# 3.
# Methods for Assessing the Effectiveness of Psychotherapy

# 3. Methods for Assessing the Effectiveness of Psychotherapy

Given the diversity of theoretical approaches to psychotherapy, as well as the range of mental health problems, therapists, and delivery settings, it is difficult to give simple answers to questions about psychotherapeutic efficacy. Although the position adopted here will be that it is possible to answer efficacy questions through scientific research, such questions are undoubtedly difficult to answer. Both because of the inherent nature of psychotherapy and because of the nature of the scientific research process, answers to questions about psychotherapy will require a complex sequence of steps and the adaptation of several research technologies to the problem of psychotherapy.

Questions about psychotherapy are complex, because, for policy purposes, they have been stated very globally. It must be recognized that general statements as to whether psychotherapy is effective are necessarily equivocal and must be tempered by information as to the specific conditions under which particular treatments are efficacious. The way in which the level of specificity of one's question affects research on psychotherapy and the methodological considerations that affect the conduct and interpretation of efficacy research are described below.

Most research assessing the effectiveness of psychotherapy has examined very specific issues. Which technique is more effective and how effectiveness is moderated by differences among patients, therapists, and settings are the typical foci of psychotherapy outcome research (see 207,287). Much of this research shows that some techniques are more effective than others, although, unfortunately, no-treatment or placebo treatment control conditions are not always included as part of these studies. Without such comparison conditions, their implications for the ultimate effectiveness questions are difficult to assess.

Questions about the effectiveness of specific types of psychotherapy usually deal with the conditions under which therapy is provided. Thus, the type of mental dysfunction, the characteristics of the patient, and the characteristics of the delivery system are the central variables being tested. Although this research can yield generalizations to policy about psychotherapeutic treatment, the inherent limitations should be recognized. Unless generalizability has been empirically established by tests conducted with a range of patients in an actual treatment site, the conclusions must be regarded as tenuous.

A number of methodological issues that arise in assessing the efficacy of psychotherapy are described in the following sections. The first issue has to do with how one measures the outcomes of therapy. There is a substantial literature describing procedures for determining the presence and strength of particular results of therapy (see 293). This literature is examined below in terms of the development of useful policy data about psychotherapy. A second set of methodological issues concerns the design of psychotherapy research and the confidence one can have that obtained changes are a result of a particular psychotherapeutic intervention. The research design problems have to do with determining the reason for outcomes and organizing research so that extraneous factors can be ruled out as the cause of treatment effects (see 41, 105). Also considered below are the problems of actually conducting psychotherapy research, including the ethical and pragmatic issues of employing random assignment procedures, A separate set of methodological problems having to do with the synthesis and interpretation of multiple efficacy studies is also described. A number of techniques have been developed for reviewing and integrating findings, and, potentially, these methods may allow more definitive assessments of the psychotherapy literature.

# MEASURES OF PSYCHOTHERAPEUTIC OUTCOMES

Measuring the outcomes of psychotherapeutic treatment has been a major focus of psychotherapy theory and research for at least the last two decades. Developing a technology for measuring outcomes is the first step in determining psychotherapy's efficacy. It involves decisions about what variables are important to assess, as well as the development of measurement techniques that can be used in actual treatment settings. Because psychotherapy takes a number of forms (e.g., treatment based on different theoretical assumptions), much of the literature deals with the selection of outcome measures that are appropriate to the goals of the treatment under study. One recent focus has been the development of measurement procedures that can be used across different types of psychotherapy. Some of this literature has proposed "batteries" of instruments (e.g., 293), whose intent is to capture (through the use of several types of measurement procedures) the core changes that result from any application of psychotherapy.

Underlying the effort to develop such common measures is a belief that no one instrument or set of procedures can measure all the outcomes of psychotherapy (21). Different psychotherapeutic techniques applied by a range of therapists to various patient populations may require different measures of outcome. Thus, functional measures of behavioral effects, although perhaps appropriate to assess behavioral therapies, might be inappropriate to assess psychodynamically based therapies. Similarly, cognitive measures might be seen as inappropriate to evaluate behavioral therapies.

Notwithstanding the unique goals of particular therapies, there seems to be support for the concept that many of the changes produced by psychotherapy can be assessed along some common matrix. Probably, this implies the use of a matrix that includes both behavioral and cognitive variables. Any single study of psychotherapy, thus, would incorporate a number of measures, not necessarily tied to the goals of the therapy. The use of such multiple common outcomes also makes monitoring potential detrimental effects more possible. Such effects may not be detectable if unique therapy-relevant measures are used.

## Measurement Criteria

To be useful in an effectiveness analysis, measures of psychotherapy outcome must be both reliable and valid (see 201). Reliability means that the measure gives the same finding over multiple uses (assuming no change in what is being measured) and provides the same findings when used by different researchers. Validity means that the measure assesses the outcomes that it is supposed to measure and provides data that are generalizable (see, e.g., 48). Neither reliability nor validity is intended as a theoretical concept; each is established by pretesting the measuring instrument.

