists when the target problems were less readily treated. The correlation was abolished by partialling out the effects of dummy variables representing target problems (Shapiro & Shapiro, 1982b). It is nonetheless of some interest that no positive association occurred with therapist experience, even after making allowance for the tractability of the treatment target. The typical duration of treatment, $M = 6.89$ hours, might appear unrepresentatively short, in comparison with clinical folklore. However, Garfield (1978) reports that traditional expectations in this regard are at variance with the empirical data, which indicate that most clinic clients remain in therapy for only a few interviews. On the other hand, there may be important differences between premature termination of a therapy expected to extend for several months and participation in the designedly brief therapy typical of our meta-analysis. The duration of therapy was uncorrelated with effect size, $r(394) = .05$, $p > .10$.

Control for Nonspecific Effects

Significantly, Cook and Campbell (1979, p. 60) cite the placebo effect and the introduction of the double-blind experimental design as their first example of a construct validity problem and the efforts of researchers to overcome it. Within the literature on psychological treatment, extensive discussion and research has focused on the comparative credibility of treatment and control conditions (Bootzin & Lick, 1979; Kazdin, 1979; Kazdin & Wilcoxon, 1976; Lick & Bootzin, 1975; Shapiro, 1981; Wilkins, 1979a, 1979b). Ideally, construct-validity considerations require the demonstration that the theoretical mechanisms underlying the differential efficacy of treatments correspond to those postulated by the treatment's rationale. Thus, all treatments under comparison, including control conditions, should be of demonstrably equal credibility to the client, so as to eliminate differential client expectation of benefit as the origin of outcome differences. Furthermore, the impact of investigator expectancies should be reduced via blind assessment procedures (client-change data obtained by personnel unaware of the treatment assignment of individual clients) and the use of measurement methods that are minimally reactive to the demand characteristics of the assessment situation. How well does the research reviewed in our meta-analysis meet these requirements?

Of the 143 studies reviewed, 70 (49%) included at least one control group receiving some form of minimal or placebo treatment. This leaves fully one half of the studies making no attempt to control for participation and expectancy effects, except insofar as these can be considered controlled for in the comparison between active treatments. The absolute (as distinct from relative) impact of treatment is only hazardously inferred from studies lacking such controls.

Kazdin and Wilcoxon (1976) have reviewed strategic issues in the design of experimental procedures to control for nonspecific treatment effects. These authors argue that the most rigorous procedure involves empirical demonstration that the control condition generates comparable client expectancies for change to those generated by the purportedly active treatment under investigation. Unfortunately, however, this empirical credibility matching was encountered in only four of our source studies, although several authors described the construction of placebo conditions with a view to enhancing credibility, without presenting data to demonstrate their success in this endeavour.

In the absence of empirical credibility matching, Kazdin and Wilcoxon advocate a *treatment-element control* strategy, in which the control procedure resembles the active treatment as closely as possible. The least preferred strategy reviewed by these authors is the *attention placebo* method, wherein any procedure is designated by the experimenter as a control for nonspecific treatment effects. In reviewing the minimal treatment conditions employed within our source studies, we found two additional classes: *Minimal-contact* conditions were confined to supplying information or instructions, usually with less therapist contact than in the active treatment groups; *counseling/discussion* control groups were considered separately from other placebo control groups where these were clearly designated as placebos by the source study author, in a design contrasting these with behavioral treatments and with no articulated theoretical rationale sufficient to warrant coding as an active, dynamic/humanistic method.

In Table 1, minimal treatments of the four types outlined above are listed separately, according to whether they figured as a treatment group for which effect sizes were obtained by comparison with another (usually no-contact) control group or as the control condition used in the calculation of effect sizes for that study. The table shows that there were only 15 instances of the treatment-element strategy favoured by Kazdin and Wilcoxon (1976); the most frequent strategy was the attention placebo (32 groups), of which the counseling/discussion control (12 groups) may be considered a variant. Table 1 shows that, considered as treatment groups, the four control strategies yielded similar effect sizes, with the exception of the minimal-contact groups, whose mean was inflated by one group obtaining a mean effect size of 3.22. To the right of Table 1 are shown the mean effect sizes obtained by all treatment groups in each study whose control condition for effect-size calculation fell in each of the four classes of minimal control group. The larger these means, the greater the advantage enjoyed by the treatment groups over the control condition of interest here. Once again, there are few differences, the apparently greater disadvantage of the counseling/discussion groups being due to one group against which a mean effect size of 4.68 was recorded. Overall, we conclude that the choice of control groups in the source studies was somewhat unsatisfactory. In terms of the discussion by Kazdin and Wilcoxon (1976), only a small minority of our source studies followed preferred strategies for eliminating the hypothesis that nonspecific effects may account for the differences obtained between treated and untreated clients. On the other hand, there is no evidence from these data that stronger controls for non-specific effects result in apparently weaker treatment effects, as would be expected were the apparent effects of treatment inflated by uncontrolled effects of this kind.

