SOME GUIDE LINES
for EVALUATIVE RESEARCH

assessing psycho-social change in individuals

ELIZABETH HERZOG,
Technical Studies Branch, Division of Research

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
SOCIAL and REHABILITATION SERVICE Children’s Bureau 1959

Provided by the Maternal and Child Health Library, Georgetown University
In its effort to contribute to the well-being of children, the Children's Bureau has always tried to encourage and contribute to research relating to child life and to services for children. One way it does this is to give research assistance or consultation to the extent that is feasible with a very small research staff. Another way is by written materials designed to be useful to those who wish to use the results of research or to engage in research activities. This report is one of a number that the Bureau has published devoted to research methods and findings.

The need for evaluating service programs is being recognized increasingly. A very large proportion of the research requests that come to the Children's Bureau ask for help in some type of evaluation.

One of the most important and most difficult types of evaluative research has to do with determining the effectiveness of efforts to bring about social or emotional change in individuals. A number of kinds of services and therapies are directed toward producing such change. Among those of particular concern to the Children's Bureau, with its aim of improving the well-being of children, are psychotherapy, social casework, group work, services aimed at the prevention of juvenile delinquency and at the treatment of delinquents, and certain aspects of parent and family-life education.

In all these fields attempts have been made to determine and evaluate the effectiveness of programs and services in bringing about social-psychological change. Apparently, however, the most sustained and varied effort has occurred in relation to psychotherapy. Accordingly, the Bureau chose this field for special investigation, branching out from it only enough to cover certain relevant studies in social casework. Our purpose was to see how the questions common to evaluative research of any kind were dealt with, and what special problems were encountered in evaluating efforts to induce social-psychological change in individuals, particularly efforts made through the medium of interpersonal relations.

By surveying the field in which research efforts have been most numerous and most varied, the Bureau hoped to derive some working principles that
can be adapted to evaluation of the services and programs in which it has the
most direct interest.

Even in the field of psychotherapy, our survey has not been exhaustive.
It has attempted to cover the range of methods and assumptions employed for
evaluating the results of efforts to bring about change in individuals without
aspiring to note every variation and angle or to review every relevant example.

This report is written chiefly for the use of administrators and others
who are considering setting up evaluative research in their agencies or are
wanting to know how much reliance to put on the reported findings of such
studies. The reason for writing it is the belief that a clear idea of the prob-
lems involved and the considerations to be weighed in approaching this kind
of research will contribute to more realistic expectations, more thoughtful
planning, and more effective use of what research technicians and research
methods can offer in our present phase.

This belief has prompted inclusion of chapter III, which is somewhat
more technical than the rest. Those not directly involved in research may
prefer to give chief attention to other parts of the discussion, especially the
first and the last. It is hoped, of course, that other chapters in addition to
chapter III will also be of interest to research workers.

P. Frederick DelliQuadri
Chief, Children's Bureau
SOCIAL and REHABILITATION SERVICE
## CONTENTS

I ABOUT THE STUDY ........................................ 5

What Is the Purpose of the Evaluation? .......................... 5

II ABOUT THE EFFORTS THAT ARE TO BE EVALUATED ...... 9

What Kind of Change Is Desired? ............................... 9

*Change from what to what?* ..................................... 9

  Change from what? ... change to what? ...

  Researchable definitions

Change known by what signs? .................................... 15

Change in whom? .................................................. 26

  Physical characteristics ... psychological

  characteristics . . environmental factors

By What Means Is Change to Be Brought About? ............... 30

  What method is used—in theory? .......................... 30

  What method is used—in practice? ......................... 32

  By whom is the method used? ............................ 35

III ABOUT THE METHODS USED FOR ASSESSING CHANGE ... 37

How Trustworthy Are the Categories and Measures

  Employed? ..................................................... 37

    How reliable are they? .................................. 37

    How valid are they? .................................... 41

At What Points Is Change To Be Measured? ...................... 50

    From what base? ........................................ 50
This page is blank in the original document.
Efforts to bring about social-psychological change in individuals are attempts to help them deal with difficulties they have encountered in social and psychological functioning. Efforts to evaluate ask: have the individuals been helped? This key question, however, is a very unstable compound. Under examination it breaks down into a cluster of questions: which ones have been helped, how much, how stable is the help, was it really the treatment or something else that helped, who says so and how do we know it is true?

Such questions are challenging enough when raised about the effectiveness of a single practitioner. When the reports of many practitioners or agencies are combined or compared, a different kind of question is added. Were they all defining help in the same way? Did they all begin with problems of the same or comparable difficulty? Were the individuals they worked with equally capable of change? Were the improvements noted comparable in kind or degree or stability?

The history of evaluative research shows increasing recognition of the questions that must be answered, increasing awareness that they cannot be answered quickly or simultaneously, and increasing efforts to lay the ground for defining them, setting priorities for them and attacking them in due order.

A review of the literature, reinforced by discussions with research people, shows a rather neat grouping of things on which the "experts" do and do not agree. They agree on the need for evaluative research, on the complexity of the problems it presents, and on the fact that so far no one has solved these manifold problems to the complete satisfaction of himself or anyone else. They agree also that even before its problems are solved, great values are to be gained from the right kind of evaluative research. Some of these values lie in its results, some in the gains derived from the process itself.

---

1 A slightly less cumbersome term, "psycho-social" will be used from here on to describe the kind of change under discussion.
On the whole, the experts agree also about the questions that ought to be answered in any sound evaluative study. The individual researcher does not always answer each one of these questions himself, nor is it always possible to do so. But on being asked, trained and experienced research people are very likely to concede that these are the ones that should be answered. They tend to disagree about the best means of answering them and about what constitutes an adequate answer.

It is now relatively well agreed that a satisfactory evaluation of efforts to bring about psycho-social change in individuals should deal directly with the following questions:

1. What is the purpose of the evaluation? (What is to be achieved by doing it?)
2. What kind of change is desired?
3. By what means is change to be brought about?
4. How trustworthy are the categories and measures employed?
5. At what points is change to be measured?
6. How fairly do the individuals studied represent the group discussed?
7. What is the evidence that the changes observed are due to the means employed?
8. What is the meaning of the changes found?
9. Were there unexpected consequences?

To different degrees and in different ways, these questions are interrelated. Some interlock so closely that one cannot be considered without simultaneously considering the others. Some depend on each other in such a way that one cannot be raised until the other has been settled. Moreover, they are questions of different orders, representing different frames of reference. Questions 4 through 7 are primarily the responsibility of the researcher. He cannot even pose the questions clearly, however, until he has answers to questions 1 through 3 which are primarily the responsibility of the practice field as represented by the agency or organization that initiates the research.

To say that the first three questions are primarily the responsibility of the field is not to imply that they can be answered by the practitioner alone—unless he is also a researcher—for the answers must be in terms that lend themselves to research. A much-bemoaned handicap of the researcher is that he is typically called in too late. By the time he arrives, the administrators or
board are likely to believe they have answered the first two questions. Only if they are willing to start with the researcher from the beginning and work out the painfully slow answers can a solid project be built. However, this process carries its own rewards—rewards that will be discussed presently.

Some questions can hardly be allocated primarily to practice or to research, for they require almost equally divided responsibility. Because of this and of the varied demands raised by all the questions to be answered in this kind of evaluative research, it has come to be almost taken for granted that an interdisciplinary team will be required. This requisite has so recently become an axiom that it deserves further comment, which will probably be more intelligible after discussion of the questions themselves, and will therefore be reserved until later.

Although the questions interlock, each one will be considered separately, with some of the reasons for needing an answer to it, and some of the problems it involves. Lest the array prove too discouraging, two points should be made here: (1) The usual experience is that as much benefit is derived from working out answers to the various questions as could be expected from the wished-for findings. (2) We do not have to wait for a complete answer to each question before satisfying some of the immediate information needs. Both of these points will appear throughout the discussion and will be considered at more length in the final comments. They are brought in here as testimony that the end of this story, though strenuous, is not unhappy.
This page is blank in the original document.
I. ABOUT THE STUDY

What is the Purpose of the Evaluation?

Although the purposes of evaluative research are legion, evaluation of efforts to induce psycho-social change in individuals is undertaken for a relatively limited number of reasons. Nevertheless, it is necessary to be quite clear at the very outset about the purpose of any study—why it is undertaken, what information is sought, how the findings are to be used, and who is to use them.

A few examples of actual requests for evaluative studies of the results secured by social casework will suggest both the range within this limited category and the need for a detailed and specific statement of purpose:

1. Numerous requests come from the field for studies to determine the effectiveness of social casework. In these requests, the stated aim is to secure sound and validated information for the enlightenment and improvement of the profession. This calls for an evaluation of casework as a form of practice.

2. A casework staff wants to know whether it is better for the same or for different caseworkers to work with different members of a family with whom they are in simultaneous contact. The stated aim is to secure a basis for usage by practitioners in a certain agency, and calls for evaluation of outcomes as related to procedures used with a selected group of the agency's clients.

3. An administrator wants to know whether the practice of his staff is up to professional standards. The stated aim is to secure a basis for decisions about personnel requirements and training, and calls for evaluation of the results achieved by the staff of a given agency, as compared with professional norms.

4. A Board of Trustees wants to know whether the service given by their agency merits continuance. The stated aim is to secure a basis for
precisely what staff, time, and funds will be required, but within rough limits an estimate can be made. It is often easier to say what cannot be done within certain limits than what can. For example, probably no honest and experienced researcher would promise to complete within his own lifetime a study fully satisfying the purpose in example 1. The reasons for this will come out in the discussion that follows, as will the variety and worth of the gains that can be made in moving toward fulfillment of that long-term purpose.
II. ABOUT THE EFFORTS THAT ARE TO BE EVALUATED

What Kind of Change Is Desired?

Change from what to what?

It seems fairly obvious that in order to find out whether a desired change has occurred, it is necessary to know what change was desired. Desired psycho-social change in individuals means change from one condition or set of circumstances to another condition or set of circumstances. To define the change requires that both from-what and to-what should be clearly specified.

When one has measles, the change desired is that the specific syndrome of specific symptoms known as measles should disappear, leaving none of the specific after-effects known to be associated with this ailment. If one has a broken leg, the desired change is that the leg heal, with none of the specific after-effects that can follow certain known procedures, omissions, or conditions. In such cases, the desired change requires removal of a specified condition and achievement of a different specified condition.

A major research predicament in evaluating the outcome of social or psychological services is that neither the "ailment" to be treated nor the goals of treatment are defined in terms as concrete and self-evident as these. This is partly because such services deal with ailments and treatment goals more complex, more elusive, more conditional, and more comprehensive than those involved in measles or broken legs. How, asks a psychiatrist, is one to classify "a case (for example) of a person of rigid obsessional character who has strong paranoid trends, presents an anxiety state as the clinical condition from which he seeks relief, and also has some psychogenic physical symptoms which he attributes to having had jungle fever ten years before? In general medicine there is no comparable confusion, for there is not the same attempt to diagnose
the entire physiochemical structure of the patient, nor, furthermore, the same
test to treat many other conditions subordinate to the main illness.”
(173, p. 435)

Change from what?—The difficulty of defining diagnoses and goals is
acutely illustrated in the two fields whose evaluative literature we analyzed.
Neither psychiatry nor social casework has developed a clear and accepted
diagnostic system which would permit definition, in professionally based terms
that command professional consensus, of the specific conditions to be changed
and the specific goals to be achieved.

In psychiatry, investigation has often shown a relatively low degree
of agreement among practitioners in diagnosing the same patients, except
within categories too broad for research utility (11, 71, 78, 219, 289, 327).
Kohn and Clausen comment, for example, that “in the present state of
psychiatric knowledge there is considerable question whether either schizo-
phrenia or manic-depressive psychosis is a single disease of common etiology or
a group of similar appearing diseases of differing etiology.” (182, p. 268) A
prospectus for a research program points out that referrals designated as chil-
dhood schizophrenia have been found to include all varieties of functional and
organic disability. One author (168), lamenting the lack of homogeneity
and reliability in present neuropsychiatric diagnostic categories, declares that
there may be more difference between two schizophrenics than between a
group of schizophrenics and a normal group—and at least one empirical study
seems to support that claim (36, 251).

No systematic tests of diagnostic agreement seem to have been made
for social casework, perhaps because it is so generally recognized that the
diagnostic categories of that field are even less satisfactory than those of
psychiatry. A number of studies have pointed to the need for sharpening the
definition of problems in social casework, and for tapping dimensions other
than those used by current problem classifications. Two, for example, con-
sider the number of clients involved in a case as part of the problem classifica-
tion, with the implication that the nature of the problem and its prognosis
may be different in cases that involve one client as compared with cases
involving several (130, 271).

Although the experts disagree about many things in evaluative re-
search, one point on which there appears to be overwhelming consensus is the
need for more satisfactory classification of the problems toward which treat-
ment or service is directed. Some regard it as the most urgent of all needs

As used in this report, psychiatry is regarded as a specialty included under psychotherapy
and psychoanalysis is regarded as a specialty included under psychiatry. Psychotherapy and
social casework are regarded as two different activities.
for research in psychiatry and in social casework.¹

In social casework, efforts to evolve more satisfactory problem classifications have been made at the Research Center of the Chicago School of Social Service Administration and at the Institute of Welfare Research of the Community Service Society in New York City. More frequent and more sustained efforts have been made to meet the need for better diagnostic classifications in psychiatry. A number of groups and organizations, after setting out to plan evaluative research in psychotherapy or psychoanalysis, changed their plans in order to do research first on developing a classification system adequate to the needs of evaluation. Some projects plan to devote up to five years solely to working out and testing diagnostic categories. These studies include diagnostic categories and also attempt to classify by degree of severity.

Not only do "the experts" agree that better diagnostic classifications are necessary prerequisites to wholly satisfactory evaluation; they also agree that the process of evolving such diagnostic categories will contribute to practice as well as to research. Many psychiatrists and social workers concur in the view that their fields would gain by the conceptual sharpening that would attend the working out of more consistent, significant, and reliable problem categories.

An example of gains to practice derived from improved diagnostic classifications is supplied by the work of the mental hospital administrators and statisticians toward improving the diagnostic classifications used in hospital records. In the proceedings of their third conference, they report that the change from old to new nomenclature brought to light and corrected a number of wrong diagnoses, and that the refined nomenclature promises to increase diagnostic precision. They admit, however, that it may also increase perceptible disagreement between doctors, since—as numerous studies have shown—there is likely to be more agreement under gross categories than under the finer sub-groupings of those categories (224).

Problems of diagnostic classification loom large also in the treatment of juvenile delinquents. Those most involved in efforts to bring about desired change in young people labeled "delinquent" are often concerned about the vast array of problems and conditions lumped under this term; and about the effects of the label itself. Some States have diagnostic centers from which juvenile delinquents are assigned to a variety of resources, according to the diagnosis made. Yet there is little agreement about the classifications used, and they sometimes seem to be determined as much by the resources available...

¹It is difficult to select references on this point because it is made by almost everyone who discusses evaluative research in either field. Among those that could be mentioned are: Carpenter (47), Freyhan (104), Gough (115), Greenwood (117), King (168), Kline (172), Levy (196), Lurie (208), Maas (211), Mehlman (219), Miles and others (227), Thorne (312).
as by the nature of the problems to be dealt with. Almost every conference on juvenile delinquency includes a plea for an adequate typology of delinquency, and there is an occasional warning that such a typology should be diagnostic, indicating the source or the area of difficulty, rather than merely describing the delinquent act as minor or major, involving property or people, etc. (334).