The use of reliability and validity criteria results in particular measurement processes' being validated to measure specific therapeutic effects. Thus, for example, a single instrument might not be reliable and valid for assessing improvement in both depression and agoraphobia; reliability and validity, at least, would have to be separately established for each condition. Moreover, outcomes have different meanings to patients, the therapist, and interested third parties. Validity may depend on who's perspective one adopts in assessing the measurement instrument. Below, some of the differences in measures designed to collect data from various individuals affected by therapy are described.

## Measures From Patients

Reports and ratings by patients of their behavior, thoughts, and feelings represent one typical, and usually important outcome measure. Often, such data are collected on questionnaires or through interviews. These measures structure verbal reports of the patient's ability to cope with various problems, and include paper-and-pencil measures of personality and adjustment (31). Although patient assessment measures yield important data, they have obvious limitations, These limitations include

social desirability effects (patients' responding to create a certain impression) and response bias **(e.g.,** positive responding to express appreciation to the therapist). Validation using other types of measures can determine the effects of such response patterns.

One type of patient measure is a functional self-report instrument which asks patients to report on aspects of their daily functioning (123). Although narrative reports of functioning may be collected as a part of any therapeutic treatment, instruments that incorporate standardized questions have been used to systematize this data collection. Such an instrument may include a series of structured questions about time lost from work, feelings of guilt, and satisfaction with therapy. Careful pretesting of these questions (e.g., by comparing responses of individuals known to be psychologically impaired with those not impaired) yields a subset of these questions. Responses can be quantified and summed to form composite scores (total and subset) of social adjustment and functioning. A number of these instruments have been developed which show high reliability and validity (see 123).

Another type of patient measure is represented by "psychometric" questionnaires and personality tests, such as the Minnesota Multiphasic Personality Inventory (MMPI). Such instruments, which are perhaps the most common type of measuring tool for both diagnostic and assessment purposes, ask a variety of questions about the respondent's thoughts. Patients, for example, those suffering from depression and anxiety (210), have been shown to respond to these questions according to particular patterns. For MMPI, a subset of questions has been developed which tends to be answered differently by those who are trying to fake responses and those who are not; thus, social desirability and other biased response patterns can be detected (see 55).

## Measures From Family and Friends

Friends, work associates, and relatives of patients are often asked to supply information about a patient's functioning, and such data concerning patient behavior and inferred mental states have been used to assess therapy outcomes (e.g., 116; see, in particular, a report by the National Institute of Mental Health, **293).** The value of these measures may be due to the great amount of information that family, friends, and work associates have about the functioning of the patient. These individuals have a chance to observe the patient in a variety of situations and may thus have a great deal of "data" on which to base their responses to questionnaires. Although these individuals have self-interests which can bias the data, they may have less motivation than the patient to reflect socially desirable responses and less need to show gratitude to the therapist.

The outcome variables included on these questionnaires also assess outcomes that may be more important to "society" than the variables included on patient self-report questionnaires. For example, questions about how much disturbance the patient causes in family life or work routines, or how often the patient has secured employment or performed satisfactorily in school are typically included in these instruments. Such questions may be more important for societal evaluations of the usefulness of therapy than questions about coping with daily activities that are typical of self-report measures. These techniques are validated both by comparing the observations of those who know the patient against one another (e. g., family members compared to work associates) and by comparing these data with other available information about the patient.

## Measures From Therapists

Measures taken from the therapist and others involved in the therapeutic process are another frequently used source of outcome data. While the therapist can provide a first-hand perspective on the therapeutic process, such data may be biased because the therapist has a vested interest in producing positive outcomes. Nevertheless, a substantial research literature on such "clinical judgments" exists, and there is some evidence that therapists can provide useful and relatively unbiased reports. Particularly in terms of assessments of patient functioning, there is evidence that therapists can provide valid data (e.g., 67,108,161,184,198).

Generally, the more specific and concrete the observations required of therapists, the greater the resulting interrater reliability and validity (186,187). These observations can range from "counts" of behavior exhibited by patients during therapy to therapist ratings of the patient's functioning with his/her family (based, perhaps, on how the patient has reported family relationships). To assess some therapies, where the concern is more with mental phenomena than behaviors, instruments have been designed to capture nonobservable behavior. While such instruments can provide one type of information about the effect of therapy, their usefulness may depend on how the goals of therapy have been described and the availability of validation data.

A focus on the specific behaviors of the patient may allow therapists to provide easily validated data about patient functioning; however, the therapist's desire to show improvements as a result of therapy may bias the use of such measures. In an attempt to remove this bias from functioning judgments, some researchers have trained members of a therapy staff who are not individuals actually involved in therapy to conduct such assessments (e.g., 209). A variety of such "blind" data collection techniques have been employed.

