Gene V. Glass, Reinhold M. Kliegl

An apology for research integration in the study of psychotherapy

first published in:
Journal of Consulting and Clinical Psychology 51 (1983) 1, S. 28-41,
ISSN 1939-2117, DOI 10.1037/0022-006X.51.1.28

Postprint published at the Institutional Repository of the Potsdam University:
In: Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 145
http://opus.kobv.de/ubp/volltexte/2009/4023/
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-40233

Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 145

An Apology for Research Integration in the Study of Psychotherapy

Gene V Glass and Reinhold M. Kliegl
University of Colorado

Criticisms of the integration of psychotherapy-outcome research performed by Smith, Glass, and Miller (1980) are reviewed and answered. An attempt is made to account for the conflicting points of view in this disagreement in terms of certain issues that have engaged philosophers of science in the 20th century. It is hoped that, in passing, something useful is learned about research of many types on psychotherapy.

The integration of psychotherapy-outcome studies that eventually led Smith and Glass (1977) to publication with Miller of *The Benefits of Psychotherapy* (1980) was born of dissatisfactions, enumerated below, with how outcome research was being pursued and used.

1. The chief occupation of the leading psychotherapy-outcome researchers seemed, at the time, to be quibbling over proper methodology; writers who commanded the most print were those most adept at writing about experimental design and statistics; and method became dogma and overshadowed substance.

2. The outcome literature had splintered and disintegrated; hundreds of unrelated efforts seemed to defy integration into anything that might address the question of the efficacy of competing therapies; indeed, it seemed to be believed that such questions could not or ought not be addressed and that incommensurability applied not only to theories of psychotherapy but also to its practice.

3. The methodology of outcome research reflected the worst features of positivist operationism: trivial quantification of outcomes (Behavioral Avoidance Tests and fear thermometers) devoid of technological importance; reliance on statistical hypothesis testing as *the* scientific method; disregard of the search for "function forms" (Meehl, 1978, p. 825) of practical and theoretical significance; belief on the part of psychotherapy researchers that they were engaged in the construction of grand theory about human behavior (instead of mapping a few "context-dependent stochastologicals" Meehl, 1978, pp. 812–3); and the tendency of researchers to ignore gross inconsistencies in findings from one laboratory to the next and to fail to draw the proper implications of such inconsistency for their field and its methodology.

4. The synthesis of psychotherapy-outcome research findings labored under the limitations of box-score (Light & Smith, 1971) counts of "statistically significant" results; it fell far short of extracting from the literature all that could be learned.

With this sense of the shortcomings of the field more vaguely felt than explicitly known (it was not until Meehl's, 1978, brilliant paper on "slow progress in soft psychology" that we began to see more clearly what seemed so futile about the course much outcome research was taking), Smith and Glass began their attempt to synthesize the huge research literature on psychotherapy outcomes. The integration they sought (a) would be based on the widest census of the outcome literature that could be accumulated; (b) would treat methodological rules as empirical generalizations whose value must be verified rather than as a priori dogma; and (c) would pit all major schools of psychotherapy against one another in a pragmatic contest where economic value and the concerns of public

We wish to acknowledge the helpful criticisms of Robert Cummins, Department of Philosophy, University of Colorado, who assisted in the early stages of thinking about this article.

Requests for reprints should be addressed to Gene V Glass, Campus Box 249, University of Colorado, Boulder, Colorado 80309.

policy would decide whether there was a winner. The first attempts (Glass, 1976; Smith & Glass, 1977) at accomplishing this task were incomplete and sketchily reported. In 1980, Smith, Glass, and Miller published a larger and more detailed analysis of the data under the title *The Benefits of Psychotherapy*. A few of the findings of that analysis are summarized here.

A total of 475 controlled evaluations of psychotherapy were found in the literature of journals, books, and dissertations. For each study, the average difference on the outcome measure between the therapy and control groups was divided by the within-control-group standard deviation among persons to form a measure of the magnitude of the effect (effect size) of the psychotherapy:

$$\Delta = \frac{M_{\text{therapy}} - M_{\text{control}}}{s_{\text{control}}}.$$

Measures on more than one outcome were frequently reported in a single study, or the same outcome might be measured immediately after therapy and at a follow-up time months later. Thus, there were many more effect-size measures than there were studies: in fact there were about 1,760 Δs from 475 experiments.

Most reviewers pursuing some synthesis using the methods of a bookkeeper or novelist are soon overwhelmed by the flood of information emanating from even a few studies. We respect the methods of statistics for their power to organize and extract information from numbers, and we feel they are not used enough in synthesizing research. A detailed coding of dozens of other characteristics of a study was also performed. A few of these characteristics appear in Table 1, an illustration of a study by Reardon and Tosi (Note 1). Each study, thus, was described by a multivariate data set amenable to any statistical analysis that might cast a little light on psychotherapy outcomes (their magnitude, their covariation with characteristics of therapies, therapists, clients, methods of research, and the like). The findings of these analyses are too numerous and complex to present here in any detail; indeed, they were too voluminous even for a 270-page book. But a few results will illustrate the approach.

The average of the effect-size measures was .85 (with a standard error of .03); that is, the difference in the means between groups receiving psychotherapy of any unspecified type for about 16 hours (the average duration of therapy) and untreated control groups was .85 standard deviation units. A .85 standard deviation effect can be understood more clearly by referring it to percentages of populations of persons. Assume that on a general measure of mental health, persons are distributed according to the normal distribution. If two separate distributions are drawn for those who receive psychotherapy and those who do not, our data lead us to expect (other considerations aside for the moment) that the two distributions will be separated by .85 standard deviations at their means. The average or median of the psychotherapy curve is located above 80% of the area under the control-group curve. This relationship indicates that the average person receiving psychotherapy is better off at the end of it than 80% of persons who do not. The average person, who would score at the 50th percentile of the untreated control population, would be expected to be at the 80th percentile of the control population after psychotherapy. Little evidence was found for the existence of the negative effects of psychotherapy. Only 9% of the effect-size measures were negative.