Change to what?—Among the gains to be achieved by sharpened definition of diagnostic classification is a sharpened definition of treatment goals—for to a large extent the goal of treatment or service is implicit in the diagnosis. If the diagnosis is measles, the goal is cure of measles; if it is a broken leg, the goal is restored use of the leg. Even with such relatively clear-cut ailments, however, the goals may become complicated and conditional. If the individual who breaks his leg has a bone disease, or is diabetic, or is ninety years old, the outcome of therapy might be judged by standards different from those that would apply if he were a healthy active boy of twelve. The goal of therapy, then, is implicit in but not fully defined by the diagnostic classification of the conditions or circumstances that are to be changed.

These examples bring out the need for progressive sharpening and differentiating, as one moves from the goals of a profession or program to the goal of treatment or service for a specific individual. There is a good deal of divergence in the way even the broad professional goal is stated either for psychiatry or for social casework. Disagreements flourish between schools of thought and also between individuals within any one school or group—and divergences often exist even between those who think they are in agreement. One attempt to express the broad goal is represented by the statement that these relationship therapies aim to help an individual live with more pleasure and less pain to himself and others. For research purposes, of course—and for practice also—a statement as broad as this requires specific spelling out, so that the extent to which the goal has been achieved can be tested and compared with other treatment results. Part of the spelling out involves criteria discussed in the following section. Unless the goal is made explicit, there is no basis for saying whether and to what extent it has been reached—in other words, no sound basis for evaluation (6, 108, 186, 212, 259, 287).

In practice, adequate spelling out of a goal usually requires that it be related both to the diagnostic classification and to the characteristics of the individual involved. The more the practitioner knows about both, the more realistically can he estimate what might be accomplished by therapy or service. Accordingly, the ability to limit goals as well as to define them is associated with professional growth. One sign of progress noted in a review of psychiatric therapies is the increasing realism of therapeutic goals, "in not expecting a complete reconstruction of every patient, and in accepting more
modest therapeutic gains.” (134, p. 245-246) In social casework also, the ability to accept limited goals is often cited as a sign of experience and professional maturity. The ability to recognize and define the limits in advance, and to adapt the treatment to their requirements, depends on the ability to make an accurate diagnosis.

Unable to accept limited goals is often cited as a sign of experience and professional maturity. The ability to recognize and define the limits in advance, and to adapt the treatment to their requirements, depends on the ability to make an accurate diagnosis.

**Researchable definitions.**—To define the change required, then, means to define the conditions which should be altered and the kind of alteration desired—that is, the diagnostic classification and the treatment goal. For a research project, the goal must necessarily be stated more broadly than for treating an individual. How broadly will depend on the purpose of the research. If it is to inquire into the effectiveness of psychiatry or social casework, then an over-all professional goal would have to be stated. If it is to inquire into the effectiveness of practice with one type of ailment or problem, the goal would need to be spelled out on that level. However broad or specific the goal, it must be spelled out clearly enough so that it is possible to determine whether or not it has been achieved. Until we can spell out in testable terms what should be different and what the nature of that difference should be, we are not in a position to know whether or to what extent the desired change has been effected. Nor are we able to compare the results of different practitioners, agencies, or methods. For in order to compare effectiveness, one must be able to compare the nature and severity of initial disorders or problems and the nature and degree of any changes that were achieved.

The working out of sound, usable categories of diagnosis and treatment goals is a major undertaking prerequisite to definitive evaluation. It is essential also to the fullest development of practice, since identification is characteristically the first step toward successful treatment.

A good deal of psychiatric research at present is being directed toward this question of diagnostic identification and classification. One important lead that has emerged is the idea that diagnostic categories probably should be “multidimensional.” Instead of seeking a single label such as “schizophrenia,” or in social casework, “marital problem” to characterize a case, it may be necessary to combine a number of aspects to represent the diagnosis. Severity of problem or amount of impairment, for example, may be indicated apart from type of problem. One study suggests that two separate factors are required to describe severity of a psychiatric problem (199). No doubt a larger number will need to be accounted for in defining problem type and subtypes.

Must all evaluative research, then, wait until adequate classifications have been evolved? The answer is that obviously it cannot all wait—and even if it could, it wouldn’t. Although analogies are often deceptive, the pattern set by the medical field is useful here. In recent years, for example, different types of treatment for polio were being undertaken and evaluated and the
results of these evaluations were being applied in practice. These evaluations were necessary to the practitioner. At the same time, however, efforts were under way to identify the virus that causes the disease—a virus that after long and costly research turned out to be not one but three types. The earlier evaluations of treatment based on less precise diagnosis were indispensable to the profession at that stage. But the refinement of diagnosis made possible by more precise definition of types helped practice to move into a more effective phase. Similarly, it is widely held that a sound typology of juvenile delinquency would pave the way for more effective treatment methods as well as for more adequate evaluative research. Nevertheless, efforts continue to evaluate the methods we have with the classifications we have, assuming that both methods and classifications will improve together.

The examples point up again the extent to which the purpose of the research determines the level on which each of the evaluative questions must be answered. If the purpose is to secure a basis for an administrative decision, it is likely to have a built-in time limitation. Moreover, in such a case it may not be necessary to spell out the diagnostic classifications with the precision required by another sort of purpose. If a Board of Trustees wants to decide whether the service given by its agency merits continuance and is convinced that an evaluative study will help in the decision, it may accept as a “given” the group of clients served by that agency, with the problems they bring to it, using the problem classifications with which it is familiar. It will then be able to work out methods for deciding whether the apparent value of the services, viewed against their costs and the amount of need for such services in the community, is worth while. In such a case, however, the results would be applicable only to this particular agency in this particular community and could not be generalized to other agencies or clients or communities or purposes. Moreover, the researchers would be under strict obligation to make very clear the stringent limitations of the study and of any conclusions that could be drawn from it.

On the other hand, if the purpose involved a comparison between the results of two agencies, or two types of therapy, or two therapists, it would be necessary to know both the kind and severity of the problems involved. Again, sound diagnostic classifications would be quite indispensable if the purpose was to evaluate casework or psychoanalysis as a form of treatment. The groups mentioned above, as they started evaluative research in psychiatry, discovered that for their purpose—increase of professional knowledge and enrichment of professional practice—it would be necessary to do research on diagnostic classifications before they would be in a position to move toward evaluation. For any evaluation that involves generalizing beyond the agency or therapist whose results are immediately under investigation, operational definitions of diagnostic classifications and of goals are highly important.
This amounts to saying that the purpose listed under example 1 on p. 5 will never be fully achieved until a good deal of research has been devoted to diagnostic classifications. But in saying so, it is well to remember that professional practice benefits directly from such research, quite aside from its contribution to ultimate evaluation of the results of practice.

**CHANGE KNOWN BY WHAT SIGNS?**

Once the change that is desired has been defined, the next question is, how do we know whether or not it has taken place? The answer offered by any evaluative study lies in its criteria—the signs that change has or has not occurred.

If the criteria are sound, clear, and feasible, the findings produced by them can be trusted, providing the application of the criteria is equally sound. If not, the findings must be challenged. Accordingly, it is a research axiom that no study can be better than its criteria—although, unfortunately, it can be a good deal worse.

Adequate criteria must afford convincing evidence of the extent to which goals have been reached. Thus, the objectives of treatment—and to some extent also the diagnostic classifications of the conditions to be changed—are inherent in the criteria of its effectiveness. These criteria are, in fact, the concrete spelling out of the change that is desired.

Adequate criteria must also be practical for research. A chronic heartbreak for the evaluator is the frequency with which significant criteria must be abandoned—either because they do not lend themselves to convincing verification or because the information necessary to apply them is not available. The unavailability may result from inadequate information or from the nature of the criterion involved. For example, many studies have had to relinquish a comparison of condition at beginning and at end of treatment because full and relevant information was not available about the nature and severity of the illness or problem at the beginning. This in turn might arise from inadequate initial diagnosis, incomplete recording, or lack of clear categories for describing the individual's initial status. As one researcher has sadly put it, "We often have to choose between the significant and the feasible." Few are as disarmingly frank as the author who observed that "the criteria finally chosen... were to an extent determined by the data in our files." Yet even fewer would claim that the criteria employed represent a free choice based on direct application of fully developed theory. And many recognize the need to strive more successfully for criteria that will be significant rather than easy (201).
A different reason for discarding desirable criteria may be that they do not lend themselves to the degree of precision desired. Unfortunately, it is often easier to be exact about minor than about major factors. It is easier to say how much more a person is eating than to say precisely how much, if at all, his anxiety has diminished. Another familiar lament in research is "We are caught in a dilemma between the significant and the exact."

Recognition of criteria problems has grown with the development of evaluative research, and their acuteness has grown with recognition. The history of the word criterion is an interesting companion-piece to the history of the quest for adequate criteria in evaluative research. Webster's dictionary gives a two-pronged definition of criterion: "a standard of judging; a rule or test by which anything is tried in forming a judgment respecting it." The criterion is the standard to be met; it is also the sign or test by which one determines whether that standard has or has not been met. Sometimes, to avoid confusion, criterion as standard is referred to as ultimate criterion variable and criterion as test is referred to as immediate or intermediate criterion variable. More often the two meanings are merged, nor is it always necessary to differentiate between them in talking about criteria—although in using them the distinction is inevitable. If we say, for example, that no study can be better than its criteria, the comment embraces both the ultimate criterion variable and the indicators by which its presence or absence is established.

The dictionary definition also notes that the word, taken over directly from the Greek, is descended from the word for "judge," which in turn is derived from a word meaning "to separate." Thus, built into the name for this crucial research component, is the recognition that judgment involves differentiation. The quest for adequate criteria in evaluative research has led toward ever-increasing differentiating or separating of elements from each other, so that in a sense the history of the criterion in evaluative research dramatizes in reverse order the etymology of the word the Greeks had for it.

The first efforts to evaluate the effectiveness of psychotherapy, like the first efforts to evaluate the effectiveness of social casework, usually offered one over-all judgment about the results of treatment. Such over-all or "global" evaluations might be in relative terms, e.g., the degree of improvement since the beginning of therapy; or in absolute terms, e.g., level of adjustment at the end of therapy. The characteristic of the global evaluation is that it arrives in one step at one sweeping judgment to cover all aspects of treatment outcome. It is a one-step application of the ultimate criterion variable.

At first it was considered enough for some qualified person to judge whether or not an individual was adjusted, cured, or helped. It soon became evident, however, that this type of global evaluation left much to be desired. Just what does improvement or adjustment mean? What are its ingredients? Are they the same for all cases? How are they recognized? Whose word
shall we take about them—the patient’s, with his stake either in proving he is well or proving he is sick? The therapist’s, with his stake in showing the success of his therapy? The community’s, with its stake in keeping the patient from doing damage or from becoming an economic burden? And even if we rely on disinterested experts, how consistent are their judgments?

Accordingly, efforts were made to single out the criteria of adjustment, cure, improvement, etc.; to spell out definitions that would carry the same meaning for different observers and to test out these definitions by having them applied independently by more than one person. Such definitions, to be adequate, had to rely on manifest evidence and not only on undocumented opinion; they must be “behavioral” or “operational.”

Thus the original global evaluation became segmented into parts. From this point it was only one step to what might be called a segmental rather than a global evaluation. That is, the intermediate or immediate criteria of adjustment or improvement, or whatever global term was employed, would be rated separately. They might be added up at the end into some weighted score or some defined level of adjustment or improvement; or on the other hand, they might be reported separately with no attempt to pull them together into a final over-all judgment of outcome. In such a case it could be said, for example, that 40 percent of the patients or clients showed a certain degree of improvement in job adjustment, that 35 percent showed a certain improvement in presenting symptoms or problems, that 50 percent showed a certain degree of improvement in family relations, etc.

This trend toward differentiating the criteria by which outcome could be judged moved from the abstract to the concrete, from the whole to its parts, with the parts becoming ever more limited, specific, subject to verification. This strong trend suggests a paraphrase of an old jingle:

“Big criteria have little criteria upon their backs to bite ’em
The small ones have still smaller, and so on ad infinitum.”

In the countless efforts to answer the nagging question “how do we know,” with all its corollary questions, a great variety of criteria have been employed in evaluations of psychotherapy. One author (317) has identified “upwards of a hundred criteria used singly and in combination,” and has by no means covered them all. Without any attempt to be exhaustive, some examples are listed below of the various types that have been employed in evaluative studies of psychotherapy or social casework. They are roughly grouped for convenience (although to some extent the various groups overlap) under the two main types mentioned above—namely, criteria employed for a “global” evaluation and those employed for a “segmental” evaluation, which may or may not be incorporated later into a global one.
GLOBAL: Ultimate Criterion Variables (May be used with or without specific indicators or operational definitions)

**Absolute**
Level of adjustment or adaptation; e.g.: good, fair, poor, not stated
Degree of mental health, e.g.: place on health-sickness scale

**Relative**
Therapeutic success or failure; e.g.: successful, partially successful, unsuccessful
Cure or improvement; e.g.: apparently cured, much improved, moderately improved, no change
recovered, improved, unimproved, dead or lost
definite improvement; partial, none, no treatment attempted
much improvement, slight, none, decrement
Degree to which problem is solved; e.g.: solved, partially solved, unchanged, worse
Movement toward treatment goal; e.g.: much, moderate, little or none, retrogression; place on movement scale

SEGMENTAL: Intermediate or Immediate Criterion Variables

**Psychological traits or conditions;** degree of or changes in:
- anxiety
- nervousness, tension
- frustration or satisfaction of "natural needs and drives," e.g., sex
- insight, awareness
- dependency
- attitudes toward authority
- self-control
- defensiveness
- breadth and depth of interests
- maturity
- integration
- organization of personality; basic personality structure
- growth and development
- perception of reality
- response to reality (problems, conflicts, crises)
- effectiveness, ability to utilize capacities
- inner vs. other-directedness; autonomy; locus of evaluation

**Expressed attitudes and opinions**
- of patient, or client, concerning:
  - self (real and ideal)
others
therapist
therapy
his present condition
_of therapist, concerning:
status of patient or client
overall results of therapy
_of collaterals, concerning:
status of patient or client
results of therapy

Clinical findings
manifest symptoms:
reduction in or disappearance of
greater tolerance for or ability to cope with
increase in, appearance of or substitution of new ones

Social and economic indices
interpersonal relations:
within the family
outside the family; e.g.: job, school, social group, neighbors, public, etc.
role changes
employment, productiveness, wages
school performance
other achievement, recognition

Events
admission, discharge, re-admission to mental institution
admission, discharge, re-admission to correctional institution
court appearance, police involvement

Physical health
general condition
specific symptoms

Some of these criteria can be used at different levels and therefore might logically appear under more than one heading. For example, in one study a global evaluation of therapeutic success or failure may be made, with the patient’s adjustment as the chief criterion of success. In another, a global evaluation of the patient’s adjustment at the termination of contact may be the end result, and a number of criteria of adjustment may be employed—e.g.: family relations, job productivity, reduction of anxiety, etc., etc. Yet another study may produce segmental evaluations of the various components of adjustment, reporting separately the client’s situation at the end of treatment with
regard to job, family, inner tensions, etc., but not attempting to combine
them into an overall score or rating. In this case, the criteria of job adjust-
ment, family relations, anxiety would be spelled out in more concrete detail
than in the case where a broad judgment of adjustment is used as the criterion
of therapeutic success. Thus the ultimate criterion variable for one study
may be the intermediate criterion variable for another. At each level the
question "how do we know" calls for further evidence.