## Measures From Community Members

Information on the outcome of psychotherapy can also be collected, more broadly, from community members or agencies. These measures may include patient data on criminal arrest rates, measures related to the patient's work, or rates of medical utilization (from nonpsychiatric illness). The range of potential community measures is very large and, though often unwieldy, seems to reflect important information that is needed to assess adequately the effects of psychotherapy. As is described later in the discussion of cost-benefit measurement (ch. 5), such variables are important to assess in order to conduct comprehensive cost-effectiveness and cost-benefit analyses (CEA/CBAs).

Many of the data collected on community variables can be assessed in relatively direct ways. The information often can be gleaned from existing court records, hospital charts, insurance claims, and similar records. To be utilized as part of an effectiveness assessment, however, such measures must be tested for reliability and validity. Oftentimes, the validity of record data is easy to establish because of its obvious relationship to desirable psychotherapy outcomes (referred to as content validity). It should be noted that much of this information can also be collected directly from patients (e.g., data on medical utilization, if it can be shown that patients report these data accurately).

Community data have also been taken to mean a patient's economic contribution to his or her community (see, e.g., 305). Thus, the patient's net monetary contribution in terms of earnings and taxes may be used as one measure of psychotherapeutic outcome (see ch. 5). These outcomes have been assessed for psychotherapy in a number of investigations (e.g., 51,114) and are described later in terms of CEA/CBAs of psychotherapy.

## Summary

The outcomes of psychotherapy can be measured in a variety of ways. It is probably impossible to develop any single measure of outcome which would reflect the diverse changes that might be brought about by psychotherapy. At present, there exists a diverse set of procedures for eliciting information from patients, family, therapists, and others. Data from each of these sources, if properly assessed for reliability and validity, provide information needed to assess psychotherapeutic effectiveness.

## RESEARCH DESIGNS

The development of reliable and valid outcome measures represents only one part of the assessment problem. One must be able to determine which of the many processes utilized in therapy are responsible for improvements and must also be able to test the possibility that nontherapeutic components are responsible for obtained effects. The particular research designs and techniques used to test psychotherapeutic efficacy are, in part, determined by what questions are being asked. These questions depend on the goals of both the patients and the therapy program. The present discussion focuses on the possible ways of carrying out psychotherapy outcome research and how questions about psychotherapy can actually be tested.

A basic element of good research design is to frame the questions to be tested in a very specific way. Whereas the global question of psychotherapy's efficacy may be obvious (i. e., "is therapy effective?"), the questions asked of research need to be more circumscribed (21,207). Such circumscribed questions usually develop as research progresses through a series of stages (105). At a formulative stage, research is based on extensive observation and summary descriptions of these observations. Then, the focus shifts to a description of patterns among data elements. At a later stage, explanations and theories of the observed patterns are formed. It is usually these later explanations that are tested in formal experiments and that represent the bulk of the outcome research literature. Although effectiveness studies can be conducted on a post hoc basis (i. e., a design constructed after the data have been collected), such research usually can be interpreted only when theory and formal experimental evidence are available (e. g., 47,1.51 ).

The purpose of the present report is to understand psychotherapeutic effectiveness; thus, the focus is primarily on data collected in actual treatment settings (in vivo). A great deal of psychotherapy research has been carried out under laboratory or analog conditions, however, and at least some researchers (e.g., 10) view the data from this research as very important. To the extent that such data provide theoretical support for in vivo findings, they are probably necessary to consider. OTA, in previous discussions of the evaluation of medical technology (202), has regarded this as efficacy rather than effectiveness research (ideal v. actual conditions). When applied to psychotherapy, however, the efficacy/effectiveness distinction seems ambiguous because of the variety of factors (therapist, patient, setting) which affect the outcome of a particular treatment and the absence of a clear demarcation between laboratory and nonlaboratory conditions. It seems desirable in the case of psychotherapy, instead of trying to differentiate between assessments of efficacy and effectiveness research, to note clearly the different conditions under which research is conducted.

In *terms* of developing research tests of therapeutic effects, some limitations of prevailing scientific logic should also be noted (see 41), Testing particular hypotheses does not permit the psychotherapy researcher to prove that a particular therapeutic effort causes a particular outcome; instead, one tests whether alternative explanations can be disregarded (47,105). Resulting inferences are probabilistic and indicate, within identifiable error rates, the likelihood that generalization can be made from the study's sample to other populations.