Table 2 reports the average effect sizes for each of 17 therapy types and placebo treatment, with the standard deviation of the effects, the standard errors of these averages, and the number of effect sizes. There exist many large differences in the size of effect produced by the therapies studied.

The highest average effect size, 2.38, was produced by cognitive therapies that go by such labels as systematic rational restructuring, rational state-directed therapy, cognitive rehearsal, and fixed-role therapy. Techniques used in these therapies include active persuasion and confrontation of dysfunctional ideas and beliefs. The second highest average effect size was 1.82 standard deviation units, produced by hypnotherapy. Cognitive-behavioral therapies such as modeling, self-reinforcement, covert sensitization, self-control desensitization, and behavioral rehearsal were third highest on the ranking of therapeutic

Table 1
Classification of Reardon and Tosi (1976)

Publication date	1976
Publication form	Journal (although the paper was available in unpublished form, both as a dissertation and paper read at a professional meeting, it had been accepted for journal publication)
Training of experimenter	Psychology (inferred from department affiliation)
Blinding	Experimenter was therapist (judged from report)
Diagnosis	High-stress delinquents (experimenter's description), delinquent or felon
Hospitalization	None
Intelligence	Average (estimated, in the absence of other information)
Client–therapist similarity	Moderately dissimilar (because of students' identification as delinquent, age and sex differences)
Age	16 (stated)
Percent male	0% (female population)
Solicitation	Self-presentation in response to advertised services (stated)
Assignment of clients	Random (stated)
Assignment of therapists	Random (stated)
Experimental mortality	0% from all groups
Internal validity	High
Simultaneous comparison	Treatment and placebo compared against control
Type of treatment	(1) "Rational-stage directed imagery" (subclass-cognitive; hypnosis or intensive muscle relaxation used as aids to induce rational, cognitive restructuring) (2) Placebo-relaxation and suggestion to feel better
Confidence of classification	Rated 4 (many key concepts associated with Ellis, Kelley, and Raimy, plus personal communication with the authors)
Allegiance	Definite allegiance toward the therapy (inferred from tone of report)
Modality	Group (stated)
Location	Residential facility (stated)
Duration	6 hours over 6 weeks (stated)
Therapist experience	3 years (inferred from status as doctoral candidates)
Outcome	Two outcomes were measured by experimenter: Tennessee Self-Concept Scale (TSCS) and the Multiple Affect Adjective Checklist (MAACL; not included because of insufficient data reporting). TSCS rated as a 4 in reactivity (self-report on measure similar to treatment). Measure was taken immediately after therapy and 2 weeks later.
Effect size	Means from TSCS (total) were obtained from a figure. Estimates of standard deviation were made using probability values. For the rational-stage directed imagery treatment, effect sizes were 1.59 for 0 weeks post and 1.59 for 2 weeks post; for the placebo treatment, effect sizes were .74 for 0 weeks post and −.20 for 2 weeks post.

effectiveness ($\bar{\Delta}$. = 1.13). Systematic desensitization, primarily used to alleviate phobias, was next highest ($\bar{\Delta}$. = 1.05). An average effect size of .89 standard deviation units was achieved by dynamic-eclectic and eclectic-behavioral therapies. Several therapy types

Table 2
Average, Standard Deviation, Standard Error, and Number of Effects for Each Therapy Type

Type of therapy	Average effect size, $\bar{\Delta}$.	*SD*	SE_M	No. of effects (*n*)[a]
1. Other cognitive therapies	2.38	2.05	.27	57
2. Hypnotherapy	1.82	1.15	.26	19
3. Cognitive-behavioral therapy	1.13	.83	.07	127
4. Systematic desensitization	1.05	1.58	.08	373
5. Dynamic-eclectic therapy	.89	.86	.08	103
6. Eclectic-behavioral therapy	.89	.75	.12	37
7. Behavior modification	.73	.67	.05	201
8. Psychodynamic therapy	.69	.50	.05	108
9. Rational-emotive therapy	.68	.54	.08	50
10. Implosion	.68	.70	.09	60
11. Transactional analysis	.67	.91	.17	28
12. Vocational-personal development	.65	.58	.08	59
13. Gestalt therapy	.64	.91	.11	68
14. Client-centered therapy	.62	.87	.07	150
15. Adlerian therapy	.62	.68	.18	15
16. Placebo treatment	.56	.77	.05	200
17. Undifferentiated counseling	.28	.55	.06	97
18. Reality therapy	.14	.38	.13	9
Total	.85	1.25	.03	1,761

[a] The number of effects, not the number of studies; 475 studies produced 1,761 effects, or about 3.7 effects per study.

yielded average effect sizes clustering around two thirds of a standard deviation; none was significantly different from the others: psychodynamic therapy, client-centered therapy, Gestalt therapy, rational-emotive therapy, transactional analysis, implosive therapy, behavioral modification, and vocational-personal development counseling.

Placebo treatments (e.g., relaxation training, pseudodesensitization therapy, minimum-contact attention control groups) produced an average effect size in comparisons with untreated control groups of .56 standard deviation units. Any unadjusted comparison of this effect with the effects of the therapies would be invidious. Placebos were often used in studies of the treatment of monosymptomatic anxieties. *In such studies*, the effect of the psychotherapy was about twice as great as the effect of the placebo.

Two types of therapy gave noticeably small effects. Undifferentiated counseling (defined as therapy reported without descriptive information or references to a theory) had an average effect size of .28. The smallest effect size in the table is for reality therapy ($\bar{\Delta}$. = .14). However, only one controlled evaluation of this therapy was found. All nine effect sizes were produced by this one study.