The multiplicity of the criteria used reflects, among other things, the
lack of clarity and agreement discussed in the preceding section. Again and
again the wail resounds through discussions of evaluative research in psy-
chotherapy: "There is no agreement about what cure is, there is no agreement
about what constitutes improvement!" Watson, after counting above a
hundred, echoed a common cry, "At present we are in the unhappy state of
not knowing what are the criteria of effectiveness of psychotherapy . . . re-
search has not yet isolated criteria on which there has been any sort of general
agreement concerning their value as indices of improvement." (317, p. 31)

Yet lack of definition is as much a result as a cause. The vast array
of criteria that have been tried reflects also the breadth, complexity, elusiv-
eness, and infinite variety of problems for which change is sought through
some form of relationship therapy, and of ways in which evidence of change
may be manifested. One hope is, of course, that greater differentiation of
diagnostic classifications and of goals will narrow the range of possibilities for
a given individual or group of individuals. No sharpening of goal or diag-
nostic classification, however, will alter the fact that treatment outcome in-
volves constellations of factors and of mechanisms in which a given element
may mean different things at different times or under different circumstances.
Because of the myriad elements involved and the multiple meanings each
can have in different patterns or processes, it is futile to count on discovering
one simple forthcoming litmus indicator for effectiveness of treatment.

A recent press release on "examples of progress" reflects the proverbial
elusiveness and complexity of criteria for the effectiveness of psychiatric treat-
ment. Progress with venereal disease and tuberculosis was reported in terms
of the reduced mortality rate for those ailments. Progress with mental health
was reported in terms of extended mental health programs and services. In
other words, venereal disease and tuberculosis were reported in terms of mea-
sured results; mental health in terms of the efforts expended. This contrast
speaks volumes, and volumes have been written about it. The rates of mental
illness cannot be used as criteria for program success since, as one report puts
it, "much remains to be determined as to standards that will be employed to
categorize individuals as 'sick' or 'well'. Until this is accomplished . . . any
figures showing prevalent rates or percentages are extremely tentative," and
"the rate obtained depends heavily on the method used." (192, p. 723)

Provided by the Maternal and Child Health Library, Georgetown University
The quest for adequate criteria has been described as an evolution, and in theory this description holds. In practice it requires sharp qualification, for it is not a neat evolution; its phases overlap. There has been a stage-by-stage development in the recognition of demands that must be met by adequate research, even though in practice these demands are not always satisfied. Today, many take for granted what was formerly assumed by very few—that if the criteria employed do not meet the most rigorous standards, then at least the extent to which the results can be generalized is sharply reduced and at most the extent to which they can be trusted at all is in question. This evolving recognition has opened the way to new research practices and emphases. And although it has not eliminated some of the old ones, it has influenced the ways in which any study must be carried out and interpreted.

It would be misleading, however, to suggest that because of the evolution in research methods, “global” evaluations are no longer attempted or are no longer justified. If the purpose requires and justifies a global evaluation, it should be used. But it will have to be used with due recognition of its limitations and with due respect for rules of evidence as currently conceived. Once more, then, the all-important question of the study’s purpose is underlined.

If the purpose is what we are calling “ultimate evaluation” (e.g., does psychiatric treatment really help and how much?) then there can be no compromise with exigencies of time, money, and staff. This kind of evaluation requires ability to compare the results of different methods, agencies, and individuals in treating different kinds of patients; and to compare the results of any treatment with the results of no treatment. Such evaluation is blocked until extensive research has been done on criteria of outcome as well as on other research questions discussed in this report. Many projects must be devoted to studying the criteria of outcome and the different ways in which they can combine to produce results of varying satisfactoriness.

Many researchers believe that premature attempts at global evaluation have impeded the development of adequate criteria. “We have allowed our concepts to become obsessed by the idea of wholeness, devaluing any part-observations because they fall short of the goal of global understanding,” says one paper on the subject (193). The authors go on to observe that a different approach has proved fruitful for the natural sciences. “The present body of chemistry, physics and animal psychology is a structure built of many small pieces of knowledge.” (193, p. 55)

This view is part of the trend already noted among researchers interested in ultimate evaluation—a trend of interest away from evaluative research per se and toward “pre-evaluative” research that will furnish the technical tools prerequisite to definitive evaluation of types and methods of therapy. Thus, for example, the Rogers group in Chicago have stated that they are
concerned, not with therapeutic success or failure, but rather with the "con-
comitants of therapy" which must be studied before success or failure can be 
assessed. An entire study may be devoted to a single intermediate criterion,
such as change in the patient's acceptance of himself, or change in his attitudes 
toward authority, or change in the extent to which his reactions are based 
on his own perceptions and convictions rather than on his ideas of what others 
think ("locus of evaluation"). Such studies do not ask whether therapy has 
helped, but instead inquire into the nature of some specific change which 
seems to be related to therapy and which, once tested, may eventually furnish 
one among many criteria of outcome (69, 113, 267, 296, 304).

On the other hand, some purposes can be served without waiting for 
extensive research to be done on the various criteria that will help to secure 
a definitive answer about the effectiveness of treatment. This is fortunate, 
since many decisions cannot wait until exhaustive research has been done on 
all the pre-evaluative problems that remain unsolved. In some instances, 
very crude and imprecise criteria can give a needed answer. To take an 
illustration from the field of delinquency, the crude criterion of recidivism 
may demonstrate that present methods of treating juvenile delinquency are 
not good enough to satisfy us, even though the criterion would be wholly 
inadequate to establish an effectiveness rate for current methods.

Present possibilities, however, allow for criteria far more satisfactory 
than this. Careful ratings by caseworkers, for example, may furnish strong 
presumptive evidence that a substantial proportion of clients in a certain 
agency are better off after contact than they were before, even though the 
criteria are not sufficient to prove how much better these clients "really" are, 
to what extent the casework contact is responsible for improvement, and to 
what extent the effective element was "casework" rather than some unique 
feature of practice in this one agency.

The primary obligation, then, is not to wait until absolute and irrefut-
able criteria have been worked out, but rather to meet squarely and fully the 
crucial criteria questions in planning, executing, and reporting a study. These 
questions are:

1. What are the criteria?
2. How are they defined?
3. How are they applied?
4. What limitations on generalizability of results arise from limitations 
in the nature and application of the criteria?

1. What criteria? It is clear from the present discussion that the 
selection of criteria merits study in itself, even though the project is not a
"pre- evaluative" study of criteria. This is one point where above all haste makes waste. The criteria selected must reflect the change that is sought. They must be significant rather than merely easy. At the same time they must be practical for research within the limitations of the proposed study.

Almost any criterion proposed will be found to have its own problems and defects. For example, in psychiatric research the relief of clinical symptoms has become suspect because new and more illusive symptoms may replace those that appear to be cured. Discharge from a mental hospital is a dubious criterion since it may be influenced as much by the attitude of the patient's family and their readiness to care for him as by his ability to be on his own. Thus, a patient who is much improved may be retained in an institution while one who is very slightly improved may be discharged. A further difficulty with this particular criterion is that in some institutions a patient may be officially discharged from the books as much as a year after release; in others the stipulated period varies and in still others official discharge is automatic upon release if no ill report is returned during the trial period (224, 272).

The opinions of the individuals most involved (the patient and his relatives) have been defended as criteria on the ground that these are the people who ought to know most about it; and attacked on the ground that they are the most interested parties. Some point out that the therapist has a strong stake in the results and therefore cannot be an unbiased judge; and also that his perception of change may result from increased information about his patient rather than from change on the part of the patient (253, 318). There are claims too that relationship therapy is a learning situation and what the patient learns may be how to say the right thing to the therapist rather than how to change his attitudes or behavior. The weight of each argument is affected both by the way the criterion is to be applied in a given study and by the bias of the commentator.

Since no criterion is likely to be free of pitfalls, the best defense is to be aware of the ones that exist, of the extent to which they are likely to bias results, and of the direction this bias will probably take—i.e. will it tend to make the outcome look better or worse than it really is? The process of selection itself is designed, of course, to find the criteria with the most advantages and the least disadvantages.

2. How defined? The quality of the criteria depends as much on the way they are defined as on the way they are selected. To get consistent and convincing application always requires clarity of definition and often needs in addition a certain amount of training and practice on the part of those who are to apply the criteria. If the opinions of laymen or experts are to be used as outright opinion material, the definitions are likely to employ evaluative
words, indicating degree of disturbance or improvement. Often, however, an attempt is made to work out definitions that describe rather than evaluate; that is, to avoid words like "desirable member" of a family or community, "good social adjustment," "better behavior" and instead to spell out the kinds of membership, adjustment, or behavior that might be characterized as desirable, good, better or their opposites (256). As far back as 1935 a practical and much used adjustment scale defined five levels of adjustment at the end of outcome in "operational" or "behavioral" terms, characterizing each one by descriptive accounts of the individual's functioning with regard to friends, work in school, job, home, and whether presenting problems had disappeared, or new ones had appeared. That is, it attempted to describe rather than to evaluate how the client was functioning in these different areas. The results of these descriptions were then used to evaluate over-all adjustment (335). Similarly, the Hunt movement scale uses "anchoring illustrations" which describe the behavior and functioning characteristic of different points on the scale (147).

A more recent example of the attempt to define criteria in descriptive rather than judgmental terms is a "health-sickness rating scale" that uses "largely descriptive" criteria to define seven dimensions used in rating on a 100 point scale—dimensions such as the degree of the patient's subjective discomfort and distress; the degree to which he can utilize his abilities, especially in work; the breadth and depth of his interests, etc. (321).

In each of these cases, the definitions are descriptive with a minimum of judgmental or evaluative terms. The various descriptions may then be placed on an evaluative scale (213).

3. How applied? The criteria used in evaluative research, of which examples have been given above, may be applied by:

(a) ratings
(b) measures
(c) a mixture of (a) and (b)

(a) Ratings (as the term is used in this report) represent an attempt to standardize opinions—that is, through careful definition and training to make sure that the terms used are applied consistently by different raters and also by the same rater at different times. Ratings may relate to change per se or to the individual's characteristics, status, or performance at some given point in time (e.g. beginning, end, and after therapy) or to concomitant situations or circumstances. Such ratings may be made by the individual for whom change is sought, by the therapist or independent professional raters, or by "collaterals" such as family, friends, boss, colleagues, neighbors, etc. The
extent to which such ratings are standardized varies widely in evaluative studies and will be discussed further under "reliability."

(b) Measures represent an effort to escape the elements of bias and impressionism introduced by opinions, however standardized, and to rely on facts that can be checked by concrete evidence of some sort. Psychological test-retest techniques represent one kind of measure. The actual results of psychological tests—the scores they produce—are matters of fact. There may be skepticism about the reliability or validity of the test—questions to be discussed later. There may be skepticism about the competence of the individual who administers them or the extent to which the test situation affects the results. But the outcome is a measure which can be compared with other results of the same test, made under similar conditions.

Measures may also be physiological and organic. The pulse, blood pressure, urinalysis, eye flicker are matters of fact rather than of opinion—although their significance or the skill with which they are obtained may be subject to dispute.

(c) Some criteria are applied by a mixture of ratings and measures—although often even the researcher forgets that rating is involved. "Verbal behavior counts" are often of this mixed variety. For example, an analysis may be made of the frequency of certain words reported in verbatim case records. Dollard devised a "Distress-Relief Quotient" that is derived from a count of the number of word units expressive of distress or of relief that are used by the client (41, 79, 164). Although originally applied to a case-worker's summaries of case records, the system has also been used with verbatim records, which makes it more reasonable although still hardly convincing. The unit count represents a measure. The decision whether a given unit represents distress, relief, or something else is a rating. The combination of rating and measure in this type of content analysis is even more clear when the units counted are those judged to represent "maturity" or "immaturity," "self-acceptance" or "self-rejection." (137, 304)

4. How limited? The criteria questions listed above must be met not only in planning and executing but also in reporting a study. It is necessary to state clearly how the criteria were selected, defined, applied, and—if appropriate—how they are combined. It is equally necessary to recognize the limitations inherent in the criteria in presenting interpretations, conclusions, and recommendations. Some studies which employ "legitimate" criteria in a sound and systematic way fall down in interpretation by generalizing beyond the proper limits of their data. At the risk of boring repetition, this point needs to be reiterated until research practice proves that it has been made once and for all. The purpose of a study may justify some compromise with
the most elaborate requirements of "ultimate" evaluative research. There is no justification, however, for failing to recognize and to state the nature and limitations of the criteria employed and to tailor any interpretations or conclusions to fit these limits.

**Change in Whom?**

For certain kinds of evaluation—especially in research that is primarily administrative—it may be enough to say that the individuals to be changed are all the clients or patients of a certain agency or a certain practitioner during a certain period. If such a study will meet the research needs, the problem is greatly simplified.

Even the simplest needs, however, require some description of the individuals included. If one takes "all clients" during a certain period, does that mean all who applied during a given time, all who closed contact during that time, or all who had active service during that time? Does one equate those who had two sessions with those who had twenty? Does one include those who discontinued treatment against the advice of the practitioner? If so, one reduces substantially the optimism of the findings; yet there are those who hold that not to do so is a distortion. Hunt, for example, remarks that "as surveyed by Knight the percentage of at least 'somewhat improved' is either 92 or 66 depending on whether or not one eliminates from the denominator those patients who discontinued treatment against the psychoanalyst's advice. Wilder argues, and I would agree, that they should not be eliminated from the denominator." (146, p. 2; 174; 325)

Again, agencies differ in their definition of active service, some beginning with the first intake interview and some dating active service from the end of exploratory study. Mental hospitals differ widely in their definitions of such terms as "first admission," "re-admission," "relapse," "discharge." Adequate description must of course include definitions of such terms. And if figures from different agencies or institutions are combined, the definitions employed must be identical for all, no matter how simple the study. For this reason, a great deal of time, effort and money is being spent to standardize the use of such terms in the field of psychiatry (4, 222, 223, 224).

These are the most elementary and superficial aspects of identifying the individuals in whom change is to be effected. A host of other aspects may require consideration, depending on the purpose and nature of the evaluation. For some purposes it may be enough to give a relatively sketchy description of the individuals in whom change is desired; for others it is necessary to include all the characteristics which are known or suspected to
influence the results of the treatment measures applied. Some technical reasons for needing different levels of description are discussed in chapter III. At this point we are concerned with the details required for full description—namely, the characteristics known or believed to relate to an individual’s capacity for change.

In psychiatry, the list of such characteristics is large and continues to grow as new studies suggest new relationships between the physical, psychological, and environmental characteristics of individuals in whom psychosocial change is sought and the effectiveness of the efforts to bring about such change. For example, a review of 1,500 articles and books dealing with schizophrenia has identified “some 40 factors for which a consensus exists regarding their prognostic value.” If these 40 factors are in fact significant, then they constitute part of the answer to the question, “who is to be changed?”

The following groups of factors have been found by some investigators to be significantly related to psychotherapeutic outcome.

*Physical.*—Among the most commonly accepted physical characteristics believed to relate to the results of therapy are sex and age. Most studies take these into account and consider them in selecting and describing a sample. Suspicions are increasing, however, that the results of psychotherapy are influenced by a variety of constitutional factors. If this is so, then the individuals for whom change is sought must be described in terms of these factors also, if any convincing comparison or broad generalization about results is to be made.