Outcome Assessment

Measures were administered blind in the case of 530 (53%) of the outcome assessments within the meta-analysis. Nonblind assessors, people who knew which group each client was in but did not themselves act as therapists, administered 183 measures (18%). We consider that such assessors are vulnerable to bias but to a somewhat less degree than the therapists themselves. The therapist treating the client administered 293 (29%) measures. Thus, over one half of our data are

Table 1
Minimal Treatments Within Meta-Analysis

	As treatments			As control conditions		
Type	*N* of groups	Effect size	*SD*	*N* of groups	Effect size	*SD*
Minimal contact	5	1.37	1.12	9	.61	.36
Attention placebo	19	.50	.67	13	1.04	.54
Counseling/discussion	8	.50	.48	4	1.59	2.06
Treatment element	6	.54	.58	9	.74	.83

vulnerable to criticism on this score, a fact whose importance is heightened by the significant correlation of our "blindness" measure with effect size, $r(1,004) = -.10$, $p < .01$, indicating larger measured effects of treatment with nonblind assessment. The vulnerability of the measures to demand characteristics was assessed via a modification of Smith et al.'s (1980) reactivity scale. Only 192 (11%) of the measures were minimally reactive (physiological measures). The majority (1,172 or 64%) of measures were coded as slightly reactive (blind behavioral ratings and standardized tests). Measures coded as highly reactive, comprising such measures as therapist ratings and simple self-reports, numbered 464. There was a significant tendency for more reactive measures to yield larger effect size, $r(1,826) = .11$, $p < .01$. Taken together with the data on blinding, these results indicate that demand characteristics are free to influence outcome in many studies and that they do indeed appear to inflate the estimates of effectiveness obtained to some degree.

In outcome research, construct-validity issues arise in relation to choosing dependent variables and putting them into operation. Our coding of the specificity of outcome measurement reflects the degree to which this directly taps the target problem under treatment (Agras, Kazdin, & Wilson, 1979, p. 68; projective measures = 1; standardized trait measures = 2; directly related psychological measures of physiological measures where the target is nonphysiological = 3; behavioral tests or physiological measures of physiological targets = 4). As shown in Table 2, the majority of outcome meaures were highly specific to the goals of treatment. The impact of specificity upon effect size was highly significant, $F(3, 1824) = 15.83$, $p < .001$. More powerful effects were associated with relatively specific measures, although the relationship was not linear. These results are generally encouraging, at least from the standpoint of the behaviorally oriented investigator. It appears that the overall impact of treatment in our meta-analysis is somewhat attenuated by the minority of measures that are not closely related to the target problem. On the other hand, personologists might point to the relatively modest effects of treatment upon standardized trait measures as a limitation on the depth and generalization of its impact.

Table 2
Specificity of Outcome Assessment and Effect Size

Specificity rating[a]	*N*	%	Mean effect size	*SD*
1	2	0	−.07	.03
2	294	16	.55	.67
3	1091	60	1.05	1.19
4	441	24	.88	1.29

[a] Defined in the text.

A further important aspect of assessment is its modality. Conventional wisdom favors multimodal assessment, the use of measures based on diverse technologies, which counters the threat to construct validity posed by reliance upon a single assessment modality. In our meta-analysis, these were coded into seven types (psychometric measures, self-ratings, direct behavioral measures, indirect behavioral ratings, "real-life" nonbehavioral data, physiological measures, and projective tests). As shown in Table 3, most studies confined themselves to only one or two of these modes of assessment, despite the mean 4.42 outcome measures obtained in each study. This limitation is somewhat unsatisfactory, although Table 3 shows that the number of assessment modes was not related to the mean effect size obtained for each group, $F(4, 409) = 1.33$, $p > .25$. These findings are subject to the minor qualification that some outcome measures obtained in the source studies were discarded because authors did not supply any data; but there is no reason to believe that such measures were any more diverse in respect to modality than those reported in full.