The therapy effects of Table 2 are controlled in the sense that therapies were compared with control groups within studies. The comparison of average effect sizes among types of therapies in Table 2 does not control for the interaction of therapy effectiveness with other variables. Persons seeking psychotherapy help do not randomly assign themselves to the different types. Some types of therapy are specifically designed for a narrow range of psychological problems. Certain therapies appeal to less seriously disturbed clients. Hence, the differences in therapeutic effect reported in Table 2 reflect variation in therapeutic effectiveness plus its interaction with client characteristics, diagnostic types, therapist experience, choice of outcome criteria, and the like. This is not to say that the comparisons in Table 2 are without value to the questions raised about psychotherapy by laymen, policy makers, and even professionals. However, more questions (often those raised by researchers) can be answered through control of some of the variation represented in Table 2.

About 50 of the studies identified for the meta-analysis involved comparison of two or more therapies directly. These studies were isolated for closer examination in the "same

experiment" analyses. Because of the small number of studies involved, the therapies were aggregated into *classes*. In 56 experiments, the outcome of verbal psychotherapies (psychodynamic, client-centered, rational-emotive, etc.) were directly compared with those of behavioral psychotherapies (systematic desensitization, behavioral modification, etc.). The experiments yielded 365 effect-size measures. Unlike the confounded comparisons one can make in Table 2, the 365 "same experiment" comparisons are equated in all relevant respects on the verbal and behavioral sides of the ledger. The findings at the most general level appear in Table 3. This difference (.19 sigma units) between verbal and behavioral therapies struck us as being quite small, and our saying so appears to have offended those who seem to believe that the psychotherapy Olympics were long since over and the laurels were theirs (e.g., Rachman & Wilson, 1980), even as it pleased those (Freudians and Rogerians) who once believed the race had been lost.

The longevity of psychotherapy effects was studied by comparing the sizes of effects measured at various follow-up dates. Two thirds of the 1,700 therapy effects were measured immediately after therapy, that is, at "0 weeks" follow-up time. Other common follow-up dates were 4 weeks (11% of the Δs), 8 weeks (4%), and 52 weeks (2.5%) after therapy. Four effects were measured more than 10 years after therapy. Effects were averaged in 19 follow-up-date categories extending from immediately after therapy to more than 300 weeks after. The average effect was regressed onto a quadratic function of time (Weeks plus Weeks Squared). The resulting least squares solution produced a multiple correlation of .78. The graph in Figure 1 depicts the relationship. The estimated average

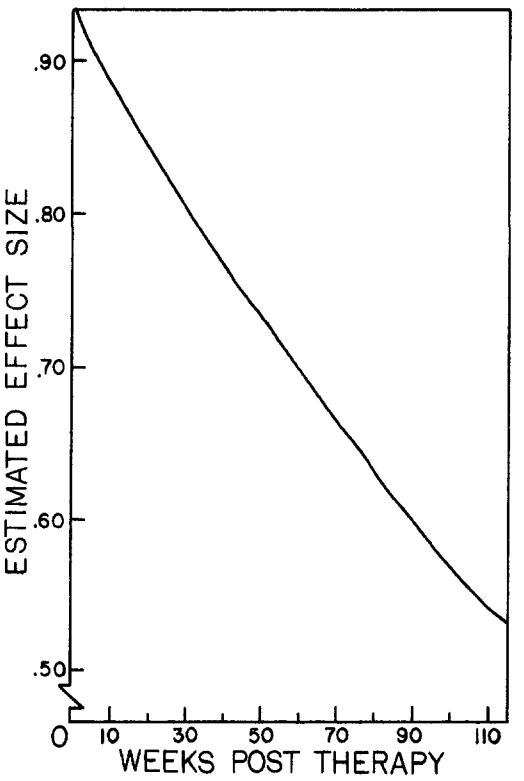

Figure 1. Relationship between measured effect of psychotherapy and the number of weeks after therapy at which the effect was measured.

effect of psychotherapy is slightly above .90 standard deviation units (therapy vs. control) immediately after therapy, and it falls to around .50 at about 2 years (104 weeks) after therapy.

The duration of therapy was recorded for each study as the number of hours of psychotherapy received by the clients. Durations of therapy ranged from 1 hour to over 300 hours. Over two thirds of all effects were measured in studies involving 12 or fewer hours of treatment; the mean duration of therapy was 15.7 hours. The effect of therapy bore a complex relationship to its duration. Figure 2 is a graph of the curve relating effect size to therapy duration. The curve has been smoothed by a fifth-order moving average. The curvilinear correlation between effect size and hours of therapy equals .29. The remarkable thing about the curve in Figure 2 is that in its most peculiar feature (viz., the dip in effectiveness for therapies of duration

Table 3
Findings of the "Same Experiment" Analysis

Findings	Class of psychotherapy	
	Verbal	Behavioral
Average effect size: $\bar{\Delta}$.	.77	.96
SD of effects: σ_Δ	.76	.87
No. of effects	187	178
No. of studies	56	56

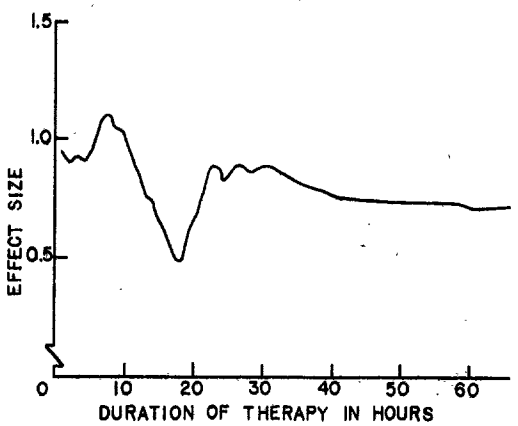

Figure 2. Relationship of effect size and duration of therapy. (Curve is derived from a fifth-order moving average of 1,760 effect-size measures.)

between 10 and 20 hours) it resembles the relationship between duration and gain that Cartwright (1955) observed for 78 subjects in client-centered therapy and which Taylor (1956) later documented in a separate study.