*Psychological.*—Psychological characteristics are also widely believed to influence the efficacy of psychiatric treatment, and a number of studies seem to support such a belief. Among the psychological attributes that have been suggested in this connection are ego strength, intelligence, capacity for insight, personal integration, attitude toward self, attitude toward therapy or help. Selection of treatment source and method may be correlated with psychological characteristics. There may be psychological differences between people who select one type of treatment and those who select another; or between those who accept any treatment and those who do not, or between those who drop out and those who continue. Moreover, the kind of treatment offered may depend on personality traits; for example, “an important factor to consider in contrasting psychoanalytic treatment success with the outcome of other modes of psychotherapy, is that only patients who are

---

judged initially to show certain strengths and ego assets are recommended for psychoanalysis.” (204, p. 272)

An important cluster of psychological characteristics relate to the specific nature of the change sought. The need for adequate diagnostic classifications in order to establish the existence, the nature, and the degree of change has already been discussed. These classifications are necessary also for describing the individuals in whom change is sought, since some kinds of illnesses or problems are more responsive to treatment than others. Accordingly, the individuals for whom change is sought must be described in terms of the nature and severity of the illness or problem involved. “Of what value are conclusions from experiments purporting to study ‘psychoneurotics’ or ‘schizophrenics’ when these terms have not yet been sufficiently delineated so as to provide homogeneous groups?” (310, p. 35)

The individual’s capacity for change is likely to be related also to the history of his difficulty. Factors that have shown significant relations to outcome are: duration of the illness or problem, previous history with regard to it, the manner in which it first became evident, type of symptoms, precipitating factors, the advantages and disadvantages it offers the person, the degree to which it cripples his life.

Environmental.—These individuals should be described also in terms of environmental factors and life situation. A number of studies have revealed marked relationships between socioeconomic status and the incidence of psychiatric disturbance, the diagnostic classification of such disturbance, the type of treatment offered, the duration of treatment and its apparent success (153, 138, 212, 245, 273). Casework agencies have found repeatedly that higher income in their clients was likely to be associated with longer duration of contact and with more favorable evaluation of outcome by the caseworkers. The role of cultural background is also gaining greater recognition. A number of studies have discussed the cultural factors that influence response to therapy or service, and also the ways in which awareness of cultural elements can help the practitioner (211, 258).

Other aspects of the life situation appear to be equally significant. The attitude of the family has been mentioned as a factor influencing the time of discharge from a mental hospital. It may also be a factor in the rapidity and stability of post-treatment progress. An inquiry into this question found, rather surprisingly, that patients who were returned to “poorer” homes tended to have favorable outcomes, while those returned to “better” homes (judged by current mental health standards) were more likely to return to the hospital after convalescence. The investigators suggested as a possible explanation: (1) greater pathology of patients returning to “better” homes (that is, the more sympathetic and indulgent family would not have a patient committed
if he were not in extremely bad shape); (2) greater tolerance of "poor" environment for behavioral deviations; (3) peculiarities of the particular sample under study. A fourth possibility, not proposed, is that the home environment rated "poor" according to current standards is more tough-minded than that given a "better" rating; and that the tough-minded approach may be more tonic for the patient.5

Attitudes and behavior of friends and colleagues may also influence prognosis, as may the presence or absence of stress situations. In fact, some go so far as to say that the environmental influences can be stronger than the psychotherapeutic ones and can obstruct or assist their operation; and that the battle of these contending forces and the strategy of coping with them have not been studied sufficiently (133). "It is striking," comments one of the more serious investigators, "how the clinical picture correlates with various life situations and stresses. . . . We do not, of course, imply that the ultimate cause of the neurosis lies in these correlates." (227, p. 89)

For all the factors listed above, some evidence exists pointing to a significant relation with therapeutic success. For most of them, as noted, a professional consensus exists. How many of them it is feasible or even desirable to include in describing the individuals dealt with in a particular evaluation of psychotherapy depends on the purpose and nature of the study.

If different methods, agencies, or practitioners are to be compared, it must be possible to describe the patients in relevant terms. Lacking such description one cannot be sure that they do not differ in respects which affect the outcome more than the treatment method does. Moreover, in comparing any group of patients or clients with a group who received no treatment at all, one would not be sure whether observed differences in improvement rates were related to treatment or to differences between the groups. This kind of comparison, which is a long-term goal of evaluative research, involves consideration of control groups, discussed in a later section (p. 62). The same key question underlies and links the problems of sample and of control: what are the significant characteristics of the individuals in whom change is sought?

On the other hand, the purpose may not require comparison of different groups, or comparison of the treated and the untreated. In this case, a much less elaborate description of the individuals treated would be necessary. Evidence about the kinds of prognostic indices listed above might, in fact, be part of the research findings. After applying the study criteria to determine what proportion of individuals had changed in the ways desired, an analysis

---

would be made to discover what patient characteristics seem to be associated with success or with failure.

Whether the treatment or service under study is psychotherapy, social casework, or services to juvenile delinquents, the general principle would be the same. If comparisons between groups, methods, or agencies are to be made, it must be shown that those who are to be changed do not differ in ways that affect the outcome more than does the treatment or service through which change is sought. In juvenile delinquency, for example, a number of variables have been cited as prognostic indices—such as, sex, age, number of previous offenses, type of offense, parent-child relations, peer relations, I.Q.

In comparing two probation services, one would have to account for these variables in each group to make sure that any difference in outcome was not due to difference in the incidence of these important factors rather than to difference in the probation services. On the other hand, if results from only one group are studied, it may not be necessary to control such factors in advance, or even to know all of them. In such a case, information about the factors that influence treatment outcome may be a major research finding—information that will be useful for future comparisons between groups.

By What Means Is Change to Be Brought About?

What Method Is Used—in Theory?

It is not enough to specify what change is sought. There must be equal clarity about the means by which it is to be effected. In the kind of evaluative research we reviewed, the means was psychotherapy or social casework. A label, however, is not a workable definition, and satisfactory research cannot proceed without workable definitions of all categories employed. For certain purposes it may be enough to say, we shall evaluate the results produced by the kind of casework practiced in a particular agency or the psychiatric treatment given by a particular clinic or the kind of psychoanalysis practiced by a particular psychiatrist. To make such a decision is to lump together a good many incomparables—but if this suits the defined purpose of the evaluation, it can be done. In that case, the means is defined by the individual or organization; it is whatever Agency X does or whatever Dr. Y does.

Often, however, the purpose is broader. People may want to know, not only about the efficacy of Agency X or Dr. Y, but also about the usefulness
of casework or of psychoanalysis. The moment such a question is raised for research, it becomes necessary to use a workable definition of casework or of psychoanalysis. Most practitioners, if pressed, can produce a definition satisfactory to themselves. But to produce consensus among a large group on a precise and workable definition of their treatment methods is more difficult.

A case in point is the American Psychoanalytic Association’s Committee on the Evaluation of Psychoanalytic Therapy. They soon discovered that they could not agree on a definition of psychoanalysis. Yet, as they put it, “In order to evaluate a subject, one must first know of what that subject consists and since apparently there were no two individuals, not only of the Committee but of the society as a whole, who would agree to a definition of psychoanalysis, the Committee was at a loss as to how they were to know just what they were evaluating.” (5, p. 17) Accordingly, they postponed their task of evaluation and set about a fact-finding survey to provide background for a long series of studies leading eventually—they hoped—toward evaluation. Their schedule asked a great many questions, including an inquiry whether, in the doctor’s opinion, he was using psychoanalysis or something else.

This group was concerned only with psychoanalysis. When it comes to defining psychotherapy, an unusually knowledgeable author remarks, “Psychotherapy has many more variants than psychoanalysis and what constitutes psychotherapy and what does not is even less clear than what constitutes psychoanalysis.” (133, p. 321) Other groups and individuals setting out to do evaluative research have also changed their immediate objective to an investigation of the therapeutic process. Notable among these are the Menninger Clinic in Kansas and the Rogers group at Chicago.

Apparently a broadly acceptable definition of social casework is no easier to achieve. The Hollis-Taylor report, for example, includes among its major recommendations a suggestion that studies be undertaken to define social work in general and each of its specialties in particular (32, 139). The need for a satisfactory definition of social casework was also apparent when a questionnaire was circulated to a group of casework agencies asking which questions they would most like research to answer. High on their list was the question, as one put it, “What is uniquely casework?” (127) In both of these instances the need was ascribed, not to research, but to the practice field.

The kind of definitions desired by the practitioners mentioned above would supply only the first step toward the kind of definition required for rigorous evaluative research. They seek agreement about the elements common to and distinctive of practice in their respective fields. The statement of the broad type is only a first step, however. It may be necessary to move a step further and say what type of psychotherapy or of social casework is being used, since within any professional field there are different specialties and different schools of thought. For instance, substantial theoretical dif-
ferences exist between "Freudian" and "non-Freudian" psychoanalysis and between "directive" and "nondirective" psychotherapy, while in casework the distinction between "diagnostic" and "functional" theory has been highly controversial. The heat of the controversy surrounding various schools of thought testifies to the conviction that these differences in theory are reflected both in practice and in the results of practice. Such a conviction can be tested only against a clear statement of the theoretical orientations under study, including a statement of what the practitioner thinks he is doing—statements to be derived, not from empirical research but rather from professional consensus.

WHAT METHOD IS USED—IN PRACTICE?

Questions about the relation of theory to practice lead directly into the next point: that before it is possible to say how well any method is succeeding, it is necessary to say not only what that method is supposed to be but also whether the material under study represents it accurately and fairly. This usually calls for a detailed study of practice.

What an executive thinks his agency is doing often does not correspond at all points with the activities of the various staff members. What a practitioner thinks he does may not be exactly what he actually does. The Committee on the Evaluation of Psychoanalytic Therapy, for example, abandoned the questionnaire approach after finding that "there may be considerable difference between what the practitioner of psychoanalysis submits in answer to a questionnaire as to what he does, and what he indicates that he does in the course of further questioning and discussion." (23, p. 387) But the difference between theory and practice may go further. The secret of the practitioner's successes—or of his failures—may lie in features of his practice that are not part of his explicit theory and of which he himself is not aware. Only by studying practice in detail is it possible to know whether the method used is actually the one assumed to be used.

A similar necessity holds for evaluating any kind of service: one must be sure that the method to be evaluated is actually the one used, and that it is used in a representative way. Attempts to evaluate probation services, for example, have often been blocked by the difficulty of finding first rate services to evaluate. Do disappointing results mean that the kind of service—i.e. probation—is not an effective means of bringing about change, or only that the service studied is not good of its kind? Or, on the other hand, is it true—as has been suggested—that the mere fact of being on probation is the effective element and the nature and quality of probation services are negligible
factors (76)? Again the question of evaluative purpose is conspicuous: to study the effectiveness of probation services as a way of handling juvenile delinquents is far more demanding than to study the effectiveness of probation services as administered by one department—though this itself is difficult enough.

As suggested above, detailed study of practice is necessary also to discover to what extent differences in theoretical orientation are reflected in practice. Strong opinions are expressed on each side of this point and each side can point to some evidence. For example, a study of psychotherapy with children reported that when records were examined the contrasts between different theoretical orientations were not as striking as is generally assumed (732). A similar comment has been made by readers of the “best casework records of the year” selected by the Family Service Association of America. It is often remarked that, on the basis of these records, one could hardly distinguish between representatives of the diagnostic and the functional orientations. Similarly, it is often observed that among psychoanalysts the label “Freudian” or “non-Freudian” gives little clue to actual procedures in practice, and that members of each school seem to follow, much of the time, principles commonly supposed to characterize the other.

There is, in fact, a good deal of speculation whether the difference between schools of thought is as great as the difference between individual practitioners within any one school. A number of studies indicate that the experience and competence of the practitioner are more influential in determining treatment outcome than is the school of practice in which he was trained (24, 63, 98, 99, 100, 126, 146, 227, 252). This raises the interesting question whether experienced practitioners grow more competent in their chosen method, or whether with increasing experience all tend to move toward some common denominator of relationship therapy in which divergent schools of thought tend to merge. There have been suggestions that a fruitful focus of study would be the elements which are common to all types of relationship therapy, including those practiced by shamans and medicine men. According to one point of view, “the differences which each school holds up as superior and unique to itself are not the causes of healing. Instead it is the points of commonness that contain the elements necessary for what is generally held to be therapy.” (72, p. 104) This view holds that in relationship therapy the type of practice and theoretical orientation as ordinarily described are irrelevant and that success or failure hinge chiefly on other factors—factors in the individual treated, in his situation, in the therapist, in the interpersonal relations between the treater and the treated. Empirical evidence exists on the other side also, with investigators claiming to have documented differences in practice that do not fade out as experience and expertise increase; and with commentators warning against wishful thinking and biased selection of data.
on either side (31, 77, 301). The salient fact here is that so far neither side has enough evidence to establish its case conclusively. Whatever one's views about the importance of theoretical orientation and practice techniques, they can be tested only by documented reports of actual practice (129).

To study practice in detail is arduous and time consuming, but it has been found repeatedly that the examination of actual practice and procedures preliminary to research has been as valuable as the final results of the study. Almost any systematic investigation of practice is likely to yield surprises, pleasant or painful at the time and useful in the end. Again and again research reports testify to the gains flowing from such self-examination. Such gains are variously phrased as "clarifying thinking about practice," "identifying the elements of our art," "bringing intake into line with department goals and function," "examining and sharpening concepts about the treatment process," "making explicit things that were only half realized," "defining glibly used and vague terms and concepts," "opening our eyes to relationships and possibilities not yet perceived," etc. (183, 259).

A number of people are working today on the problem of breaking down psychotherapy into its component parts and describing those parts so accurately that a second observer, trained to perceive the same components, will describe them in identical terms. Various kinds of content analysis have been employed in attempting to study the therapeutic process, sometimes as steered by the therapist, sometimes as revealed in interaction between therapist and patient. Some have tried to analyze the type of response the therapist offers: active or passive, directive or non-directive, approving, disapproving, neutral, etc. One study includes an assessment of the various roles adopted by therapists. A Johns Hopkins' project attempts, among other things, to analyze the terms in which the therapist conceives and formulates the treatment problem and goals—for example, whether the diagnostic formulation included, in addition to clinical description and narrative biography, the meaning and motivation of the patient's behavior; whether the strategic goals of therapy were "personality-oriented" or "psychopathology-oriented." The University of Michigan Therapy Project has made notable efforts to break down vague concepts into researchable elements. In this project, "depth of interpretation," is defined as the degree of disparity between the view expressed by the therapist and the patient's own awareness of his emotions. The elusive variable "warmth of therapist" is subdivided into degree of commitment, effort to understand and degree of spontaneity. These scattered and fragmentary examples are cited merely to show that, although attempts to

---

define the means of change are relatively recent, a variety of dimensions has been used (31, 70, 176, 204, 243, 291, 324).

Those who are working at this problem are the first to recognize that no one way has yet succeeded in catching all the elements that must be controlled in order to represent accurately and consistently the therapeutic process. It is always possible to hope that a wholly new approach may prove simpler and still adequate to the demands of evaluation. Such a hope is the more inviting since some of the current efforts to describe the therapeutic process make monumental demands for observing and recording. These demands, which are of an order different from the evaluative questions being discussed here, can only be mentioned at this point. Yet, once the importance of knowing actual practice is granted, the mechanics of acquiring that knowledge must be recognized as a very serious problem of evaluative research. (Some of its aspects are mentioned in chapter III.)