Table 3
Multimodal Assessment and Effect Size

N(modes)	*N*(studies)	*N*(groups)	Mean effect size	*SD*
1	64	177	1.02	.84
2	54	167	.89	.53
3	21	54	1.16	.88
4	3	12	.87	.41
5	1	4	.88	.11

Outcome research should demonstrate the maintenance of treatment effects over time. Of our data, 77% were obtained immediately posttreatment, and only 6% were obtained 4 or more months after the end of treatment. Although there was no relationship between extent of follow-up and the effect size obtained, $F(3, 1824) = 1.57$, $p > .10$, the mean follow-up of .79 months is clearly unsatisfactory, indicating that controlled comparative-outcome research of the kind reviewed in our meta-analysis has generally failed to address the issue of maintenance of therapy gains. In mitigation, however, it should be noted that a few longer term assessments were excluded from the meta-analysis because data from untreated controls were no longer available. Practical and ethical constraints militate here against the demonstration of longer term treatment effects, particularly in clinical populations.

External Validity

Cook and Campbell (1979) define external validity in relation to problems of generalizing to particular target persons, settings, and times, and of generalizing across types of persons, settings, and times. In the present context, the central issue is that of representativeness: To what extent do the persons and settings of comparative-outcome research represent those of clinical practice?

Concerning representativeness of samples, the practical constraints upon researchers can lead to sampling procedures that limit the generalizability of the findings. Only 67 (17%) of the treated groups comprised individuals who had sought the aid of a clinician or whose problems were of clinical severity. The participation of all but 42 (11%) of the groups was solicited by the investigators. Of the groups, 238 (61%) had a mean age of 20 years or less; the great majority were undergraduate students; 282 (82%) of the groups were university or college educated. Only 35 (6%) of the groups comprised individuals who had received a psychiatric diagnosis. Only 30 groups (7%) presented with anxiety and depression similar to the generalized neurotic problems commonly seen by clinicians; in contrast, 126 (30%) of the groups were treated for performance anxieties such as test and public-speaking anxiety and 106 (26%) were treated for physical and habit disorders. These data point to serious limitations upon the external validity of the studies reviewed.

Table 4 shows the correlations of client variables, aside from target problem, with effect size. Kazdin (1978) argues against the unquestioning assumption that analogue studies yield an inflated estimate of treatment efficacy. In respect to client characteristics, for example, he suggests that the "real" clinical client may be more highly motivated to achieve improvement, thus evincing more compliance and positive expectancy than students solicited to participate in return for course credit. Table 4 suggests little difference in outcomes between clinical and nonclinical populations, between solicited and other participants, or between young, highly educated (i.e. college student) samples and other participants. On the other hand, those few studies of patients receiving a formal psychiatric diagnosis yielded significantly lower effect sizes than obtained for the nondiagnosed participants involved in the great majority of studies reviewed. Furthermore, differences in effect size among clients with different target problems were highly significant, $F(29, 384) = 3.53$, $p < .001$, and accounted for over 20% of the variance in effect size, $\text{eta}^2 = .21$. As shown in Table 5, the largest effects were obtained for phobias, $M = 1.28$, and the smallest for anxiety and depression, $M = .67$, with intermediate results for physical and habit problems, sexual and social problems, and performance anxieties. Thus it does appear that the overall efficacy of treatments included in our review is inflated by the underrepresentation within the data of clinical populations suffering depression and anxiety,

Table 4
Correlations of Client Variables With Effect Size

Variable	*r*	*N*(of groups)
Severity	.04	393
Source (solicited vs. others)	.05	394
Age (years)	.04	393
Education (high vs. others)	.00	344
Diagnosis (present vs. absent)	−.09*	414

* $p < .05$.

Table 5
Target Classes and Effect Size

Target class	*N*	%	Mean effect size	*SD*
Anxiety and depression	30	7	.67	.62
Phobias	76	18	1.28	.88
Physical and habit problems	106	26	1.10	.85
Social and sexual problems	76	18	.95	.75
Performance anxieties	126	30	.80	.71

which are typically the most frequent recipients of psychotherapy in practice.

An important aspect of representativeness concerns the setting in which treatment is carried out and assessments of outcome are made. The 71% of our groups that were treated in laboratory settings yielded data that were unrepresentative in two respects: Not only was the treatment itself carried out in a setting unlike that of clinical practice but laboratory-based outcome assessments cannot be deemed representative of assessments made in clinical settings (Kazdin, 1978). Although nonlaboratory data yielded slightly smaller effect sizes, this difference was not significant, $t(388) = 1.43$, $p > .10$.