Some empirical findings about how psychotherapy-outcome studies are done and the relationship between methods and study results were as interesting as the findings concerning therapy efficacy themselves. For example, outcome studies differ markedly in the methods of measuring therapeutic gains. At one end of the continuum lie measurement techniques patently subject to bias and distortion through personal influence, such as unblinded therapist ratings or clients' behavior (e.g., touching rats) enacted in the presence of the experimenter. At the other end lie outcome measures little subject to self-serving pressure by the experimenter or the clients' desire to reward the therapist by appearing to have benefited from treatment (e.g., grade-point average; galvanic skin response). When the 1,760 effect sizes were sorted into five categories on this "reactivity continuum," the average effect sizes in Table 4 resulted.

The findings of Table 4 were surprising and sobering. Reactivity of method of measurement proved to be the highest single correlate of effect size that Smith et al. observed. Are the effects of psychotherapy in the published literature puffed up by a good bit of self-congratulation and self-deception? Smith et al. (1980) were provoked to the observation that

> The measurement of outcomes seems to have been abandoned at a primitive stage in its development. Rating scales are thrown together with little concern expressed for their psychometric properties. Venerable paper-and-pencil tests invented for diagnosis and with roots planted vaguely in no particular theory of pathology or treatment are used to hunt for effects of short-term and highly specialized brands of psychotherapy. A superfluity of instruments exists, and too little is known about them to prefer one to another. (p. 187)

One worries that some outcome experiments were designed purposely to make one type of psychotherapy look good and a second type look bad. One way of getting at such a motive is to classify each therapy effect size by whether the experimenter felt allegiance (expressed or clearly inferred) to the therapy or felt allegiance to a different therapy competing with it in a comparative study. The average effect sizes under these two circumstances plus a third category of unknown allegiance are reported in Table 5.

The difference in average effect size between therapies favored by the experimenter and those not favored is 50% (.66 vs. .95). It might be argued that the efficacy of particular

Table 4
Average Effect Size Classified by Each Value of Reactivity of Measurement

Reactivity scale value	Examples of "instruments"	Average effect size	SE_M	No. of effects
1 (Low)	Galvanic skin response, grade-point average,	.55	.06	222
2 (Low average)	Blind ratings of adjustment	.55	.04	219
3 (Average)	Minnesota Multiphasic Personality Inventory	.60	.04	213
4 (High average)	Client self-report to therapist, E-constructed questionnaire	.92	.03	704
5 (High)	Therapist ratings, behavior in presence of therapist	1.19	.06	397

Note. Average reactivity for all cases: 3.46. Correlation of reactivity and effect size: linear, $r = .18$; curvilinear, $\eta = .28$.

Table 5
Average, Standard Deviation, Standard Error, and Number of Effect Sizes for Experimenter Allegiance

Experimenter allegiance	Average effect size, $\bar{\Delta}$.	*SD*	SE_M	No. of effects
Allegiance to the therapy	.95	1.46	.04	1,071
Allegiance against therapy	.66	.77	.04	479
Unknown or balanced allegiance	.78	.86	.06	213

therapies won them their followers; it could be counterargued that therapies of all types have their adherents.

Finally, studies were classified as high, medium, or low in "internal validity" (a combination of random assignment, differential mortality patterns, regression, and other principles of the Campbell–Stanley type). The variation in study findings related to experimental internal validity appear in Table 6.

No reliable differences in magnitude of effect can be accounted for by differences in design validity. There was a slight positive relationship between effect size and internal validity, showing that the better designed studies produced larger effects. This trend is opposite of that implied by some critics who maintain that only poorly designed studies show psychotherapy to be effective.

The analyses of the data set are continuing. Researchers with interests in special topics have acquired it and are performing more refined analyses. At some point in the future, a new generation of studies may be added and the analysis may continue.

Responses to the Meta-Analysis

The publication of the meta-analysis of psychotherapy outcome research attracted praise (Abeles, 1981; Frances & Clarkin, 1981; Simon, 1981) and condemnation (Crown, 1981; Eysenck, 1978; Kazrin, Durac, & Agteros, 1979; Presby, 1978; Rachman & Wilson, 1980). That this one study could have prompted such different evaluation (Eysenck—"an exercise in mega-silliness"; Simon—"Classics in science are rare, but I predict this volume owns a distinguished destiny.") suggests not that one reviewer is perspicacious while the other is unaccountably blind, but that the work can be viewed simultaneously as more than one simple thing (unlike, say, an experiment of the effects of glyoxylate on CO_2 fixation in photosynthesis). Because the Smith et al. study resembled things that some psychotherapy researchers are traditionally interested in, they naturally found it difficult to evaluate in other than traditional terms. They might usefully have pointed out that the Smith et al. study failed to solve a problem in which they were interested, but that would have been quite a different matter from saying that it failed to solve *any* problem. In a bad-tempered sort of way, biologists may feel that sociology is a waste of good grant money, or sociologists may scorn history as fradulent gossip; but such provincial grumbling is not serious criticism, especially if those who speak it have failed to recognize that different inquires have been deliberately shaped by different purposes. Most misapprehensions of the purposes of the Smith et al. study are due, perhaps, to the tendency of some critics to assume, in the absence of a clear statement of rationale, that the purpose of the study was similar to their own purposes or those of psychotherapy-outcome studies. We hope to correct these misunderstandings here and provide the belated rationale.