**By whom is the method used?**

The preceding paragraphs bring home—if that were needed—the extent to which the method is the man and the man is the method. It is impossible to discuss the reasons for studying practice without discussing the role of the practitioner. His importance, however, calls for further comment. It is generally conceded that a study of psychotherapy must include the therapist himself as a key variable. Some hold that the therapist is considerably more important than the type of therapy he happens to practice. An obvious cluster of questions about him concerns his training: What profession and what school of thought within that profession? How much training? Where, and what kind? These questions, unlike many proposed in this report, are relatively easy to answer. No more difficult, and perhaps even more vital, is the question of experience—how much has he had, and what kind?

It was mentioned, in discussing method, that some observations have shown more difference in practice between experienced and inexperienced psychiatrists of the same orientation than between experienced psychiatrists of different orientations. Without attempting to assess the relative importance of experience vs. theoretical orientation, it is clear that in describing the means by which change is to be brought about, the experience of the practitioner is a significant item. Experience is a function of the individual and not of his theoretical orientation. Yet it relates directly to method, challenging the researcher to discover what features of practice are associated with experience rather than with theories of treatment, and how these relate to treatment outcome.
The same practitioner will, of course, vary his methods to suit the case, and the more experienced he is the more variation may be expected. Moreover, evidence indicates that—regardless of experience—individual practitioners tend to have different rates of success with different kinds of patients or problems (324).

It is often argued that consideration of the practitioner should include not only training and experience, but also general approach and personality make-up. It has even been suggested that different personality types are attracted to different professions or to different schools of practice, so that the difference between specific methods or orientations in psychiatry or in social casework may in fact represent the difference between the type of person who chooses to practice one or the other. Whether this is true or not, the application in practice of any theory is certainly influenced by the personality and viewpoint of the practitioner.

A number of points have been made about factors which need to be considered in describing the means by which change is to be brought about. These points are, of course, strongly interrelated. Theoretical orientation, training, experience, personality make-up are interactive and are all part of the way in which theory is translated into practice. To what extent each one will be accounted for in any study will depend on the purpose and scope of the study. The blanket requirement is not that each one be part of the study plan, but rather that if any be left out it be by design and not by inadvertence.
III. ABOUT THE METHODS USED FOR ASSESSING CHANGE

How Trustworthy Are the Categories and Measures Employed?

How RELIABLE ARE THEY?

Before depending on any measurement or rating, it is necessary to assess its reliability. That is, it is necessary to determine to what extent the differences it reveals arise from inconsistencies in the measuring or rating rather than from differences in what is measured or rated. Or to put it differently, it is necessary to make sure what degree of consistency can be assumed for the instruments and the way they are used.

There are a number of methods for determining reliability. The most frequent procedure used in evaluative studies of casework or psychiatric treatment is to have the same material observed or coded independently by different individuals and to compare their results. If agreement is sufficiently high and consistent, the instrument is considered reliable. The definition of "sufficiently," the number and nature of the analysts, and the method of computation are matters to be decided by a competent technician on the basis of the requirements and limitations of the study.

The reliability of the materials used for a study is obviously as important as the reliability of the processes by which they are analyzed. One check on the accuracy with which events and behavior are reported is to determine the extent to which independent accounts of it agree about what happened. Many evaluative studies devote considerable effort to establishing the reliability of individuals independently reading the same records, but skip the step of establishing the reliability of the records themselves. Admittedly, it
would often be difficult or impossible to test the reliability of the records used. It is merely assumed that the material analyzed represents an accurate—though often incomplete—account (180). Such an assumption often has reasonable grounds. If it is made, however, it should at least be recognized as an assumption, and some estimate offered of its probable justification. Efforts to test the reliability of recording are likely to be extremely arduous and time consuming—as witness some of the work done with individual interviews and also with small groups (13). Nevertheless, considerable efforts have been made to check and improve on the accuracy of recording, with a view to reliability as well as to completeness of information (12, 106, 185, 198, 227, 259, 261, 264, 274, 318).

Electrical recording of interviews appears to offer fewer psychological obstacles than was formerly assumed. Numerous studies have shown that patients and clients have far less objection to it than anxious practitioners often expected. One or two instances have actually been reported in which a patient found comfort and help in expressing himself to the recording machine during the therapist’s absence. On the other hand, electrical recording is extremely expensive and time consuming, and it does not eliminate the need for later analyzing and coding. Moreover, even complete verbal transcription fails to capture all the significant elements of an interview. Its advantages and disadvantages make a fascinating chapter in methodology, but one which is beyond the limits of this discussion (23, 87, 108, 177, 187, 277, 295).

Reliability in evaluative studies of the type discussed here depends partly on completeness and accuracy of data, partly on explicitness and concreteness of definitions, partly on the qualifications of the coders, and partly on their training in the use of the categories employed. No amount of research technique can restore missing information or compensate for inaccurate records. However, if adequate data are available, careful construction, definition, and pretesting of classifications, combined with careful training of qualified coders or raters, can usually achieve the required degree of reliability. This does not mean that high reliability is easy to achieve but merely that it can be done, provided adequate clarity of and consensus about definitions have been established. This may be a long and arduous process.

It is interesting to consider the contrast offered by the reliability of diagnostic classifications commonly employed in practice and of the various classifications employed in evaluative research. As reported in chapter II there is proverbially low agreement between practitioners in the use of diagnostic classifications in psychiatry and social casework. That is, the natural reliability of the classifications and of the classifiers is low. On the other hand, it is usually possible to achieve relatively high reliability—in measures devised for evaluative research in these same fields.
Merely to state the contrast suggests at least part of the explanation: consensus, clarity, and purpose. The requirements for reliability are agreement on the meanings of categories and clear, full, explicit, concrete detail in their definitions. Often this means working out a general or abstract definition of the category, plus concrete examples of the variations that fall under it and of the different degrees or subcategories into which it is divided. Given these prerequisites, plus sufficient training and background on the part of the raters, reliability becomes one of the more solvable although not one of the least demanding of the researcher's problems.

But the prerequisites are very definitely not given in clinical diagnosis. As we have seen in chapter II there is often neither clarity nor consensus regarding the diagnostic classifications of psychiatry or social casework. Accordingly, studies that use current diagnostic classifications rather than definitions especially worked out and agreed on are likely to show low agreement (or reliability) among practitioners. And at least one study reports that the greater the experience of the practitioner, the lower the reliability in using diagnostic categories (10).

The reasons for lack of consensus and clarity are inherent in the purpose and history of the categories. The categories used in research are set up by one individual or group with the express intention of achieving reliability. Once established, they are used by one group, under controlled conditions, with no reason for departing from the definitions agreed upon.

In practice, however, diagnostic categories have a very different origin, evolution, and application. A rough and ready approximation of reliability is assumed, but the demand for very high and consistent reliability is likely to occur relatively late rather than with the first use of a classification. And the most persistent efforts to increase reliability are likely to be sparked by the wish for adequate statistics and research—as in the case of the mental hospital administrators and statisticians (222, 223, 224).

In evaluative research so far, less effort seems to have been devoted to achieving reliability in diagnostic classifications than in ratings or measures of outcome—perhaps because so much prerequisite work remains to be done on the description and definition of the changes that are desired.

Reliability tests, by definition, focus on the means and agent of measurement rather than on what is measured. When ratings or codings are involved, it is necessary to be sure of consistency not only between different raters, but also in the same one at different times. It is not enough to determine reliability in advance—although this is necessary. In addition, a continual check must be made to be sure that reliability is maintained. This is especially true if coding or rating continues over a substantial period of time, allowing unrecognized variations to develop. It should be added that reliability can never be assumed—it must always be tested. Perhaps one reason
for the low agreement among practitioners is that testing for agreement is not
normally a part of practice.

Some studies substitute a “conference judgment” for the usual reli-
bility check. Two or more individuals analyze the material independently
and then compare their results. If they do not agree, they re-analyze to-
gether, if necessary calling in additional consultation. By this method the
original definitions are constantly invoked, keeping them fresh and clear.
For further assurance, the conference method can be combined with a
systematic check on reliability between two or more coding “teams.” (262)

When psychological tests are used, the question of reliability is some-
what different. Such tests are checked for reliability before they are gen-
erally accepted for use. If their reliability is not trusted, at least by the
investigators, they are not likely to be employed in an evaluative study.

It seems unnecessary to dwell at length on reliability, for several
reasons. One is that the need for reliability, once pointed out, is likely to be
self-evident. It is easy to see that study results cannot be trusted if cate-
gories are not consistently applied. If the same patient may be classified as
schizophrenic or manic-depressive and the same result as excellent or disap-
pointing, depending on who codes the material and when, then the results
become either meaningless or deceptive. If the same degree of social adjust-
ment in a parolee may be classified as excellent by one rater and as fair by
another, then the study’s final report on the results of probation loses
significance.

The problem of reliability is of course greater with more refined cate-
gories, as the mental hospital statisticians and administrators found in connec-
tion with their revised classification (222-224). If the categories are broad
and obvious there is less danger of unreliability than if they reflect subtle
differences. It is proverbial, for example, that reliability is higher for the
extremes than for the middle points of a seven point scale. For this reason,
when meticulous pains cannot be taken to achieve and test reliability, some
studies use only broad and obvious categories. Some even use only extremes,
analyzing the cases for which natural reliability is high and discarding the
middle ranges. Such a device makes it possible to say what proportion were
unquestionably helped or not helped, leaving a considerable number in doubt.
Its chief usefulness is not as a means to evaluation per se, but rather as back-
ground for determining what characteristics or conditions are strongly cor-
related with therapeutic success or failure. Thus it may offer a good deal of
usefulness to pre-evaluative research.

Another reason for not dwelling at length on reliability problems is
that they are among the most tractable in evaluative research. They can be
solved by methods already available. Also—or perhaps therefore—they are
among the best known problems of evaluative research. Almost the first
words the nonresearch person learns about research are “sample” and “reliability.” They seem, in fact, to have become catchwords, so that if questions about sample and reliability are answered satisfactorily, it is assumed that everything about the research is satisfactory.

The nontechnical meaning of the word “reliability” may be partly responsible for its siren hold on consumers and producers of research. A writer known for his strong commitment to statistical method has commented, “To the layman at least, ‘reliability’ conveys more than that the test correlates highly with itself, and I am inclined to believe that even to many psychologists it is subtly misleading.” (52, Introduction, p. xv)

That satisfactory reliability is an indispensable research requirement, then, needs to be recognized but does not need to be publicized. What does need emphasis is that satisfactory reliability is, as one researcher put it, “necessary but not enough.”

**HOW VALID ARE THEY?**

The validity of a category or measure depends on the extent to which it actually captures and assesses what it is supposed to measure or rate. Thus, validity concerns *what* is measured and the meaning of the results. Reliability, on the other hand, concerns how the measuring or rating is done. Satisfactory reliability gives assurance that the findings are not accidental products of inconsistency in the application of research procedures and instruments. Satisfactory validity gives assurance that the findings mean what they appear to mean.

The problem of validity invades every aspect and every detail of the evaluative process, especially the selection, definition and application of criteria. (Chapter II, p. 15) It has been pointed out that a number of the research questions discussed here are overlapping, and none is more intertwined with all the others than the question about validity. It calls for separate discussion, not only because of its crucial importance, but also because recent research history has tended to blur the definitions and relative significance of reliability and validity.

Reliability is prerequisite to validity. If data analysis is distorted by inconsistencies or individual vagaries in recording or classifying, no firm basis exists for considering validity. Yet no amount of reliability in itself establishes validity. For example, there may be high reliability on evidence that certain psychiatric symptoms have disappeared after treatment. This does not demonstrate, however, that their disappearance is valid evidence of cure or even of improvement. It is well known that remission of psychiatric symptoms may
be accompanied or followed by the appearance of substitute symptoms. Accordingly, reliable evidence that presenting symptoms have disappeared can be viewed as valid evidence of improvement only if it can be demonstrated (a) that the disappearance of the specific symptoms does in fact imply improvement; (b) that no substitute symptoms have arisen to take their place.

In evaluating psycho-social change, it is often possible to amass reliable evidence, but very difficult to demonstrate its validity. For example, one frequently used criterion of psycho-social improvement is better job performance. The validity of findings about job performance will depend (a) on the extent to which job performance actually reflects improvement and (b) on the extent to which the indicators selected and the way in which they are applied actually reflect job performance.

To start with (b), there is room for debate on the validity of indicators used to reflect job improvement. Neither a quantitative nor a qualitative measure is likely to be enough in itself. Increased quantitative production at the sacrifice of quality may not represent improvement. Improved quality at an exorbitant sacrifice of quantity may not represent improvement. These problems can be handled, but they suggest that to work out valid indicators of job improvement is by no means simple.

At the same time, improved job performance may not always be a valid indicator of psycho-social improvement. That it is likely to be one is an opinion widely shared among professionals and laymen. Nevertheless, under certain circumstances better performance on a routine manual job could conceivably result from an individual’s giving up aspirations to a more satisfying career for which he is in fact well suited. There might be an open question whether lowering his occupational sights was accompanied by psychological changes detrimental to his total emotional economy. In this case there might well be a difference of opinion whether improved job performance is necessarily valid evidence of improved general adjustment. If so, even though evidence of improved job performance is reliable beyond question, its validity as a major criterion of improvement would remain in doubt.

Another frequent criterion is improvement in family relations. Under certain circumstances, however, apparent improvement in family relations might result from a client’s decision to cease struggling against a destructive family situation. In such a case, reliable evidence of less family bickering might not in itself be valid evidence that family relations had improved essentially; nor would it demonstrate the validity of improved family relations as a criterion of overall improvement on the part of the client.

These examples are cited merely to point up the difference between reliability and validity. Either of the two criteria mentioned might be a valid indicator, if taken in conjunction with other evidence pertaining to validity. Validity, then, differs from reliability also in requiring assessment as
part of a total constellation of factors, for each of which reliability can be established separately.

Psychological tests of emotional or personality factors afford an interesting example of the problem of validity: do they really test what they are supposed to test (161)? Most of the tests in general use have been standardized to varying degrees—the estimate of degree depending sometimes though not always on the viewpoint of the observer. Concerning a few tests there is relatively strong agreement that they are standardized to such a point as to yield consistently the same results on the same material. This is less true of others, especially the projective tests. However, those who are most skeptical about the usefulness of psychological tests for evaluative research, challenge them less on grounds of standardization—i.e., reliability—than on grounds of validity. Very few tests indeed have actually been validated by comparing the test scores with independent assessments of the traits presumably measured. Thus there is little evidence of the relation between test scores and real life behavior. “Who knows,” asks Hunt, “how well changes in the Rorschach will predict changes in social behavior, reports of distress relieved, and social acceptability?” (146, p. 238) An occasional small attempt to validate a projective test against real life behavior has had disappointing results (28). The findings may be different, however, in at least one very large project to assess the validity of a personality inventory test, the Minnesota Multiphasic Personality Inventory.

Psychological tests are sometimes used in evaluative studies of efforts to bring about psycho-social change in individuals. Those who rely fully on them tend to assume that they are not only reliable but also valid. At times they present the results of the tests as if these results in themselves constituted the answer to the key evaluative question. They discuss, not the validity of the tests, but the meaning of the findings produced by them.

Those who question dependence on psychological testing for this kind of research often grant that it may be useful for group purposes, even though they question its value as a measure of psycho-social change in individuals. Probably the prevailing view is that psychological tests, especially projective tests, can be potent and penetrating adjuncts to clinical diagnosis of personality and psychopathology, but that so far none has reached the point of broad acceptance as an evaluative measure of psycho-social change in individuals. It is interesting that recently one or two investigators have advocated using psychological tests of personality as interviews rather than as measuring instruments. That is, to use the test for getting a sharp, clear picture of the individual, to supplement other diagnostic evidence; to use, not the test scores, but their content as the data (279, 344).