The inherent variability of human behavior, thoughts, and feelings often makes the findings of psychotherapy research equivocal. A great deal of variability will exist under any conditions, and these variables must be separated from the effects of the treatment. In addition, because patients can be affected by many variables (i.e., factors other than psychotherapy) that cannot be controlled, there may be numer-

ous alternative explanations for any apparent improvements produced by psychotherapy. The number of alternative explanations can be reduced somewhat by careful design of psychotherapy research and by the inclusion of control conditions that hold all elements constant, except some aspect of a treatment. A variety of experimental designs have been developed to reduce the plausibility of a number of standard alternative explanations. These research designs are not unique to psychotherapy (see 27,41, 94,165), but their use in assessing the efficacy of psychotherapy raises a significant set of unique problems. Several types of research designs that can be used to assess psychotherapy are described below.

## Therapy Versus No-Therapy Designs

The classical research design, as applied to psychotherapy, assigns patients either to receive psychotherapy or not to receive psychotherapy according to a random selection procedure. It can be expected that randomization will distribute differences in patient characteristics (e.g., level of mental dysfunction, amenability to treatment) equally between the psychotherapy and no-psychotherapy conditions. Typically, the psychotherapy condition is referred to as the "experimental" condition, while the no-psychotherapy condition is referred to as the "control" condition. The functioning of each subject in the experiment is measured following psychotherapy (or following an equal interval of time for control subjects). Although measures can be taken at other times (e.g., pretherapy), additional design problems (having to do with the effect of taking a test or responding to a questionnaire more than once) result with such procedures.

The measurements of functioning (including cognitive and behavioral outcomes) obtained from the treatment condition contain three possible "effects:" 1) "effects" of treatment, 2) "effects" due to whatever nontherapy factors are affecting the patient, and 3) "effects" of the haphazard fluctuations in functioning measures caused by imperfections in measurement instruments and the variability of behavior. The measurements of patient functioning obtained from the control condition are used to "subtract" the "effects" due to treatment from the "effects" due to other factors. A control condition is necessary to perform the above "subtraction" because it provides the only empirical way of knowing how nontherapy factors and measurement problems affected the outcome.

When this type of therapy versus no-therapy research design is used, explanations of apparent improvements in patient functioning that are actually due to factors other than therapy itself can be rejected with reasonable confidence. As long as patients have been assigned randomly to therapy and no-therapy conditions, it can be assumed (within identifiable probability limits) that the obtained "effects" are due to the psychotherapeutic treatment. If patients are not randomly assigned, but are "matched" on various characteristics, such an unequivocal statement is not possible. It can be argued that differences existed between experimental and control group subjects that were not controlled by matching and that these characteristics are responsible for differences obtained between therapy and no-therapy groups. The absence of a control group makes such inferences about causal factors extremely difficult to develop.

## Therapy Versus Therapy

The basic rationale used to distill the effects of psychotherapy from effects of measurement and factors unrelated to therapy can be extended to test for the superiority of one psychotherapy over another. In such a therapy comparison study, patients are assigned randomly to therapy A, therapy B, . . ., and, perhaps, to a no-therapy or delayed-therapy group. In effect, such a design results in the use of multiple treatment groups. The use of no-treatment control groups is not ruled out, but such groups are often not employed in therapy comparison studies because the purpose of such studies is the assessment of the best therapy. It should be noted that comparisons across therapies are often made without use of an experimental research design. When that is done, patients who have received different therapies are compared without regard to the selection factors that in-

fluenced which patients received which therapy. In such nonexperimental research, differences may be due to a variety of factors (e.g., preexisting differences between patients in each group), and these variables need to be controlled before such data are useful.

In another common design used to compare therapies, two potentiality effective therapy techniques are presented separately to two groups of subjects and in combination to a third group (yielding individual and combination therapy conditions). Statistical analyses are used to separate the effects of the therapies and to compare them with no-therapy conditions. A particular concern of the statistical analysis is to test for interaction effects (e. g., where therapies combine to produce an effect that is different from the sum of the two effects alone). A variant of this type of design has been used to assess the joint, as well as separate, effects of chemotherapy and psychotherapy (e.g., 142).

## Therapy Versus "Placebo" Therapy

One problem that has plagued much psychotherapy research (as well as other research on medical interventions) is that some of the effects obtained by psychotherapy researchers may be due to placebo effects. The "aura" of being in therapy and the expectancy that one is finally about to be "cured" may be a form of treatment (136,256). The problem of separating these effects from those of formal therapies is analogous to the use of sugar pills in controlled medical research and involves the use of placebo-control conditions. In such conditions, the patient may receive attention from a therapist, but therapeutically meaningful discussion is avoided and no specific techniques are used (e.g., 176). It is easier to employ such control procedures when testing the efficacy of behavioral, or *even* psychodynamic, therapies, in which the therapist plays an active and directive role.