Reporting Adequacy

We note quite severe inadequacies of data presentation; for only 60% of the data were means and standard deviations presented within the source papers. A further 24% of effect sizes were calculated on the basis of supplied means together with error terms obtained from analysis of variance and related statistics; the remaining 16% were inferred from minimal data, such as probability values and sample sizes, or by probit transformation of dichotomous data (Glass et al., 1981). Thus, a substantial minority of these published treatment-outcome studies failed to report basic descriptive statistics.

In addition to the problems raised by nonreproducible treatment methods, the failure to monitor treatment interventions, and the lack of evidence concerning the credibility of treatment and control conditions, many data on treatment, client, contextual, and design variables were missing from the meta-analysis as a result of simple reporting omissions within the source papers. These lacunae severely limited the number of variables that could be included within multiple regression analyses and, thus, hampered our attempts to examine the relationships among study parameters in their impact upon effect size. Table 6 shows that few studies specified whether or not clients were concurrently on medication; the blindedness or otherwise of the data gatherer was specified in a bare majority of cases. Several of the remaining major instances of missing data concern therapists; this may reflect the behavioral orientation of many of the source-study authors and a consequent lack of concern with the impact of the therapists. Many important design variables were afflicted by substantial proportions of missing values. The data of Table 6 represent quite conservative estimates of the extent of source-study authors' underreporting. They exclude data inferred during coding (such as the age of college student samples). Furthermore, our selection of parameters for coding within the meta-analysis was based on prior acquaintance with reporting standards current within the literature, together with pilot coding exercises. Missing values therefore represent the failure of source-study authors to meet quite modest

Table 6
Missing Values

Variable	% missing
Treatment variables	
Mode	1.7
Therapist experience	36.5
Duration	4.3
Client variables	
Education	16.9
Age	5.1
Severity/screening	5.1
Source	4.8
Contextual variables	
Setting	6.0
Medication	88.4
Design variables	
Assignment of clients	6.3
Assignment of therapists	27.5
N of therapists	25.6
Attrition of treated group	18.9
Attrition of control group	18.7
Pretreatment equivalence	23.2
Blindedness of data gatherer	45.0

reporting standards. Furthermore, most of the missing data would require minimal journal space, so that space restrictions cannot account for them, although brief reports were often needlessly scanty and imprecise in these matters.

Conclusions

Our overall finding of an effect size approaching 1 standard deviation unit appears quite large, relative to findings in other domains of behavioral and social research (Cohen, 1977; Smith et al., 1980). But how good is the evidence upon which this estimate is based? The present review suggests that its statistical conclusion and internal validity are rather better than its construct and external validity. In respect to construct validity, the prevalence and somewhat inflated outcomes of reactive and nonblind outcome assessment, coupled with the general failure to control adequately for nonspecific treatment effects, give cause for concern. There is also insufficient multimodal outcome assessment. These deficiencies pale into insignificance, however, when set against the overwhelmingly unrepresentative nature of the data upon which our meta-analysis was performed. Typically, these studies concerned behavioral and cognitive therapies directed toward subclinical, focused target problems, such as performance and social anxieties, rather than generalized anxiety and depression. Data were usually obtained in laboratory settings, with a planned treatment duration of some 7 hours, with college student clients whose participation was solicited, and with the majority of the treatments conducted in groups by graduate student therapists. Some consolation may be derived from the generally small measured impact of these aspects of unrepresentativeness upon the effect sizes obtained. The fact remains that, on features relevant to all Kazdin's (1978) dimensions of generality, most of these data are quite unrepresentative of clinical practice, and the quantity of evidence with acceptable generality is vanishingly small. The virtual absence of follow-up data is another factor seriously limiting the clinical utility of the findings. In respect to our final concern in the present review, the reporting adequacy of the source studies is somewhat mixed; although many treatments were quite highly reproducible, little attempt was made to monitor the actual conduct of therapy, and there were several gratuitously missing data concerning the implementation of the research. A substantial minority of authors reported basic descriptive statistics inadequately.

Of course, the unrepresentativeness of our data might result from the relatively stringent design criterion (two or more active treatment groups plus a control) applied in selecting studies for consideration. For example, assignment to control conditions may be less feasible or considered ethically unacceptable with clinical populations, and longer term follow-up data on untreated controls are especially difficult to obtain. On the other hand, the less exacting design criteria of Smith et al. (1980), requiring only the comparison of two or more groups, did not result in much more representative data. Thus, it appears that contrast-design outcome research is indeed generally unrepresentative of clinical practice. This places severe limitations upon its implications for policy concerning service provision and training (Shapiro, 1980; Vandenbos, 1981). The methodological profile of this research revealed by meta-analysis highlights a pressing need for the redirection of the efforts of investigators toward realistic clinical studies.