It is also generally agreed that if any testing at all is to be used in studies of individuals, only a large battery of tests is satisfactory for evaluative pur-
poses. If an adequate battery is beyond the resources of the study, advises one commentator, better use none at all (323). Aside from time and expense, one difficulty in using a considerable number of tests is that often they do not agree with each other, so that the investigator must find ways of deciding which one to believe (28).

Because of their current status, the use of psychological tests in evaluative research often serves to test the test rather than the results of therapy. If the tests agree with other evidence they are good tests; if not, they are defective. The reason for all this is the question about their validity. There is not yet sufficient proof that they really measure what they are supposed to measure (22, 65, 67, 89, 135, 286, 328, 329, 340, 345).

The various examples given suggest three levels in the validity problem: first, is the criterion selected a valid criterion of what is to be measured (e.g., is improved job performance a valid criterion of therapeutic gains); second, is the indicator selected a valid reflector of the criterion (e.g., is increased production a valid criterion of improved job performance); third, are the various valid segments of the study combined in such a way as to preserve their individual validity and achieve validity of the whole?

The third question becomes crucial when separate criteria are combined in one over-all score or index of outcome. The more differentiated the criteria employed, the more serious is the problem of combining them. If separate assessments are made on an array of criterion variables, some sociological, some psychological, some crudely circumstantial, what weight should each one have in the final summing up? To assign numerical scores to ratings does not convert them into objective facts. It merely provides a device for comparing or combining them. The legitimacy of the scoring system depends on the quality of the logic used in designing it. And if a numerical index is used, there is no escaping the assignment of weights to each component. To give each a weight of one is as much a procedural decision as to assign weights of 3, 5, or 10.

If the logic is sound, it may be justifiable to combine scores from different ratings into one over-all index of "adjustment" or "improvement" or "therapeutic success." Sometimes, however, numbers are arbitrarily allotted to heterogenous findings and arbitrarily added up and the result is called an index of adjustment. This is as if the number of rooms in one's house, the number of cylinders in his car, the number of suits in his closet and the number of his memberships in desirable clubs were added up and called an index of economic status. It is true that very useful indices have been evolved for rating economic status. Their utility, however, has depended on the process that went into selecting and testing them.

Giving equal weights to separate criteria is sometimes more convenient than convincing. In one study, for example, patients were rated on a five-
point scale for each of the following: work adjustment, social adjustment, marital adjustment, clinical findings, insight, patient's subjective complaints—each rather sketchily defined in a few words. The over-all rating was the average of the scores for each of the six criteria. Regardless of what else in such a study is right or wrong, anyone who doubts that the six criteria are of equal importance in all cases will look very critically at the results of this arithmetic. And anyone who reads an account of the method is likely to question the investigators' description of the final score as "objective." There is no objective basis for the initial selection of the criteria, the manner in which they were rated on the five-point scales, or the assignment of equal weights to all six in the final averaging.

There are a number of ways to guard against unjustifiable combining of criteria. One is to avoid lumping elements to which valid weights cannot be assigned. That is, to use the "segmental" evaluation—reporting separately such elements in change as improved job performance, improved family relations, remission of specific symptoms, etc. Another is to recognize and make explicit the process and the theory by which weights are assigned in combining various elements of change. Another is to arrive at a clinical judgment for each case, in which specified elements are considered systematically, but with no formal method of giving the same specific weight to a given element in every case. The cases can then be grouped according to types or levels verbally described, without assigning scores. One group, for example, might show marked improvement in elements A and B but not in C and D; another might show improvement in all; another in none, etc. This does not eliminate the weighting of separate factors but it does make the weights more flexible and less absolute. Also, it protects research consumers—and research producers too—from "the unwarranted confidence produced by numerical scores." (22, p. 47)

One source of possible discrepancy between reliability and validity is the built-in assumptions of those who analyze the data. Treatments and services directed toward producing psycho-social change in individuals are based on professional assumptions which are often widely held, and widespread assumptions can be the source of high reliability, irrespective of validity. There was a time, for example, when one could have achieved a high degree of reliability in findings by medical practitioners about the presence or absence of the four body humors in patients suffering from various diseases. Later medical theory challenged the validity of this highly reliable finding. Now, some see signs of a swing toward a theory somewhat analogous to that of the body humors (216). At any stage of belief reliability can be achieved among raters who share the same assumptions and apply the same definitions; but the contradictory findings so reliably reported cannot all be valid (234).

This example reminds us also that, in assessing psycho-social change, the
question of validity rests ultimately on opinion. Some key problems in
evaluative research relate to this ultimate dependence on informed opinion—as can be seen by considering evaluations in other fields, in which validity can be checked against demonstrable facts. The problem, for example, might be
to evaluate the efficiency with which a certain type of steel could perform in
a machine subject to known degrees of pressure, heat, and friction. It would
be possible to create laboratory conditions reproducing the specified pressure,
heat, and friction. Sound evaluation would need to take into account numerous
other factors—cost, deterioration through time, behavior under certain
abnormal conditions, etc.—but within reasonable limits, these could be con-
trolled and determined. The familiar and painful point is that the test of
validity itself would be a matter of determined fact and not a matter of
opinion. If the material stands up under specified stresses, then there is no
doubt that it stands up. If it cracks or crumbles, then there is no doubt that
it cracks or crumbles. One can see and test and measure the changes.

Lest we over-romanticize the conditions of evaluative research in the
natural sciences, it should be admitted that the picture given here over-
simplifies the metallurgist’s problem. If response to stress and strain could be
determined with complete accuracy, a number of airplane accidents presum-
ably would not have happened. It remains true, nevertheless, that for certain
types of problems the responses of steel can be evaluated more dependably and
less controversially than the responses of human beings. There is not likely
to be serious uncertainty or difference of opinion about whether the steel
is in one gleaming, unmarred mass or is scattered in fragments on the floor.
One needs merely to think of “adjustment,” “cure,” “mental health” to
recognize that their presence or absence cannot be confirmed in the same
manner by reference to “objective” findings.

To recognize this difference without being thrown off base by it is a
primary requisite in the kind of research discussed here. One obstacle to
balanced recognition is partly semantic. In an earlier section (chapter II,
p. 24), it was pointed out that criteria may be applied by ratings, by measures,
or by a combination of the two. The results of measures are often referred
to as “objective” and the results of ratings as “subjective.” These two words
unfortunately have acquired honorific and derogatory connotations leading
to a good deal of unconscious distortion. “Objective” has come to be re-
garded as synonymous with “scientific,” and “subjective” as synonymous with
“emotional,” “impressionistic” or “unscientific.”

The consequence has been a tendency to disparage information which
is not “objective” and cannot easily be quantified; and a corollary tendency to
regard anything expressed in numbers as “objective.” An assumption has
grown up that numbers are more true than words, even though their apparent
precision may be spurious, and an associated assumption that information
which is not quantified cannot be science. This conception is in part responsible for the slighting of the "grubbing stage of intense observation" (295) necessary for learning to identify and control the variables essential to sound evaluative research, since the early stages of intense exploration are usually not conducive to quantification (116, 278).

To a considerable extent the point of view that excludes preliminary exploration from "scientific research" arises from a misconception of the natural sciences as absolute, wholly quantitative, wholly neat and well ordered. The existence and the dangers of "equating scientific methodology solely with quantitative procedures" have been discussed in an article by the psychologist, Hadley Cantril, and three associates. "Scientific inquiry and scientific method," they say, "are ... not to be confused with investigations limited solely to a so-called 'quantitative approach.' An over-concentration on problems of measurement as such can easily side-track the investigator." Pointing out the extent to which advances in the natural sciences rest on initial observation, speculation, exploration, they add that "scientific inquiry will be strangled if the investigator feels that he cannot be scientific without being one hundred percent quantitative." (45, p. 492-493)

This misconception of the natural sciences is less widespread than it was, thanks to the popularizing of such concepts as relativity, the principle of indeterminacy, the field of forces, and to repudiation by natural scientists of the myth that science is strictly quantitative. "It seems to me," writes one of them, "that the worst of all possible misunderstandings would be that psychology be influenced to model itself after a physics which is not there any more, which has been quite outdated." (250, p. 134)

Nevertheless, the cult of objectivity still plagues social research. One of its consequences has been a need to blur the difference between objective and standardized information. Often the quantified results of ratings are referred to as "objective," apparently in an effort to prove they are "scientific." For example, a group of students were asked to rate themselves on a 10-point scale of "nervousness" experienced before examination. The scores derived from this self-rating were then referred to by the investigator as "objective" measures. Actually, they were dubiously standardized forms of highly subjective material. In another study, the behavior reported by clients during therapy was rated as showing little, some or a good deal of maturity, responsibility, control. The ratings were converted into numerical scores which were then referred to as "objective" data, presumably because they were expressed in numbers.

Greater clarity about the differences between the standardized and the objective would have two desirable effects. First, it would promote clearer comprehension of the nature of findings produced by standardized ratings or by arbitrarily derived scores. There would be less tendency to accept them
as "objective facts." A second and perhaps more important result of sharper differentiation between the objective and the standardized would be the recognition that nonobjective data are not necessarily second class data (83, 206, 284). The need to assume that standardized ratings can be made objective by converting them into numerical scores arises from the mistaken belief that if they are not objective they are not scientific. If the difference between objective data and standardized opinion can be faced, the way is opened for recognition that standardized opinion can be scientifically respectable—providing it does not masquerade as "objective" fact.

It has been a matter of surprise to the writer of this publication that objection to indiscriminate equating of quantification with objectivity is often taken to mean objection to counting or to statistical techniques. Certainly the intention is not to oppose those indispensable research requisites. It is rather to plead for (a) greater clarity about what is counted and what the resulting figures or scores mean and (b) recognition that non-quantified and non-quantifiable materials may be legitimate parts of research data. To this extent the viewpoint presented here does diverge from that of the eminent statistician who declared that "what I cannot control I must ignore."

What, then, is the ultimate evidence on which validity rests? How can one prove that a person is better off (or worse off or unchanged) after a period of service or treatment than he was before? For the time being, at least, it appears that "objective" proof is not available. The most "operational" or "behavioral" definition of outcome is based on someone's conviction about what is desirable or undesirable, what is adjustment or maladjustment, what is improvement or deterioration. The "ultimate" criterion of success or improvement is an opinion. Moreover, in the course of time, the opinions held by the most enlightened are subject to change. Validity requires, then, that this ultimate criterion be explicit and theoretically tenable, that the basis for all measures and categories be explicit and theoretically tenable, and that the use of all measures and categories be reliable. This kind of validity cannot be fully attained until all the research questions discussed here can be answered adequately.

To recognize the extent to which findings can be useful even before this degree of validity has been reached, one must return to the basic aims of treatment. Usually an individual seeks service or treatment in order to live with more satisfaction and less pain for himself and for others. Psychological satisfaction and pain are subjective elements. If the patient himself, his therapist, and all who come in contact with him are convinced that he feels and causes more satisfaction and less pain, to all intents and purposes he is better off than before. Accordingly, if measures or ratings can be evolved that reveal changes concerning which all of these assessors would agree and which would also be agreed to by a dispassionate qualified observer, the pre-
umption of validity is strong. This presumption would be tempered to the extent that contrary or conflicting evidence appears. The great insistence on exactness in defining categories and consistency in applying them is to make sure that everyone really knows what everyone else is talking about so that what is equated is really equal and what is contrasted is really different.

Lack of validity at any point can disturb the validity of the whole; and we have not yet achieved ability to test, control, and demonstrate validity at every point. Such a statement is not a counsel of despair but a counsel of clarity. The answer to the validity dilemma, as to many others, lies in increasing efforts to clarify rather than to slur over distinctions, to recognize the limits of what can be achieved by the means so far available, and to build on current achievements in order to widen current limits.

Evaluative efforts in casework and psychiatry offer examples both of expressing and of suppressing the validity problem. One report, for instance, after elaborating the evidences of reliability, dismisses validity with a footnote reference to another paper about the same study, in which the problem of validity is "discussed." This "discussion," when checked, turns out to consist of a statement that we have to assume that caseworkers know their business. A different treatment of the problem occurs in a study often cited with respect by evaluators:

A critical appraisal of our data in terms of accuracy and reliability reveals the fact that in describing interpersonal relationships, sexual adjustment, basic conflicts, "ego defenses," etc., the material is not only incomplete but is by no means unbiased. An investigation of dynamic factors is unavoidably shaped by the examiner's skill, personality, and theoretical orientation. From the psychodynamic standpoint the most significant facts were least accessible, and when obtained were apt to be fragmentary and difficult to evaluate. Factual data concerning age, sex, number of siblings, years of schooling, etc., were reliable but were of course relatively unimportant.

In the analysis and interpretation of the data there were apparent contradictions. This probably indicates that our criteria for evaluating such factors (and our anamnestic information) are inadequate to detect any but gross differences (227, p. 103).

Because the authors of this study so sharply define its limitations, and take them into account in interpreting their data, the reader tends to place full confidence in the report, which has served as a model and a stimulus to the field. Limited findings can be of tremendous usefulness provided only that the limitations are clearly stated—not just in a terse foreword which is
forgotten by the time the conclusions roll around, but in connection with the interpretation of findings. In turn, clarity about limitations provides both impetus and basis for attacking them.

At What Points Is Change To Be Measured?

FROM WHAT BASE?

Studies of change are, by definition, of the “before and after” type. In order to assess the “after,” it is necessary to be clear about “before.” Accordingly, evaluative studies of change require a firm baseline to which the “after” can be compared. The necessity for it has been brought out in discussing the need to stipulate “change from what?”

Findings are often reported as direction and sometimes as degree of change: “little or no improvement,” “somewhat improved,” “cured.” Such ratings imply the relation of “before” and “after” but do not reveal how bad “before” was or what it represented. That is, they are relative ratings, showing movement but not status. Other findings are reported as a description of “after” without reference to “before”—for example, by rating the patient or client on degree of sickness or health, or degree of adjustment after treatment. These are absolute ratings, showing status but not movement. Neither type gives an adequate base against which to measure change.

At least two very serious efforts to improve the baseline are under way. The Menninger Clinic, as mentioned in a previous section, has been working on a measure of mental health. This measure of health status would place the patient on a health-sickness scale, so that a rating made before and after treatment would show how much change had taken place in between. That is, the amount of movement would be revealed by comparing two measures of status. And the Community Services Society in New York City is working toward a measure of status which would provide anchoring points for the movement scale already developed at that agency by J. McVicker Hunt and his colleagues (147, 299). The need for such anchoring points has been evident for some time, and the investigators who collaborated in developing Hunt’s original scale have remarked more than once that without them it was incomplete. Their conviction of the necessity for describing status as well as movement is recognition of the need for establishing a firm base against which to measure change. An effective measure of status plus an effective
measure of movement could make it possible to tell both "from what" and "how much" change had occurred.

**AFTER WHAT INTERVAL?**

It is usually regarded as axiomatic that assessment of change should compare information at least from the beginning and the end of treatment. As noted above, practical considerations often leave the initial picture implicit rather than explicit, but this is a compromise researchers make perforce and with regret.