Table 6
Average, Standard Deviation, Standard Error of the Mean, and Number of Effect Sizes Classified by the Internal Validity of the Study

Rated internal validity	Average effect size, $\bar{\Delta}$.	*SD*	SE_M	No. of effects
1 (Low)	.78	.80	.05	224
2 (Medium)	.78	.83	.04	378
3 (High)	.88	1.42	.04	1,157

In the *Benefits of Psychotherapy*, Smith et al. took as object field and explananda the *literature (i.e., printed documentation) of psy-*

chotherapy-outcome research, the methods of study employed by researchers, and the use of this literature and methodology by professionals, researchers, laymen and policy-makers. The choices of concern are pragmatic judgments, value judgments; and the authors took some pains to distinguish the desired end products of their inquiry from those that a psychotherapy researcher might falsely assume them to be (Smith et al., 1980, pp. 24–27). The literature on psychotherapy outcomes is distinct from psychotherapy outcomes themselves by a sequence of translations too obvious to enumerate. The importance of this literature is attested to by the effort that psychotherapy researchers invest in writing it, refereeing it, reading it, and reviewing it. The inquiry that takes as primary objects the documented outcomes of psychotherapy and the methods of psychotherapy-outcome researchers and that takes as explananda the policy decisions that can be justified by the literature of outcome research will necessarily use taxonomies and methods different from the inquiry that takes as object field the interaction of psychotherapist and client and as explananda the outcomes of this interaction. The distinction can be made clearer by example: It would have been meaningless for Smith et al. to have spoken of a published study as being "anxious" or "depressed," whereas the attribution would be perfectly appropriate made by a psychotherapy researcher about a client in an outcome experiment. The construction that we imposed on the psychotherapy-outcome research literature was chosen for a specific purpose, that is, to determine how and in what ways the judgments, decisions, and inclinations of persons (scholars, citizens, officials, administrators, policy makers) ought to be influenced by the literature of empirical research on the benefits of psychotherapy. This purpose was chosen in part as a remedy for an habitual inability of psychotherapy researchers to rise above partisan squabbles and theoretical hot-dogging when attempting to inform policy makers. We join Meehl in the opinion that "most so-called 'theories' in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless" (Meehl, 1978, p. 806). Hence, we could not have shared many of the concerns that motivated the authors of the primary studies we integrated, their concerns being primarily the construction of big theory about changing behavior. A naive form of rationalism believes that the best policy is synonymous with the best science. One would have thought that Oakeshott (1962) would have put this belief forever to rest. Too many psychotherapy researchers seem to believe that "rational" policy is impossible until encompassing theories are discovered and confirmed.

An Apologia for Smith et al.

Empirical inquiry of all sorts shares a basic structure. One can distinguish (a) the selection and definition of an *object field* (the events or things that are to be explained, understood, predicted or whatever), (b) the construction of a *taxonomy* (the definition of constructs, words and symbols, "slicing up the raw behavioral flux," as Meehl, 1978, p. 808, put it), and (c) the development of a *methodology* (techniques of measurement, analysis of information, collection of evidence, and the like). The articulation and refinement of these three elements defines a particular science. It is a regretable egocentric failing of many scientists that they are unable to reflect self-consciously on the historical choices that have bequeathed to them their particular "science," but instead believe that logic demands they pursue their inquiries precisely as they are pursuing them. Scientists readily acknowledge that the selection of an object field is a choice that is part arbitrary and part historical circumstance. There are fewer scientists, however, who believe the same for the choice of a methodology. A substantial part of meta-analysis is concerned with the investigation of methodological assumptions (which are genuinely refutable conjectures, Meehl, 1978, p. 810, and not axioms at all). Thus, meta-analysis shakes up a scientist's faith in his method. Habermas (1971) argued convincingly that the knowledge-constitutive interests that determine, in part, the selection of a certain methodology for science can be derived from the structure and pragmatic needs of the society in which the science exists. For example, experimental

design of the Fisherian sort, the specification of independent and dependent variables, the identification of cause–effect relationships, and the criterion of replicability, are principles growing out of the wish for technological control, whether it be control of pi-mesons, doorstops, machines, or human beings. Within a technological society, these methodologies are seen as paying greater dividends than, for example, passive observation, thought experiments, and Verstehen (empathic understanding). The notion that "logic" itself has led a particular group of scientists to a "proper" methodology has been severely and justly criticized (in particular Feyerabend, 1978; Kuhn, 1962). It can not be logically proved that biology is a science and archaeology is not. Voodoo may have as much potential for enriching our knowledge of disease as does the paradigm of Western medicine we have inherited. The selection of a particular methodology can not guarantee the success of a research program.

The misapprehensions certain critics have held concerning the Smith et al. study and many of their specific criticisms stem, we believe, from the critics' failure to discern the rationale we used in adopting object field, taxonomy, and methodology or from their insistence that Smith et al.'s choices in regard to these three elements were illegitimate. Clearly, a scientist's choice of a taxonomy or a methodology is subject to criticism and evaluation (e.g., the taxonomy of astrology is bloated, inconsistent, and contradictory, or the playing-card-anticipation technique of parapsychology is sometimes poorly controlled against visual clues). Even one's choice of object field and explananda are never above reproach, as anyone proposing to the National Institute of Mental Health a study of the influence of distant planets on marital accord would quickly learn. The belief persists among some psychologists that the proper pursuit of psychological science demands attention to a particular object field (e.g., experimental intervention and not naturalistic observation). That belief is arguable.

These remarks on the roots of methodology (viz., that they lie in pragmatic human interests, not in logic) were made to help clarify why meta-analysis takes methodologies as part of its object field. Meta-analysis tests and casts doubt on unwarranted methodological principles. It tests the rationality of holding beliefs about various methodological principles, while subscribing to some of the same methodological principles it tests. In this sense, it breaks no new ground. What prevents the endeavor from being uselessly circular or self-refuting is that methodological principles never exist in abstraction from an object field. Consequently, it is possible to select as an object field the usefulness of methodological principles applied to a different object field (e.g., psychotherapy). Indeed it is done all the time; it is merely disconcerting to some to see it done empirically rather than as a matter of mathematics or logic prior to empiricism.

From the many complaints and crotchets that critics have registered against the Smith et al. study, three general concerns emerge as serious and troubling. Each, we maintain, ceases to be a forceful criticism of the work when viewed in the context elaborated upon in the above remarks. The three criticisms are the quality of study problem, the uniformity problem, and the incommensurability problem. Glass, with Smith and McGaw, discussed each of these at some length in *Meta-Analysis in Social Research* (1981, pp. 217–26) and with specific reference to psychotherapy-outcome research in *Benefits of Psychotherapy* (Smith et al., 1980, pp. 27–35). The frequency with which independent critics raise these issues and the rigidity with which the positions are held in the face of counterarguments suggest that the misconceptions are deeply rooted in the critics' conceptual systems, which include the nature of science itself.