Several commentators have striven to understand the modest progress of current therapy research toward the goal of identifying effective ingredients within clinical practice (Agras et al., 1979; Frank, 1979; Garfield, 1981; Horowitz, 1982; Rachman & Wilson, 1980). Despite the varying emphases of these authors, the following common themes are underlined by this meta-analysis:

1. All research design necessarily involves compromise between conflicting priorities and requirements, in the context of resource constraints (Waskow, Note 1). For example, the design requirements of internal and construct validity may often conflict, in the practical implementation of a study, with the requirements of external validity. As revealed by the present review, contrast-design outcome research has been marred by the recurrent tendency of investigators to make the

same compromises (e.g., sacrificing external and construct validity to internal validity). Our confidence in the data as a whole would be much greater had different investigators resolved their common dilemmas in more varied ways. Data overwhelmingly tending to share the same fault (such as the lack of effective control for nonspecific effects noted here) is more vulnerable to alternative explanations, in terms of the factors left uncontrolled by that fault, than data with more varied deficiencies requiring multiple alternative explanations for its interpretation to be challenged (Smith et al., 1980). Some of the blame for this uniformity of error must lie with sociocultural factors such as the reward and opportunity structures within which investigators conduct research (Agras et al., 1979).

2. It is costly of time and resources to conduct a clinically realistic contrast-design outcome study meeting current methodological requirements, and therefore research efforts must be diversified to include complementary research strategies (Agras et al., 1979; Horowitz, 1982). The analogue laboratory study of the type predominating within our meta-analysis should be viewed as but one such strategy. Other strategies sharing some claim to resolve issues of cause–effect relationships crucial to the identification of therapeutic ingredients include single-case experimental designs (Hayes, 1981; Hersen & Barlow, 1976) and experimental-process research, in which treatment techniques are systematically varied over the course of therapy and client changes in response to these are analyzed via both extensive and intensive designs (Shapiro & Shapiro, Note 2). Descriptive and correlational studies should also find a place within a methodologically diverse and evolving research effort, integrated via its sustained concern with a given scientific question (Agras et al., 1979).

3. The weaknesses of meta-analysis identified by critics, such as overgeneralization, indiscriminate inclusion of low-quality data, and idiosyncratic and unacceptable conclusions, are largely if not wholly inherent in the research endeavours reviewed by a meta-analysis (Shapiro & Shapiro, 1982a). The long-held and laudable aspiration of those behaviorally oriented investigators who have been most critical of meta-analysis (e.g., Eysenck, 1978; Rachman & Wilson, 1980) has been the development of a firm basis for clinical practice in replicated outcome studies. The establishment of such a research base depends upon generalizing across studies, and meta-analysis does this more systematically and sensitively than traditional literature reviews, especially where careful disaggregation is employed (Shapiro & Shapiro, 1982a; Strube & Hartmann, 1982). If the conclusions of such an analysis are uncongenial, this merely reflects what the literature in question can tell us. If the impact of therapy techniques upon outcome appears relatively modest, for example, then investigators should take such a finding seriously. They might consider methodological improvements, such as thorough efforts to control the ubiquitous nonspecific effects. Alternatively, they might pursue the substantive implications of the uncongenial result obtained. For example, experiments could be designed to identify and demystify the so-called "nonspecific" factors, together with other likely influences upon outcome, such as the client's readiness for change (Frank, 1979). If the image in the mirror displeases you, little is gained by throwing the mirror away.

Reference Notes

1. Waskow, I. E. *Presidential Address.* Society for Psychotherapy Research: Oxford, England, July, 1979.
2. Shapiro, D. A., & Shapiro, D. *Comparative therapy outcome research: Reflections on meta-analysis.* Paper presented at the London Conference of the British Psychological Society, December 1981.

References

Agras, W. S., Kazdin, A. E., & Wilson, G. T. *Behavior therapy: Toward an applied clinical science.* San Francisco: Freeman, 1979.

Arkin, R., Cooper, H., & Kolditz, T. A statistical review of the literature concerning the self-serving bias in interpersonal influence situations. *Journal of Personality,* 1980, *48,* 435–448.

Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review,* 1977, *84,* 191–215.