Probably, because of the nature of psychotherapy and the difficulty of specifying the precise ingredients of therapy, it is impossible to control all placebo effects (207,280). The relationship that a therapist establishes with a patient is acknowledged (e. g., 84) to be an important component of therapy, and it is not clear how such effects should be distinguished from treatments per se. It should also be recognized that while placebo effects may inflate the true effects of therapy, they may often distort the data obtained in control conditions. Few control conditions can be "pure" in the sense that patients do not interact with a therapist. If for no other reason than to monitor the patient, those in control conditions must usually be supervised by a therapist. The effects of this supervision may make therapy appear to be less effective.

Control conditions may be introduced to assess the effects of such factors as therapist "demand characteristics. " An ardent researcher or therapist may communicate to patients what the results of an experiment are "expected" to be (e.g., 240). Experiments using se] f-report measures obtained from patients are especially prone to this problem. To examine the strength of these demand characteristics, psychotherapy researchers can use designs in which patients are told that therapy will produce a temporary worsening of functioning, when in actuality the researcher expects no worsening. Alternatively, the researcher may tell patients in some conditions that therapy should not be effective for several months, when gradual improvement actually is expected. If the "demanded" effects are not found in these conditions, the researcher can have some confidence that demand characteristics are not causing therapy effects. Although these controls may not be employed in every outcome study, they are often used in developing research designs.

## Quasi-Experimental Designs

For a variety of reasons, including ethical problems of withholding treatment (see discussion below), the experimental designs described above are often not employed. The real or perceived difficulties of assigning patients on a random basis to therapy or no-therapy conditions make other types of comparisons necessary. There are a number of quasi-experimental designs that can be used in such circumstances (see 41), Such designs are sometimes considered to be poor substitutes for "pure" experimental designs, but this may not be the best way to view them. Quasi-experimental designs, if care-

fully constituted, eliminate the most important plausible alternative explanations for the results of an experiment and can provide useful information. Quasi-experiments based on these designs, however, may involve complex statistics and require the collection of many more data (perhaps from fewer subjects) than a "true" experiment would.

One common quasi-experimental design in psychotherapy is referred to as an intensive design (97). Typically, intensive designs use fewer subjects than experimental designs, but compensate by obtaining more measures on a frequent basis and by obtaining clear "baseline" measures of patient functioning. The intent is that sudden improvements will be dramatic and closely correlated with the onset of therapy. To the extent that these changes are sudden, many of the alternative explanations can be ruled out (despite the lack of a control group). The use of intensive designs, however, depends on the type of therapy and the outcomes that are expected. Only therapies that posit observable and rapid changes (typically, the more behavioral therapies) can be tested with such designs.

Return-to-Baseline Designs.—In this design, patient functioning is measured several times before therapy (the baseline period) and several times during therapy (the manipulation period). Therapy is then withdrawn abruptly (reversal to second baseline), and it is expected that there will be a return to lower levels of functioning similar to those found during the first baseline period. Repeated measurement continues during this period and during a return to therapy (second manipulation). Therapy is then phased out with the hope that improvements will be maintained.

An example of this type of design is Allen, Hart, Buell, Harris, and Wolf's (3) study of a single child. The child had isolated herself from other children, causing severe psychological problems. The amount of time that the child spent with adults and other children was measured during the baseline period. A simple therapy that provided reinforcement for her interactions with other children was then provided. Measurement continued during this period, as well as after therapy. Later, there was an abrupt withdrawal of the reinforcement (reversal to second baseline), followed by its reinstitution. The results showed very noticeable changes in the girl's interactions. In the absence of other factors (besides therapy), the effects were seen as demonstrating therapeutic efficacy.

Intrasubject Multiple Baseline Designs.— To obtain information analogous to that obtained from a separate no-therapy control group, it is also possible to separately treat and assess functioning in different areas of the patient's life. Thus, treatment is designed to improve one aspect of functioning, and the patient's other behavior is used as a control. Such a design, while having advantages because changes can be compared with only one subject, suffers from the obvious problem that effects may generalize across function areas. In fact, for some therapies generalization of effects could be expected, and an intrasubject multiple baseline design would lead to "missing" the effects of therapy.

An example of this type of design is provided by Morganstern (189). The effectiveness of aversive conditioning on obesity (pairing nauseous smoke with eating certain foods) was examined by conditioning a patient to one type of food, then after several days, to another. Because eating of other foods was found not to change until aversive conditioning was applied to the particular foods, it was possible to infer that the therapy produced changes.

There are a number of variations on this intrasubject multiple baseline design. Multiple subjects can be used, and different combinations of treatment can be tested. This design does not eliminate alternative explanations due to placebo or demand characteristics, but can eliminate some problems inherent in nonrandomized experiments. Thus, additional subjects can serve as controls to assess the importance of factors such as spontaneous remission (improvement without treatment). However, the use of this design may be severely limited by the types of problems (very specific) and the therapies (mostly behavioral) which one tests.