The Quality-of-Study Problem

The dogged rejection of the psychotherapy-outcome meta-analysis appears to be rooted in part in the guiding principles that a particular scientific community has adopted. The unifying principles of this community are primarily methodological as opposed to substantive. Uniformity or consistency, the hallmark of this view, is achieved primarily by insuring potential intersubjective testability of reported results. The acceptance of this principle is a priori to any research. The be-

lief or the conviction is held above question that there is a way of doing "correct" and "good" scientific work. To be sure, there are and always will be arguments about flawed designs, inappropriate controls, invalid measurement, and the like. Disagreements about findings are settled (if indeed any attempt at all is made to settle them) by replications of the object studies with alternative designs and measures. Disagreements about the adequacy of a methodological assumption are settled by giving an alternative assumption a chance in replication. Meta-analysis treats methodological assumptions of object studies as part of an object field in itself, that is, as a posterioris. This different point of view permits one to suspend judgment about studies that others might regard as "a priori bad." Rachman and Wilson (1980, p. 253), for example, impute to others doubts about the entire dissertation literature, while professing broadmindedness on their own behalf. Indeed, some reviewers of psychotherapy research ignore all dissertation reports on the grounds that they are undependable, a judgment that is surely a matter of empirical test rather than a priori conjecture. If design "flaws" are crucial, they will show a correlation with study findings expressed as effect sizes. The weaknesses of method need not be judged a priori.

Viewing the Smith et al. meta-analysis this way, the argument "garbage-in—garbage-out" (Eysenck, 1978) is not only trite but beside the point. Such a judgment reveals that its author can not conceive of treating design properties as something that can be empirically studied rather than merely debated. In this respect, Eysenck's (1978) claim that Smith and Glass advocated "low standards" for research quality and "abandoned scholarship" can be understood as the opinion of one to whom methodology is dogma.

> I would suggest that there is no single study in existence which does not show serious weaknesses, and until these are overcome I must regretfully restate my conclusion of 1952, namely that there still is no acceptable evidence for the efficacy of psychotherapy (Eysenck, 1978, p. 517).

It is relevant that Eysenck (1981) continues to dispute evidence on the causal link between smoking and lung cancer and stakes his argument on narrow methodological grounds long after reasonable people have acknowledged that a less-than-perfect study (U.S. Public Health Service, 1963) nonetheless produced an eminently credible conclusion.

The empirical status of methodological principles can only be studied where there are sufficient studies under various methodological circumstances to permit estimation of the relationship between the principle and study findings. Such principles cannot be studied in a single object study (focused on psychotherapeutic processes) for the same reason that it would be impossible to draw strong theoretical conclusions about human behavior on the basis of experience with one person. Hence, methods remain a priori for object studies, and one must respect them. However, at the level of meta-analysis, the necessity and justification of the methodological principles of the object study become the point of concern. One can be grateful for some less-than-perfect designs since the relationship of the methodological principles with the study findings cannot be studied unless the principles are satisfied to varying degrees. Rather than "garbage-in—garbage-out," meta-analysis examines that which is garbage when judged by a priori standards. "Garbage-in—information out" might be nearer the truth.

The a priori considerations of a meta-analysis are different from those entering the design of a therapeutic processes object study. However, meta-analysis subscribes to the same methodological dogmas as does psychotherapy research at the object level (measurement validity, control of extraneous influences either by statistical correction or experimental arrangements, and the like). A number of criticisms have been directed at Smith et al.'s work on this level. For example, an integrative analysis ought to be representative; relevant studies should not be overlooked. Smith et al. were said to have missed some studies that Rachman and Wilson (1980, pp. 251–2) regarded as "major, well-controlled investigations." Moreover, they argued that these oversights disadvantaged studies of behavioral therapy. Such a claim is fully appropriate and bears checking. Since meta-analysis takes as object field a static and tangible body of documents, it is often simple to include or exclude certain studies or characteristics of studies and perform new anal-

yses, somewhat in the manner of astronomy, where the same data are subjected to many analyses. The data base constructed by Smith et al. has already been borrowed, expanded, trimmed, or recoded by a dozen different researchers seeking to answer new questions or old questions with better methods (for examples, see Andrews & Harvey, 1981; Shapiro & Shapiro, 1982).

The Uniformity Problem

The assertion was repeatedly made that Smith et al.'s work was invalidated by their mistaken belief in some sort of "myth of uniformity," that is, that they believed persons, therapists, therapies, and pathologies to be somehow all the same and not worth distinguishing among.

> Smith and Glass subscribe to what Kiesler (1966) called the "uniformity assumption myths" of psychotherapy research and evaluation. Nowhere is this more damaging than with respect to measures of therapy outcome. . . .
>
> Other objections to the Smith and Glass meta-analysis involve the confusing mixture of patients and problems. No attempt is made to distinguish between the effect of this medley of treatments on schizophrenics, or alcoholics or adolescent offenders, or under-achieving college students, or phobic psychiatric patients, or subnormal patients, or patients suffering from migrane, or from asthma. . . .[1] Evidently we have travelled a great distance from Paul's (1967) recommendation that we evaluate the effectiveness of the particular technique for the particular problem of the particular person, all the way to a spreading sludge of diverse problems. (Rachman & Wilson, 1980, pp. 253–254)

Or consider Bandura's (1978) criticisms (delivered with nearly the same dip in the inkwell as his condemnation of Smith and Glass, 1977, for having mixed apples and oranges) of what we regard to be the finest single evaluation of psychotherapy outcomes in the literature:

> A widely publicized study by Sloane, Staples, Cristol, Yorkston, and Whipple (1975) comparing the relative efficacy of behavioral therapy and psychotherapy, similarly contains the usual share of confounded variables, unmatched mixtures of dysfunctions, and inadequately measured outcomes relying on amorphous clinical ratings rather than on direct assessment of behavioral functioning. As is now predictable for studies of this type, the different forms of treatment appear comparable and better than nothing on some of the global ratings but not on others. With such quasi-outcome measures even the controls, who receive no therapeutic ministrations, achieve impressive improvement. Based on this level of research, weak modes of treatment are given a new lease on life for those who continue to stand steadfastly by them. (p. 87)