Bootzin, R. R., & Lick, J. R. Expectancies in therapy research: Interpretive artifact or mediating mechanism? *Journal of Consulting and Clinical Psychology,* 1979, *47,* 852–855.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.*

Chicago: Rand McNally, 1963. (Also published as *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.)

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.) New York: Academic Press, 1977.

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

Cook, T. D., & Leviton, L. C. Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality,* 1980, *48,* 449–472.

Cooper, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology,* 1979, *37,* 131–135.

Eysenck, H. J. An exercise in mega-silliness. *American Psychologist,* 1978, *33,* 517.

Frank, J. D. The present status of outcome studies. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 310–316.

Franks, C. M., & Wilson, G. T. *Annual review of behavior therapy: theory and practice* (Vol. 6). New York: Brunner/Mazel, 1978.

Garfield, S. L. Research on client variables in psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (2nd Ed.). New York: Wiley, 1978.

Garfield, S. L. Evaluating the psychotherapies. *Behavior Therapy,* 1981, *12,* 295–307.

Glass, C. V., McGaw, B., & Smith, M. L. *Meta-analysis in social research.* Beverley Hills, Calif.: Sage Publications, 1981.

Glass, G. V., & Smith, M. L. The effects of class size on achievement. *Journal of Education and Policy Studies,* 1979, *1,* 2–16.

Hayes, S. C. Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology,* 1981, *49,* 193–211.

Hersen, M., & Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press, 1976.

Horowitz, M. J. Strategic dilemmas and the socialization of psychotherapy researchers. *British Journal of Clinical Psychology,* 1982, *21,* 119–127.

Kazdin, A. E. Evaluating the generality of findings in analogue therapy research. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 673–686.

Kazdin, A. E. Nonspecific treatment factors in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 846–851.

Kazdin, A. E., & Wilcoxon, L. A. Systematic desensitization and non-specific treatment effects: a methodological evaluation. *Psychological Bulletin,* 1976, *83,* 729–758.

Kazdin, A. E., & Wilson, G. T. *Evaluation of behavior Therapy: issues, evidence and research strategies.* Cambridge, Mass.: Ballinger, 1978.

Kraemer, H. C. Coping strategies in psychiatric clinical research. *Journal of Consulting and Clinical Psychology,* 1981, *49,* 309–319.

Lick, J. R., & Bootzin, R. R. Expectancy factors in the treatment of fear: Methodological and theoretical issues. *Psychological Bulletin,* 1975, *82,* 917–931.

Luborsky, L., Singer, B., & Luborsky, L. Comparative studies of psychotherapies. *Archives of General Psychiatry,* 1975, *32,* 995–1008.

Mahoney, M. J. Experimental methods and outcome evaluation. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 660–672.

Rachman, S. J., & Hodgson, R. *Obsessions and compulsions.* Englewood Cliffs, N.J.: Prentice Hall, 1980.

Rachman, S. J., & Wilson, G. T. *The Effects of psychological therapy: Second enlarged edition.* New York: Pergamon Press, 1980.

Rosenthal, R., & Rubin, D. B. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences,* 1978, *3,* 377–415.

Shapiro, D. A. Science and psychotherapy: The state of the art. *British Journal of Medical Psychology,* 1980, *53,* 1–10.

Shapiro, D. A. Comparative credibility of treatment rationales: Three tests of expectancy theory. *British Journal of Clinical Psychology,* 1981, *20,* 111–122.

Shapiro, D. A., & Shapiro, D. Meta-analysis of comparative therapy outcome research: a Critical appraisal. *Behavioral Psychotherapy,* 1982, *10,* 4–25 (a).

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

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

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

Strube, M. J., & Hartmann, D. P. A critical appraisal of meta-analysis. *British Journal of Clinical Psychology,* 1982, *21,* 129–139.

Vandenbos, G. R. *Psychotherapy: Practice, research, policy.* Beverley Hills, Calif.: Sage Publications, 1980.

White, K. The relationship between socioeconomic status and academic underachievement. (Doctoral dissertation, University of Colorado, 1977). *Dissertation Abstracts International,* 1977, *37,* 5076A. (University Microfilms No. 77-03250).

Wilkins, W. Expectancies in therapy research: Discriminating among heterogeneous nonspecifics. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 837–845. (a)

Wilkins, W. Heterogeneous referents, indiscriminate language, and complementary research purposes. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 856–859. (b)

Received February 24, 1982
Revision received April 13, 1982 ■