## Program Evaluations

The design considerations described above implicitly refer to psychotherapy as a unitary treatment that can be applied or not applied to form experimental and control conditions. However, as chapter 2 points out, psychotherapy treatments are comprised of a number of factors. The theoretical orientation of the therapy, usually the basis of a label for the treatment, may not adequately describe the treatment as it is actually delivered. Depending on the nature of patients' problems, therapists' orientation and skill, as well as aspects of the delivery setting, therapy may have different outcomes. One possibility is to treat each of these factors as an independent variable and to construct multifactor designs. In such designs, subjects are randomly assigned to different therapies, therapists, and settings. A more recent trend, though, has been to conduct such outcome research through program evaluations (see 212). In such evaluative research (e.g., 9, 159,252), one evaluates a complex of treatment variables that have been organized as a program (e.g., a community mental health center (CMHC)).

Thus, for example, [a]Program evaluation study of a CMHC tries to assess how and to what extent patients who receive treatment at the center are aided. The CMHC, in addition to being community based (which is hypothesized to be an adjunct to treatment), may offer patients a number of therapies, and patients may be treated by multiple professionals and paraprofessionals. Under such circumstances, where joint effects of these treatments are expected and where it is extremely difficult to separate—for research design purposes—the components of treatment, program evaluation yields a design that may be more compatible with the actual circumstances. Evaluative research does not preclude the conduct of experiments where individual aspects of the treatment are assessed.

The designs for program evaluation studies of psychotherapy can include aspects of the true and quasi-experimental designs described above. In general, the same methodological considerations for research designs apply to program evaluations (229,298,303). There is, in fact, a substantial literature on the use of the experimental designs in program evaluation (e.g., 230,249). The literature describes both how complex variables such as psychotherapy can be conceptualized and the conditions for implementing randomized designs. Similarly, the literature on quasi-experimental designs has been related to program evaluation (e.g., 41, 230). The principal difference between these designs and traditional research design is in how one conceptualizes the treatment. In a program evaluation study, the treatment includes a number of elements. These elements, at least in the initial stages of such an evaluation, are not separately tested.

Even though program evaluations are designed specifically to aid in policy decisionmaking, there are a number of endemic problems. Just as it is difficult to organize a traditional psychotherapy outcome study (i. e., randomly or otherwise assign patients to treatments), it is difficult to organize program evaluation studies. In fact, because the randomization units are more complex, such evaluation studies are often very difficult to carry out well, and there is a substantial literature about implementation failures. In addition, program evaluation studies may not resolve the underlying conceptual problems involved in assessing psychotherapy. It may be difficult to determine through a program evaluation what factors were responsible for the success or failure of the program.

## Difficulties in Conducting Research

There are a variety of problems which make it difficult to conduct psychotherapy outcome research. Some of these problems, which relate to the inappropriateness of some methods to test particular therapies (e.g., intensive designs) and the problem of multiple factors affecting outcome (e. g., the role of therapist variable), have already been described. There are also problems, however, which are perhaps more important, having to do with the pragmatic and ethical difficulties of conducting experimental research.

While the advantages of experimental methods to develop unequivocal data are well known and widely accepted, there are obvious ethical problems connected with decisions to withhold treatment (e.g., 45,229) or to make treatment available on a random basis. Especially if one believes that therapy is efficacious, it is difficult not to allow certain patients to receive treatment for research purposes. Several considerations are important in thinking about this dilemma.

One important consideration is the necessity of conducting research, especially that which employs randomized control groups. Such studies (which, in medicine, are referred to as randomized clinical trials) make it possible to assess causality in the most unambiguous way (see, e.g., 36,41,165). Although, from a theoretical point of view, the value of randomized control group studies has long been recognized, there is now some evidence that such studies better enable researchers to detect inefficacious treatments (see, e.g., 94). If less adequate designs are used (i.e., designs not employing randomized control conditions), decision errors may result. Thus, it may appear, perhaps because of the effects of other variables, that the treatment is effective when it is not.

One resolution to the ethical problems presented by needing to withhold treatment is suggested by the nature of the dilemma. If a treatment has not been demonstrated to be effective, then it may not be unethical to deny treatment to some individuals. The treatment, in the absence of empirical data, may not be accomplishing anything. Actual treatment providers (i.e., therapists) may not share this view and, therefore, may be reluctant to participate in this form of experimental research. At the very least, practitioners might want to be able to supervise control group subjects and, in effect, provide partial treatment.

One pragmatic resolution to the ethical dilemma of this research is to provide control group subjects with access to treatment once the experiment has been concluded. This is often referred to as the "waiting list" approach. It is made a more attractive option in some studies (which compare several therapeutic approaches) by offering the delayed control group subjects the form of treatment which the experiment demonstrated as best. Obviously, this option can be provided only when therapy is of a relatively short duration. It is, however, easy to implement when resources are very limited. If only a small group of patients can be given treatment at any one time, a waiting list (established by random assignment) may be practicable (although it might be viable only when patients do not have other treatment options).