Many believe that at least one assessment should be made during the course of treatment, and most would agree that there should be a followup study some time after treatment has ended. A rating of results made at the close of treatment is in a sense a prediction about its effects in the client's subsequent day-by-day living. If these effects are not stable, but evaporate when the individual is no longer sustained by a therapeutic relationship, then their success is obviously less than if they endure over a long period of time. Ideally, efforts to bring about psycho-social change in an individual should have cumulative effects, so that the values gained from them increase as he is able in daily life to build into his attitudes and behavior what has been learned during contact. (An exception here would be the kind of supportive therapy that aims only to help the individual from day to day, with the recognition that he probably cannot be helped to the point of getting along without this support. Such cases—like all cases—must, of course, be evaluated in relation to the therapeutic goal.)

The fact that therapeutic contact is a learning situation adds another reason for making a followup investigation after an interval. Critics of psychiatry, casework, and various types of individual counseling sometimes say that what people learn in treatment is merely how to say what the psychiatrist, caseworker or counselor wishes them to say (72). Such skepticism invites demonstration that even after a considerable time has elapsed the fruits of this learning experience are evident in the individual's feelings and behavior. This is the proof of the pudding. Follow-up studies, of course, are not made chiefly for the avowed skeptics. They are wanted more by the practitioner, eager to learn what treatment has what results, and by the administrators, board members and general public who are actual or potential directors, supporters and subjects of treatment and counseling.

Since few would dispute the need to make assessments at the beginning and end of treatment and after a post-treatment interval, it is striking to find as recently as 1951 the categorical statement: "In the published research on
psychotherapy, no investigator has reported the use of the same techniques for evaluation of treatment at the time of initial contact with the patient, at the close of therapy and at time of followup at a subsequent date.” (320, p. 293) By 1957 this statement was no longer strictly true, although the challenge came from a very select minority of recent and current projects.

Followup studies.—Assuming that such studies are necessary to adequate evaluation, they raise a number of questions—some practical, some pertaining to professional theory and ethics. Many have believed that, although such studies would be desirable and useful, they were too "dangerous" to undertake. Practitioners have often feared that they would do harm to the patient or client, and also to public reliance on the confidentiality of professional relationships. This kind of fear has diminished greatly. In fact, the amount and speed of the change should nourish optimism about other "insuperable" problems in evaluative research.

Nevertheless, the followup study usually requires great caution and many precautions. It should be, and usually is, undertaken with extreme care in the approach and interviewing. Many agencies and individual practitioners review the cases of those who are to be approached and if there seems reasonable ground for thinking that the followup could be damaging to an individual, he is removed from the sample (180). Moreover, even when the research design stipulates that the interviewer should know nothing about the case, if special circumstances indicate that special caution should be observed in connection with some feature of the case, the interviewer may be briefed about that one point in advance. That such precautions may bias the sample must of course be recognized, and the probable effects dealt with both in analysis and in reporting.

Responsible research people are extremely careful to preserve confidentiality—especially, by not approaching an individual in the presence of others and by not, in trying to locate him, making the kind of inquiry that would let others know anything about him he might wish to conceal. They place top priority on the need to protect the former patient or client from any effects that would be damaging or unwelcome to him.

So far there has been little if any evidence of social or psychological harm resulting from a carefully planned followup study. On the contrary, a number of those who have been interviewed in such studies have felt that the experience was helpful to them. Our survey, however, did not reveal any studies specifically designed to investigate the effects of followup interviews, so that information about them is for the most part a byproduct. The possibility that they may be harmful remains highly theoretical, since the little evidence available points in the other direction.

Aside from basic questions about the professional correctness of followup studies, a number of practical questions are raised about them. An im-
important one relates to feasibility: will it be possible to locate the desired respondents, and—once they are found—will they be willing to participate?

**Locating the sample.**—Finding people after a lapse of time is often difficult, and the difficulty is likely to increase as time goes on. Nevertheless, a great many studies have shown this difficulty to be considerably less than might be expected. In planning a study, it must always be assumed that after some months or years a certain proportion of the desired sample will be unlocatable, or can be located only with great difficulty and at great expense. The Community Service Society sent an investigator all over the United States to locate the principals in 38 cases (181). On the other hand, many studies do not require that every member of the original sample be located. Some provide a list of alternates, some allow for sample "decimation."

Often it is possible to define what steps shall be taken before an individual is considered "unlocatable" and to make up the sample of those who can be found. If this is done, allowance must be made in the study design for loss of the unlocatables. If a certain sample size is called for, there must be a reservoir of cases to draw from, to substitute for those not located; and this reservoir must be chosen in the same way as the original sample. If "all" cases of a certain type or agency or period are to be studied, the kind of analysis planned must be realistic in terms of the number finally located. That is, plans for elaborate breakdowns and comparisons need to be based on the number that can actually be interviewed rather than on the number who had received the treatment or service under study.

The ability to locate the desired individuals is likely to vary with geographical region and type of community. In a large city where mobility is high, there may be considerable difficulty in finding them. On the other hand, in a recent followup study of 500 adoptions, made in Florida 10 years after the adoption petitions had been granted, over 70 percent of the original sample were located within the State. Difficulties encountered in locating people for followup studies underline the value of keeping adequate records on cases, including full names and addresses of relatives. It is always necessary, of course, to make a painstaking comparison between those who have and have not been located, so that any differences which might be significant for outcome can be reported and taken into account in the analysis.

The purpose and plan of the study will have to determine how much time and money should be devoted to locating cases. And obviously, in this as in all respects, the study plan will have to be adapted to available resources of manpower and funds. In husbanding resources, it is sometimes helpful to remember that the work of locating people need not be done by the research staff, but can be assigned to others—if they are available—working closely with the researchers.
Will they participate?—The evidence so far indicates that the majority of patients or clients do not resist or resent an attempt to discover how they are getting along some months or years after treatment, providing due care is exercised to avoid any possible exposure or embarrassment for them. If this is done, and if the approach is carefully planned and carried out, most of them seem ready and willing to put their experience to work for the benefit of others. Some seem to welcome the interview, and many say they think it has helped them. (The followup interview appears to offer values for practice, quite aside from research. The Jewish Family Service in New York City has been experimenting for some years with a followup interview, and has found it useful enough to become a regular part of practice with extended counseling cases (130)).

There are of course varying degrees of acceptance or resistance to participating in such a study. On the whole, even those who are unenthusiastic or antagonistic at the outset seem likely to warm up considerably by the end of a well conducted followup interview. Certainly some refusals are likely, although the usual experience is that they are fewer than expected. Whatever the number, it is always necessary to compare what can be learned about those who refuse with what is known about those who do participate, and to estimate and report the probable effect of the refusals on the study findings.

The number of refusals can be influenced by a wide variety of factors, including the skill of the interviewer, the population under study, the nature of the evaluation, current public attitudes toward the treatment or service to be evaluated, and specific features of the approach, some of which are discussed below. One group of investigators, approaching clinic patients after an interval of from 3 to 12 years, found that the more disturbed individuals and the ones initially hostile to the clinic were the most likely to object to participating (169). Others, however, have found that even those who were formerly hostile and dissatisfied were often quite ready to participate—some, apparently, because they were glad of a chance to “let off steam” and some because their feelings had changed during the years.

Whatever the factors making for refusal, those favoring participation seem on the whole to be considerably stronger. Followup studies may be seriously hampered by inability to locate the desired participants. But we have found no reports of studies that had to be abandoned because people would not participate, or even of studies in which the refusal rate was high enough to throw serious question on a carefully qualified report of the findings. Where serious questions arise, they are likely to stem from some other source.

Those who fear that people will be unwilling to participate often underestimate the appeal of an opportunity to render service, to have something useful to others come out of the trials one has suffered or the benefits one
has received. At the same time, if people are acutely troubled by problems related to the subject of the study, the hope of getting help sometimes serves as added inducement to participate.

**How long an interval?**—Investigators disagree about the optimum interval between the end of treatment or service and a followup study, although arguments on both sides are convincing enough that few seem dogmatic about their own solutions. Only a study made after a considerable interval can give information about the stability of changes evident at the end of contact. But the definition of “considerable” varies widely. An investigator recently referred to a followup after 6 months to one year as “long term.” Other studies have been made after intervals up to 10 and even 20 years (86, 105, 233, 269, 322).

The usual objection to the very long term followup is that the longer the intervening time, the greater is the opportunity for other influences to enter a person's life and thus the more difficult is it to demonstrate that any changes which occur are ascribable to the treatment given so many years ago. Some research people would avoid the very long term followup for this reason (143).

On the other hand, if the followup is made after a period of less than a year, there has not been sufficient time to test the outcome. This is the more true because of so-called sleeper effects—which may be either good or bad. At times, after what seems to have been an unsuccessful contact, a client or patient begins to experience gains that were imperceptible while he was in therapy. Caseworkers occasionally learn of such developments by accident, or by having the apparently unsuccessful client refer someone to the agency because he has been helped so much. Psychoanalysts are familiar with and often count upon the continuing gains a patient may experience long after the active analysis has ended. This kind of effect requires time to manifest itself, just as time is required for the unfavorable effect—the wearing off of what had appeared to be highly gratifying gains (35, 233, 322).

Allport, in his foreword to the Cambridge-Somerville study, cites an interesting example of sleeper effects—the case of a “cynical lad” who at age 17 gave a negative report on the effect of the efforts to help him made by “Miss A” who had worked with him some years earlier. Later, at age 21, an unsolicited communication from this same youth spoke “in the highest terms of Miss A’s influence upon his life. The reversal in his evaluation points up sharply a basic issue: when in the course of an individual’s life shall we assess the effects of character-building influences? It takes many years for some seeds to germinate.” (260, p. xiii)

To the writer of this publication it seems that one year is minimal for the followup period; and that 5- and 10-year followup studies will be necessary.
to establish claims to real effectiveness for psychotherapy or social casework. The results of intervening experiences can be checked through shorter term followup studies and also through replication of studies using large samples and whatever means of control can be established. (See Controls, p. 62)

Repeated followup contacts are preferable to a single one for another reason—namely, that every individual, with or without treatment, fluctuates in his behavior and outlook on life. Followup studies do not usually take account of these fluctuations, which may be confused with improvements due to treatment or with treatment failure (28, 133, 343). It takes a great many followup studies at varying periods of time, including a great many individuals representing a wide variety of initial problems, to give a trustworthy picture of results. This fact is usually conceded in principle. In practice, however, the followup study is often skipped entirely. Or, if undertaken, it is often inadequate in the length of the period, the number of individuals interviewed, or the number of interviews with each one. Ideally, a series of followup studies at one or two year periods, would be desirable—provided the effects of repeated study could be handled adequately. No such project has come to our attention, however. Different kinds of treatment or service will of course require different periods for followup study.

Who is involved?—It is usually assumed that, whoever else is or is not involved, the individual who received the treatment or service should be—unless this is contra-indicated by special circumstances such as his age or condition. Some studies attempt also to get information from close relatives, friends, colleagues, teachers, etc. The advantage of further evidence from other and possibly less biased sources is obvious. The disadvantages of collateral interviews are also obvious. They can be held only when they involve no possibility of violating confidentiality or of doing other damage, however slight.

Who interviews?—If the followup study is conducted by interview, the character and training of the interviewers is of special importance. Even if a preconstructed schedule is used, with simple yes-and-no or checklist answers, the interviewer in this kind of followup study is in a highly responsible position. Many practitioners and researchers think that only mature, experienced, and thoroughly trained interviewers should be employed in such studies, and that the training and experience should include not only interviewing as such but also some casework, psychology, or psychiatry.

What arrangements?—Most followup studies undertaken for evaluating relationship therapy include direct face-to-face interviewing, although they may also use psychological tests and information drawn from collateral
sources. Interviewing by correspondence, although less costly and often less difficult, as a rule seems to be less satisfactory than a face-to-face contact. Nevertheless, a number of investigators have resorted to mail questionnaires and inquiries, when direct interviewing was not feasible. One study used long distance telephone interviews with people who were too far away for in-person interviews. This method was obviously regarded as second-best by the investigator, and was resorted to only because it brought in data that could not be obtained in any other way.

The type of interview will depend on the purpose of the study and also on the school of thought of the study director. In any case, however, if direct interviewing is to be done the question is likely to arise—should it be announced in advance by letter or telephone, or should the interviewer appear without advance notice? Reports differ on this point, and the approach used will depend partly on the nature of the inquiry and the amount of cooperation to be requested. Some investigators say that, if the respondent can be found when he is alone, it is better to approach him unannounced rather than to give advance notice through a letter or a telephone call. This makes it possible to present credentials and give a full explanation of the study and its purpose, while allowing him to satisfy himself about the appearance and apparent motives of the interviewer (162). For certain studies, however, the disadvantages of this approach outweigh the advantages. Aside from being more costly and more time-consuming, it runs the risk of catching the respondent at an inopportune time, causing him embarrassment or anxiety, which may affect the content of the interview as well as the feelings of the respondent. If advance notice is to be given, a telephone call, when feasible, is often preferable to a letter, which can seldom meet any special questions or anxieties that may be aroused by the request for an appointment. The pro's and con's for each method will have to be weighed individually for each study.

In view of the generally favorable response to the "help-others" appeal, it may seem paradoxical that occasionally payment is offered to respondents for participating in a followup study. This has not usually been found necessary, yet some who have offered fees report favorably about the results (169). According to one report, the amount offered seemed far less significant than the fact that any remuneration was offered, and the gesture was effective both with those who refused the payment and with those who accepted it, regardless of the individual's income or need for money (243). It is possible that the offer of token reimbursement adds to the study both impersonality and dignity which are reassuring to people who are being asked to discuss their private affairs. Fees have been used too little to permit any solid generalization. It does appear, however, that on the one hand they are by no means necessary; and on the other hand they may reduce the number of refusals—
perhaps less because of the money reward than because of its psychological connotations.

How Fairly Do the Individuals Studied Represent the Group Reported On?

How is the sample selected and defined?

The need for an adequate sample is so generally taken for granted that it no longer calls for argument. Probably it is enough merely to state that (1) the group to be reported upon (i.e., the "population") must be clearly specified and (2) either the total population or an adequate sample of it must be studied. An adequate sample must be representative—that is, it must possess, within reasonable limits of error, the characteristics of the population it is to represent, in the proportions found within that population. It must be representative in the first place to give a legitimate basis for generalizing from the sample to the population on which the study will report. In the second place, if comparison is to be made between two groups, the sample must give a legitimate basis for such comparison. This section of the report is concerned only with the first consideration, leaving problems of comparison to the following section.

The "notorious unreliability" of testimonial anecdotes about therapeutic success or failure comes from bias in the sample. One has no idea what the relation is between the incidents reported by unsystematically selected tales and the general experience. Adequate sampling insures that representative experiences of a representative group are reported.

It is not proposed to enter here into technical discussion of sampling techniques. Suffice it to say that the sample is a primary consideration and, for any substantial statistical problem in sampling, a technician must be consulted. The great difficulty, however, is not in the statistical problems of sampling, but in determining and dealing with the characteristics which must be accounted for if the findings of a study are to be generalized beyond the population actually sampled for the study. Some of these characteristics were brought out in considering the question, "who is to be changed?"—characteristics of individuals, of their environments and life circumstances, of the problems in which change is sought. (Chapter II, p. 26)
If all members of a defined population are to be included, there is no sampling problem. The problem would be simply to decide whether the group is large enough to give meaningful results with the type of analysis proposed. This is a question for a statistician to answer. The same would hold if one selected every other member or every nth member of an entire population to be studied; or if it is possible to draw a random sample from the entire population—the method of randomizing and the size of the sample to be checked with a statistician.