Criticisms of these types reflect a fundamental misunderstandng of scientific inquiry. In a scientific inquiry, things that will be distinguished and things whose distinctness will be ignored are embodied in the choice of object field and taxonomy. To a physicist, a lamina of which the center of gravity is sought may perfectly well be assumed uniform in density; an engineer testing materials for breaking strength would surely not make the same assumption. Organisms assumed to be uniform in some sense by a physiological psychologist would not necessarily be so regarded by a histologist. The elaboration of all nonuniformity is an endless task that no sensible scientist ever attempts; knowledge demands simplification. The criticism that a particular inquiry fails to draw a set of critical distinctions can only be defended in the context of the object field and explananda being investigated.

Smith et al. have not the slightest belief in uniformity; if they had, they would never have employed statistics, that prima facie proof of disbelief in uniformity (likewise used, incidentally, by everyone who criticized Smith et al. for believing in uniformity). The irony of the criticism is twofold. We would have liked to have drawn many more distinctions among therapies, instances of a particular therapy, clients, therapists, and the like than we did. Unfortunately, such distinctions are largely impossible in an investigation of the documented literature of psychotherapy outcomes because the distinctions are not recorded by psychotherapy researchers in the primary studies. And they are not drawn there, we claim, because of the naive beliefs of logical empiricists that understanding of psychotherapeutic processes can be communicated in the operationist shorthand that defines the contemporary culture of research journals. Furthermore, Smith et al.'s belief in *non*uniformity of human thought and action far surpasses that of their critics (Glass, 1979). Many psychotherapy researchers appear to believe that the intransigent variabil-

[1] Results are reported separately for categories such as these in Smith et al. (1980).

ity and unpredictable quality of human behavior will be tamed by a multifactor analysis of variance design leading to simple generalizations of the form *this* type of therapy with *this* type of client will yield *this* type of outcome (Kiesler, 1966; Paul, 1967). Or they seem to believe that consistency and uniformity will emerge once we undertake the "direct assessment of behavioral functioning." As regards all contemporary theories of human behavior, they strike us as naively simple and grossly uniform. In our defense we cite Cronbach (1975), Bowers (1973), and Gergen (1973).

The Incommensurability Problem

Research integration is faced with the task of comparing studies that might have differed in their authors' intentions in respect to object field, taxonomy, and methodology. Dissimilar intentions would seem to imply a fundamental incomparability, but it does not. This aspect of meta-analysis probably gives rises to the angriest criticisms. Rachman and Wilson pounced on Cooper's (1979) stipulation of the conditions that must be met for a meta-analysis to make sense: The research studies to be integrated must "a) share a common conceptual hypothesis or b) . . . share operations for the realization of the independent or dependent variables, regardless of conceptual focus," (p. 133). Rachman and Wilson took Cooper's dicta as proof that Smith and Glass's (1977) integration of outcome studies was illogical and useless—that it failed even to meet meta-analysis's own requirements. Rachman and Wilson's use of Cooper against Glass and Smith is ironic and mischievous. Having coined the term "meta-analysis" (Glass, 1976) and first applied it to the integration of psychotherapy-outcome research (Smith & Glass, 1977), Glass and Smith might be justified in claiming some proprietary rights for deciding what it means when *they* use it. Of course, there are no proprietary rights to ideas in science; but observers will see inconsistencies where they do not truly exist if they fail to distinguish what one person judges proper research synthesis to be from what someone else thinks.

In the framework of science identified in this article, Cooper would set as a condition of comparability the identity of object field, taxonomy, and methodology. Thus, Cooper's stipulation is clearly inappropriate for the purposes of meta-analysis; indeed, if taken seriously, it would virtually preclude any meta-analysis. The more common criticism of meta-analysis (viz., that it errs in mixing "apples and oranges") is best understood by distinguishing *theoretical* from *practical* commensurability. The former is a long-standing point of debate in the philosophy of science, and the best that can be said of progress toward the solution of the problem is that there has been little. It occupied Kuhn (1962) in *The Structure of Scientific Revolution*, who despaired of simple answers; and it will undoubtedly be some time before a popular position emerges. Practical commensurability, on the other hand, is trivial by comparison. Conceptually, it poses problems in the philosophy of values, where some of the most productive thinking in philosophy has been pursued in recent decades. Fortunately, practical commensurability depends in no way on theoretical commensurability.

No matter what Smith et al. might have found, their conclusions would not have provided anything but the most dubious evidence for or against the validity of any theory of human behavior, whether biological, behaviorist, cognitivist, or what-have-you. The equality of benefits—if that were observed—surely would say nothing about the equality of theories in respect to heuristic power or capacity to grow. Nor, one must add quickly, would the superiority of effects for A versus B imply the theoretical superiority of A. The meta-analysis addressed a set of practical policy questions. Smith et al. (1980) erred when their rhetoric slipped momentarily off these questions and into the domain of traditional psychological theory:

> We did not expect that the demonstrable benefits of quite different types of psychotherapy would be so little different. It is the most startling and intriguing finding we came across. All psychotherapy researchers should be prompted to ask how it can be so. If it is truly so that major differences in technique count for little in terms of benefits, than what is to be made of the volumes devoted to the careful drawing of distinctions among styles of psychotherapy? And what is to be made of the deep divisions and animosities among different psychotherapy schools?