Whatever the justification for random assignment to treatment and control conditions, explicit guidelines (established by the Department of Health and Human Services and professional societies) must be followed to protect patients' rights. One central principle of these guidelines is that participation be voluntary, based on patients' (or legal surrogates') "informed consent" (see, e.g., 6). Informed consent procedures require that subjects be informed as to the purposes of the experiment, known risks, data to be collected, and how the data will be used. Subjects also have the right to leave the experiment at any time and, usually, must be told of other treatment resources. The procedures also require that review panels be established to approve research protocols and monitor the conduct of this human research.

Although such procedures may create differential attrition across conditions of an experiment, and perhaps result in only certain types of patients' or treatments' being involved in research, it does not rule out the use of experimental designs. Under a variety of circumstances, individuals will agree to participate in experimental research, and the problems may have more to do with the researchers or therapists involved than with patients. As noted above, until unequivocal evidence is available about psychotherapy and more resources are available for such treatments, it seems necessary to conduct such experimental studies. These studies are critical to providing unequivocal tests of the theoretical hypotheses suggested by basic research. For many therapies, without experimental evidence, there will be no way of resolving questions about alternative explanations for their effects (see 207). In order to carry out such

studies, it may be necessary for the Government to develop special rules for their conduct. Thus, for example, special reimbursement procedures may need to be applied so that treatment and research costs can be separated.

Given the difficulties of conducting in vivo experimental research, it is perhaps not surprising that much of the best controlled research has been laboratory analogs to clinical settings. In these analog settings, psychotherapy is offered to patients who often have less severe dysfunctions than are seen in psychotherapy clinics. However, these patients can be assigned to no-treatment and placebo conditions because the psychotherapy researcher controls the program and because the loss of the therapeutic benefit has less impact. This type of research may yield more rigorous experimental findings, but has low external validity; that is, the patients (often undergraduate students), therapists (often doctoral students), and procedures (more theoretically oriented and less eclectic than is typical) are not representative of those used in "real" psychotherapy (e.g., 162,169). It should be noted that some researchers suggest that this is the best way to conduct rigorous investigations of the effectiveness of psychotherapy and that the innovative techniques developed and tested in such settings can be transferred to more "messy, " real world settings *(10,* 135, 136). In essence, they argue that the differences between actual and analog settings are not as great as might be anticipated and that generalizability can be demonstrated for these analog studies (at least for some therapies and conditions of treatment ).

## INTEGRATING FINDINGS FROM PSYCHOTHERAPY RESEARCH

Although the key to evaluating the efficacy of psychotherapy lies in the conduct of well-designed research which uses multiple measures that are both reliable and valid, this may not be sufficient for policymaking purposes. No one research design (including true experiments and program evaluations) will enable the development of definitive information about the effects of psychotherapy, nor will the measures used in any one study be adequate for all purposes. In part, this is because of the current state of theorizing about psychopathology, which encompasses a number of approaches each with different ideas about research. The problems, however, are not only theoretical, but also reflect the limitations of the scientific process with respect to developing unambiguous conclusions on the basis of individual research studies. Given the diversity of criteria against which such psychotherapy research must be evaluated, as well as the divergent theoretical views, methods are required to synthesize, aggregate, or integrate the findings of multiple studies.

Despite the fact that it might be desirable to have a few studies which "settle" the effectiveness question, for the most part, it is necessary to treat cautiously the results of individual studies. Their results must be judged against other research designed to test the same or similar hypotheses. Traditionally, such judgments have been made through literature reviews, where scholars analyze a body of research. These scholarly analyses are reviewed by peers and published so that other researchers can comment. Because of the scientific standards underlying peer review (see, e.g., 39), most such reviews reflect honest efforts to weigh the appropriate evidence.

However, even though the evidence may be reported accurately, it is fairly easy to be selective about the research data one includes. It is rarely possible, especially when considering the host of problems for which psychotherapy has been applied, to include all potentially relevant research. Any researcher who conducts a review makes a number of implicit and explicit choices about what will be included. As well, choices are made about what elements of the studies will be discussed and those that will be ignored; in effect, the author "slants" the review to support a particular hypothesis or viewpoint. Clear criteria are not always provided for judging the

methodological adequacy of the studies included in the review.

To make this process more scientific, or at least more systematic, a number of procedures have recently been developed and applied to assessments of the psychotherapy outcome literature. These procedures, although they have their own limitations and are by no means unanimously accepted, represent attempts to make sense of the burgeoning and often contradictory research about psychotherapy. Two such procedures are discussed below: "box-score analyses" and "meta-analyses."

Box-Score Analyses.—The procedure of "box-score analysis" begins with identification of a population of research studies (see, e.g., 162). Usually, the reviewer establishes certain standards and excludes studies that are not sufficiently rigorous by methodological criteria, or are otherwise not appropriate. The latter category might include studies that are well designed but lack sufficiently reliable or valid outcome measures. It might also include studies that do not have patients assigned randomly to conditions. The difference between this procedure and that used in literature reviews is not sharply defined and is only the greater degree to which criteria are explicitly described in box-score analyses and used to select research for the review.

All studies meeting the reviewer's criteria are culled from some defined set of sources (e. g., journals) and are then sorted into categories. Typical categories might be dysfunction treated, therapy technique, and/or training of therapist. Finally, the reviewer evaluates each study's outcome (e. g., "yes, " "no, " "equivocal") and tallies scores using the predefine categories (e. g., dysfunction, technique, therapists).

This method of categorization and then tallying of effects is designed to systematize the literature review process. However, it still leaves a great deal of room for individual judgment and typically uses a rather simplistic measure of treatment outcome. Importantly, it does not take account of the strength of findings within particular studies. A box-score analysis does not take account of the methodological rigor of a study, except in global fashion (e. g., by excluding studies without a particular kind of control condition). It is, nevertheless, a useful procedure, and a number of important box-score-type reviews of the psychotherapy literature are reported in chapter 4.

Meta-Analyses.—As described by Smith and Glass (267) and Smith, Glass, and Miller (268), "meta-analysis" is a procedure for integrating research findings (see also 239). It is a more recent, and undoubtedly controversial, set of procedures (see ch. 4). Meta-analysis employs statistical techniques for aggregating data and for determining relationships between causal variables and outcomes. Usually, the first step in a meta-analysis is the precise description of a population of studies on which the analyses will be based. In this respect, a meta-analysis is conducted similarly to a box-score analysis.

In a meta-analysis, however, studies are then coded on a set of variables that are thought to be related to outcomes—the number of these variables is at the discretion of the analyst. The experience of the therapist, the patients' symptomatology, the quality of the research design, and the setting of the treatment are examples of the categories of variables that would be coded in an analysis of psychotherapy. These measures are later correlated with the outcomes (usually quantified) and used as the basis for organizing outcome results in terms of aspects of the studies. In Smith, et al. 's (268) work, an outcome measure for psychotherapy based on the size of the effect (standardized) was computed for almost 500 studies. The studies came from a population of controlled (i.e., comparison group) investigations. Although a global effect size measure, as calculated by Smith, et al., does not differentially weigh studies according to the quality of the measures employed or the design, such factors are controlled by coding each study in terms of instrument and design validity. If, for example, there is a difference between the size of the effects in studies that used poor measures and those that used good measures (i. e., low valid v. high valid), then one knows that a bias of some sort is operating.

Meta-analysis uses systematic methods to uncover trends in the available research literature.

Assuming that there are some good studies and coding criteria can be agreed on, it should prove to be a useful tool in understanding the research on the effects of psychotherapy. A potential problem, however, is that by focusing on available research, one ignores the fact that only research that reports positive and/or statistically significant findings may be published. Although some believe (241 ) that this could be a problem only when there are a few studies reporting significant findings and when the magnitude of the differences between groups is very small, obviously, the quality of available research is a critical factor in the usefulness of meta-analysis. Other problems, such as the compatibility of studies, are described in detail in the next chapter as part of a substantive review of Smith, et al.'s work.

## FINAL COMMENTS

This chapter has described the methodological strategies that can be used to assess psychotherapy. Implicit in this discussion was an assumption that psychotherapy represents a researchable intervention that can be evaluated using scientific criteria of measurement and design. Valid measures of psychotherapy outcome can be developed, and designs that allow relatively unequivocal assessment of psychotherapy can be constructed. Although there are some who maintain that it is not possible to assess psychotherapy because of its inherent complexity, it is not clear what types of information would be excluded by the scientific analyses described here.

Whether the methods described here have been applied appropriately and to what ends they have been used, however, is a different question. Despite the possibility of conducting research on the outcomes of psychotherapy, as is noted in the next chapter, well-conducted research is inadequate to answer at least some of the important questions about psychotherapy (207,277). The reasons for the lack of research are not clear, but probably have to do with both substantive features of psychotherapy and a complex set of pragmatic factors.

These factors include the difficulty of conceptualizing the multiple factors that are part of psychotherapy, as well as attitudinal differences on the part of researchers and therapists. In part, the development of program evaluation techniques may facilitate the conduct of policy-relevant research. In addition, more attention, by researchers, therapists, and the Government to the issues of conducting ethically acceptable controlled experiments, may help to develop better research. Finally, attention to the issue of conclusion-making, based on an analysis of multiple investigations, may lead to better understanding of exactly what is achieved by psychotherapy under what conditions.