These are the kinds of sweeping questions that too

often evoke trite and thoughtless answers. Perhaps we can avoid both. We regard it as clearly possible that all psychotherapies are equally effective, or nearly so; and that the lines drawn among psychotherapy schools are small distinctions, cherished by those who draw them, but all the same distinctions that make no important differences. Those elements that unite different types of psychotherapy may prove to be far more influential than those specific elements that distinguish them. (pp. 185–186)

This statement, drawn from Smith et al.'s chapter of conclusions, crosses over from the inquiry on what claims can be rationally supported from the literature of psychotherapy-outcome research to the inquiry into human behavior, its antecedents and consequences. Add to this the fact that some toes were trod upon and the angry reactions of some critics become more understandable. Perhaps Smith et al. should forgive their critics a few misapprehensions about object fields and explananda when they occasionally lost track themselves of the precise questions they could and could not answer. They did, however, properly address questions of the form, "Does a person who professes to administer Gestalt psychotherapy achieve benefits importantly and demonstrably superior to those produced by a therapist who calls his or her approach 'cognitive behavioral'?" The validity of answers to this question is supported by the answers to several auxiliary questions, for example, how does psychotherapy in practice compare with psychotherapy in print? How does "publication bias" distort one's perception of the efficacy of psychotherapy? These questions are clearly different from the question, presumably of most concern to several critics who were upset by the Smith et al. study, whether theory of human behavior A, B, or C is more valid. To underline how separate are the questions addressed in the psychotherapy meta-analysis from those addressed by various theories of psychotherapy, the first author confesses that in regard to theories of human behavior his predilections agree with Meehl's (1978): "I do have a soft spot in my heart . . . for psychoanalysis." (p. 829). What makes this personal revelation more than gratuitous is that there is nothing in the least inconsistent about believing, as the first author does, that psychoanalysis is by far the best theory of human behavior and acknowledging that there exists in the Smith et al. data base not a single experimental study that would qualify by even the shoddiest standards as an outcome evaluation of orthodox psychoanalysis. (For more on the distinction between evaluating pragmatic claims and testing theories in the context of psychoanalysis, see Scriven, 1959.)

Conclusion

This article is an apology to those scientists who have worked at psychotherapy research over the years and who saw things which they may not have approved of done to their efforts for reasons that were unclear. In our defense, we insist that our purposes were legitimate and violated no canons of science, even if we were derelict in stating the purposes clearly. In an attempt to correct any misapprehensions our work may have fostered, we have offered the remarks in this article to our critics in the spirit of, as the French say, "to understand all is to forgive all."

Reference Note

1. Reardon, J. P., & Tosi, D. J. *The effects of rational stage directed imagery on self concept and reduction of psychological stress in adolescent delinquent females.* Unpublished manuscript, Ohio State University, 1976.

References

Abeles, N. Psychotherapy works! *Contemporary Psychology*, 1981, *26*, 821–23.

Andrews, J. G., & Harvey, R. Does psychotherapy benefit neurotic patients? A reanalysis of the Smith, Glass, and Miller data. *Archives of General Psychiatry*, 1981, *38*, 951–962.

Bandura, A. On paradigms and recycled ideologies. *Cognitive Therapy and Research*, 1978, *2*, 79–103.

Bowers, K. S. Situationism in psychology: An analysis and critique. *Psychological Review*, 1973, *80*, 307–36.

Cartwright, D. S. Success in psychotherapy as a function of certain actuarial variables. *Journal of Consulting Psychology*, 1955, *19*, 357–63.

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.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, *30*, 116–27.

Crown, S. Review of *The Benefits of Psychotherapy. The British Journal of Psychiatry*, 1981, *139*, 199.

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

Eysenck, H. J. *Causes and effects of smoking.* Beverly Hills, Calif.: SAGE, 1981.

Feyerabend, P. *Against method.* London: Verso, 1978.

Frances, A., & Clarkin, J. Review of *The Benefits of Psychotherapy. Hospital and Community Psychiatry,* 1981, *32,* 807–8.

Gergen, K. J. Social psychology as history. *Journal of Personality and Social Psychology,* 1973, *26,* 309–20.

Glass, G. V. Primary, secondary and meta-analysis of research. *Educational Researcher,* 1976, *5,* 3–8.

Glass, G. V. Policy for the unpredictable: Uncertainty research and policy. *Educational Researcher,* 1979, *8,* 12–14.

Glass, G. V, McGaw, B., & Smith, M. L. *Meta-analysis in social research.* Beverly Hills, Calif.: SAGE Publications, 1981.

Glass, G. V, & Smith, M. L. Meta-analysis of research on the relationship of class-size and achievement. *Educational Evaluation and Policy Analysis,* 1979, *1,* 2–16.

Habermas, J. *Knowledge and human interests.* Boston: Beacon Press, 1971.

Kazrin, A., Durac, J., & Agteros, T. Meta-meta analysis: A new method for evaluating therapy outcome. *Behavior Research and Therapy,* 1979, *17,* 397–99.

Kiesler, D. J. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin,* 1966, *65,* 110–36.

Kuhn, T. S. *The structure of scientific revolutions.* Chicago: University of Chicago Press, 1962.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review,* 1971, *41,* 429–71.

Meehl, P. E. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 806–34.

Oakeshott, M. J. *Rationalism in politics, and other essays.* London: Methuen, 1962.

Paul, G. L. Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology,* 1967, *31,* 109–18.

Presby, S. Overly broad categories obscure important differences between therapies. *American Psychologist,* 1978, *33,* 514–5.

Rachman, S. J., & Wilson, G. T. *The effects of psychological therapy: second enlarged edition.* Oxford: Pergamon Press, 1980.

Scriven, M. The experimental investigation of psychoanalysis. In Hook, S. (Ed.), *Philosophy, scientific method and psychoanalysis.* N.Y.: New York University Press, 1959.

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

Simon, J. Review of *The benefits of psychotherapy. American Journal of Psychiatry,* 1981, *138,* 1399–1400.

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

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 1977, *32,* 752–60.

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

Taylor, J. W. Relationship of success and length in psychotherapy. *Journal of Consulting Psychology,* 1956, *20,* 332.

U.S. Public Health Service. *Smoking and health.* Report of the Advisory Committee to the Surgeon General. Washington, D.C.: U.S. Government Printing Office, 1963.

Received March 26, 1982
Revision received April 19, 1982 ■