Controversy about the adequacy of a sample is likely to concern its relation to the population it is supposed to represent. If a sample consists of every other patient in a certain clinic during a certain time period and if the findings are generalized only to the patients of this clinic, probably no one will challenge the sample per se. However, if the sample consists of all the patients diagnosed schizophrenic in this same clinic during a certain period and if the findings are then generalized to patients diagnosed schizophrenic anywhere, the sample will be utterly inadequate. In other words, the population of patients diagnosed schizophrenic in one clinic cannot be assumed to be representative of the general population of schizophrenic individuals. Some of the reasons for this have been discussed in chapter II. If one is to generalize the results of evaluative research in psychotherapy to the results of psychotherapy in general, or of one type of psychotherapy, or of psychotherapy with all members of one diagnostic classification of patients, then one must be able to demonstrate that the population sampled really represents the broader population being discussed. The practitioner must depend upon the sampling expert for this demonstration. But the sampling expert must depend upon the practitioner for information about and definitions of significant variables which are likely to influence treatment outcomes and which may vary in different populations.

Useful evaluative—and pre-evaluative—studies can be made without generalizing to so broad a population. Moreover, repeating a limited study with a different population can greatly enhance the value of each one. That is, if the same study methods are repeated in a number of different populations with the same results, the likelihood of more general applicability is increased. But the kind of evaluation ultimately sought, applicable to a broad and variable population, involves very serious sampling problems. How serious they are is brought out, once again, by comparison with the physical sciences. Among many who comment on this point, Nathan Kline says: "In the physical sciences, one ingot of standard 24-karat gold, for example, is as good as another if conditions are identical; or one beam of white light under standard conditions can be expected to behave just as any other would. The assumption has been blithely made that one group of schizophrenics (or any other diagnostic group) is as adequate as another in determining attributes or the
effects of procedures.” (172, p. 477) He summarizes a number of the reasons why such an assumption cannot be carried over from the physical to the biological and behavioral sciences: that it is seldom possible to isolate “pure” examples of what one wishes to study; that the interrelatedness and lack of functional independence in the biological and behavioral sciences exceeds anything in the physical sciences; that in the biological and behavioral sciences "organisms probably behave as something other than the sum of their individual parts, even if these could be completely investigated," and—the familiar refrain—that the classes, types and groupings of individuals having psycho-social problems possess little of the concreteness and testability of classifications in the physical sciences.

It sometimes seems to be assumed that the random sample offers a simple solution to all sampling problems. Randomizing within a specific population, however, does not make that population representative of a broader population. It does not give a basis for generalizing beyond the population actually sampled, unless it can be demonstrated that no systematic differences exist between the group sampled and the broader population concerning which generalizations are desired. In other words, a random sample of one agency's clients is not necessarily representative of the clients in other agencies, or of people who need help but have not tried to get it. Obvious as these points are when stated, they appear to be forgotten more often than might be expected (116).

A number of the comments made about sampling for evaluative research in psychotherapy apply with equal force to evaluation of any effort to induce psycho-social change in individuals. For example, a nagging problem in efforts to evaluate the effectiveness of probation for juvenile delinquency is the difficulty of sampling the individuals in such a way that consequences of their family background, their social environment, their current life circumstances and situations should not be mistaken for the consequences of the probation services offered them.

Various experts make various suggestions about coping with sampling problems in evaluative research in psychotherapy. One suggestion is that all possible factors be accounted for and that those which cannot be controlled at least be reported on (172). Another is that until it is possible to identify and control all significant variables, "large scale collection of data based on clinical opinions with relatively simple statistical analysis may produce results which are just as valid as the more statistical treatment of small sample data derived from rigidly controlled experimental situations.” (310, p. 3)

The suggestion to use primitive methods on large samples rather than more refined methods on very small samples sounds almost "reactionary" today, when so many researchers are experimenting with methods that can be applied only to very small samples. A full content analysis of verbatim
records of hour-long interviews held daily or even weekly over a period of months is hardly feasible for a really large sample; nor is it possible to administer a full battery of psychological tests or to make an elaborate series of ratings based on fully recorded material for many hundreds of long cases. One hope held by those who favor elaborate research on very small samples is that when the research methods are perfected they can be used and tested with larger and more representative samples. Another is that exhaustive investigation of small samples will reveal indices simple and strong enough to apply on very large samples.

The meticulous work with very small samples falls chiefly in the area of pre-evaluative research (designed to produce the tools for ultimate evaluation), and as such offers both immediate usefulness for practice and promise of ultimate usefulness for evaluation. The great and chronic proviso is, of course, that no generalizations be made beyond the limits of the data. Even the more sophisticated researchers are sometimes accused of generalizing too broadly on the basis of samples that are either unrepresentative or too small to support the conclusions offered—or both (88).

Some of those engaged in pre-evaluative research are also considering ways of obtaining larger samples. One suggests, for example, that if children are used as subjects it might be possible in a large urban public school system to overcome a considerable number of sampling problems (255). Another proposes, for evaluation of psycho-analysis, the use of various psychoanalytic institutes and societies, as offering “the greatest concentration of former analysands to be found anywhere,” adding that though the samples would not be representative it would constitute a sizeable piece of a representative sample of the analyzed (23). An even more ambitious proposal, discussed under controls (p. 69), calls for a standard nationwide sample to be drawn upon as needed.

Those whose purpose is to obtain a prompt and convincing answer to an evaluative question will probably find it advisable for the present to rely on relatively simple methods (always based on competent technical advice) without devoting great effort and expense either to sampling by traits or to geographical coverage. Until key variables are known and methods developed for dealing with them, even the most adroit statistical procedures for sampling by traits or attributes within one population cannot be counted on to produce a sample truly representative of a broader population. Important characteristics which are not accounted for may lead to systematic differences and so distort the results and produce misleading information. Until there is more clarity about these characteristics, it seems advisable in such studies to keep to the simplest type of sampling and to state clearly the limitations of whatever method is used. For the most part this means using a random sample—or all—of the population to be studied, and not attempting to gen-
eralize beyond that population and its characteristics, or to compare or combine different populations.

Excellent studies can be produced with such safeguards, using the imperfect knowledge already at hand. But until we are better able to describe the individuals in whom change is to be effected, we are not in a position to benefit fully by intricate techniques that depend on sampling by traits. For the sampling expert must depend upon the practitioner to inform him about the key variables that need to be considered in describing a population or in making inferences from one partial population to another.

What Is the Evidence That the Changes Observed Are Due to the Means Employed?

What controls, if any, are used?

Even if it is possible to demonstrate that changes of the sort desired have taken place in individuals by the end of psychotherapy or casework contact, this in itself does not demonstrate that the changes were caused by the efforts of the practitioner. To establish a causal relation between observed changes and the means employed to produce them is extremely difficult. Often the best that can be done is to establish a strong presumption of a causal relation. This, in fact, is what most evaluative studies in psychotherapy or social casework—or for that matter in other forms of personal counseling or in the treatment of juvenile delinquency—have settled for. How strong the evidence or the presumption of causality must be will depend on the purpose of the study. No study, however, can escape the obligation to be clear about the conditions necessary to establish such a connection, and the extent to which this particular study does or does not meet them.

The classic device for demonstrating that observed changes have been caused by treatment is the untreated control group. If two groups are identical and are subjected to identical conditions, with the sole exception of the variable under observation, then any differences displayed by the two may reasonably be attributed to the presence or absence of that variable. Applying this to evaluative research in psychotherapy if two groups are identical, and are subjected to identical conditions except that one has been treated and the other has not, and if the treated group shows favorable changes not apparent in the other group, then there are grounds for
claiming that the treatment is responsible for the changes. If the comparison is made repeatedly between treated and untreated groups which are identical except for treatment, and the results are the same, then the claim is strengthened.

The need for such evidence is reinforced by the repeated finding that about two out of three patients appear to be helped by any one of numerous kinds of psychotherapy and the repeated claim that about two out of three who do not have treatment appear to improve or recover without it. It is generally recognized that some ailments and problems are self-limited—like the common cold which “can be cured in about two weeks if carefully treated and if left alone runs its course in about a fortnight.” In certain cases, fortunate timing of therapy can produce an apparent cure and unfortunate timing can produce an apparent failure. In certain cases, factors in the life situation can decisively accelerate or impede recovery. An adequate control group would help to determine how much improvement is due to therapy and how much to spontaneous remission of symptoms. It would help also to discover what problems and what patients are the ones most likely to improve or recover without therapeutic intervention.

The wish for an adequate control group and the reasons for wanting it run like a refrain through discussions by research people seriously committed to the evaluation of psychotherapy—in statements such as the following:

J. McVicker Hunt points out that even if the necessary descriptive classifications were perfected, “we should still have information only about the first evaluative question, namely is there change associated with receiving psychotherapy? There need be no causal relationship, and . . . until we find out how frequently the changes associated with psychotherapy would occur without it, we cannot logically attribute them to psychotherapy.” (146, p. 239)

Kenneth Appel comments that “the therapeutic statistics of psychiatry appear to justify only the conclusion that the essential factors in cures are still unknown. Nevertheless, one gains the impression that therapy does something and is effective.” (9, p. 1153)

“The value of any type of psychotherapy remains to be conclusively demonstrated,” declares an article by three investigators, adding that the figure two-out-of-three (plus or minus about 10 percent) crops up so regularly and with such diverse treatments that cynics might conclude “psychotherapists make their living off the spontaneous remission rate. Yet every psychotherapist has had patients whose improvement followed so closely upon occurrences in the therapeutic situation as to make it highly unlikely that this was due to chance.” 7 (253, p. 343)

7The “two-out-of-three” remission rate has become such a refrain and rests on such debatable ground that it merits further comment, which is given in the Appendix.
The wish for adequate controls is reinforced also by studies in other areas where a control group has been used. For example, a battery of tests was given to two experimental groups and two control groups, before and after a mental health course. All showed significant changes. The authors note that if no control had been used the results could have been interpreted as highly gratifying (14).

A neat and simple demonstration of the value of comparison groups was afforded recently by a review of figures on enrollment in schools of social work. There has been some concern about the decline of enrollment in social work schools, and a good deal of soul-searching on the part of the profession. Comparison of the figures with those of other professional schools, however, showed that enrollment in schools of social work had not declined more than in other professional schools. Thus, though the fact of the decline is not altered, the implications drawn from it are changed by evidence that the decline reflects a tendency evident across the board and does not necessarily constitute, as had been assumed, a reflection on the social work profession (163).

In the field of juvenile delinquency, the Cambridge-Somerville study points up both the desirability and the problems of a control group (260). Its attempt at elaborate matching was extremely costly in time, money, and subjects for study. The therapists involved believed that about two-thirds of the children had "substantially benefited" by the treatment, yet no significant difference in number of court appearances was found between the treated boys and an untreated group. Accordingly, if benefit related only to juvenile delinquency as measured by court appearances, the treatment might be considered ineffective, despite the therapists' opinion. It was found, however, that many of the treated boys had benefited in ways not relating to juvenile delinquency.

**Difficulties of establishing adequate control.**—The controlled experiment, comparing groups that are all but identical except for the variable under study, is usual and accepted in the natural sciences. It is far from usual in research evaluating psychotherapy or social casework. And although its desirability is generally accepted, a variety of difficulties have blocked the use of untreated groups as controls.

Some of the obstacles grow out of problems already discussed. It would be necessary to show that the treated and untreated groups have no significant differences with regard to diagnosis and prognostic traits—including environmental factors. One means would be by elaborate matching of individuals. But matching on more than two or three variables is seldom feasible and in any case the relevant factors are not well enough identified, agreed upon, and controlled. Somewhat easier than matching individuals, al-
though still extremely difficult, is the matching of groups, with a view to obtaining similar distributions of traits in each group, even though a specific individual in one group may not exhibit the same cluster of traits as a matched individual in the other. One suggestion calls for matching individuals on the two or three most important traits and matching groups on all the rest. Even if the matching process were less difficult in itself, however, the needed information about some of the significant traits could be secured only by a process closely akin to treatment—leaving the control group something less than a true control.

The alternative to matching would be to compare the treated groups with a random sample of the same or a closely similar population who are untreated. But in most cases, for reasons discussed in preceding sections, it is very difficult to assure that the “control” group comes from a highly similar population. Various ways of meeting this problem have, however, been tried—unsatisfactory, for the most part.

One example is the “own-control” method used by the Rogers group. This method supplied the “control” by having some patients wait sixty days for treatment, testing them before and after this waiting period for signs of psychological change. One difficulty with this method was that at the end of the sixty days some of the prospective patients decided not to go into therapy after all. The findings of the study indicate that those who moved into therapy at the end of the waiting period tended to get worse rather than better during the interim; and that those who, after waiting, decided not to go into therapy tended to improve. This contrast suggests that—aside from other incomparables—the treated population differed initially from the untreated, so that the two-thirds who improve under treatment and the two-thirds who improve without treatment, represent a different two out of a different three (80, 120).

The finding also serves as a reminder that those who drop out of treatment differ from those who continue, and therefore should not be counted either as a control group or as “unimproved.” Other evidence reinforces the indications that those who discontinue treatment cannot be equated either with the untreated or with the unsuccessfully treated (16, 94, 180, 232, 281).

Although the results of the “own-control” study are most illuminating, the waiting period used in this study and several others seems entirely inadequate. Moreover, to wait sixty days after deciding on treatment and arranging to secure it within a stipulated time is not necessarily the same as not wanting it, or as wanting it and not being able to secure even a prospect of getting it. Thus the “own-control” group cannot be equated with the untreated. In addition, this method ends by removing its members from the control status, leaving no possibility for the followup comparison which is indispensable to any real assessment of effects (44).
Another suggested way of achieving a control group is to regard the people on an agency or clinic waiting list as controls. Critics of this method object that either the period is too short for adequate study and in the end the "controls" go into treatment, or else these individuals are unfairly deprived of the treatment they had been led to expect. Another suggestion has been to use as controls the people who have withdrawn from a waiting list. The evidence of the "own-control" experiment suggests, however, that those who withdraw before treatment is started differ in discernible ways from those who wait for it, and therefore cannot be considered an adequate control group.

It is often said that the only way to obtain a true control group of untreated patients or clients would be to go one step beyond the waiting list method and definitely withhold treatment from a random sample of applicants. The objection that this procedure would violate professional ethics is sometimes met by invoking the needs and canons of pure science, and sometimes by the argument that it is usually impossible to treat all applicants at any rate; if so, there is no harm in being systematic about who is to remain untreated (255).

The question of professional ethics is one of values and will have to be decided on that basis. It seems unlikely that many practitioners will feel comfortable with random refusal of service in the near future. Quite aside from professional qualms, however, there is some doubt whether this scientific "purity" would in fact produce the pure control desired. The people to whom treatment is denied will not necessarily forego it indefinitely. Some will seek it elsewhere—and the ones who do will be different from those who do not, thereby biasing the control. Some would have dropped out in any case and so (according to the "own-control" evidence) differ initially from the treated. Moreover, there is no knowing what effect the original act of getting on the waiting list and looking forward to treatment may have had; or the effect of being denied service over an extended period of time after being placed on the waiting list.

The Cambridge-Somerville and the New York City Youth Board studies were able to use this method because they sought out their clients and offered services to them for study purposes rather than operating in the usual manner of a service agency (260). In such cases, the offering agent is able to select his subjects at will and no guilt or criticism attaches to withholding treatment from those who have not sought it. He is hampered, however, by refusals to accept or continue with the offered service and by the freedom of the "control" group to profit by services offered elsewhere, all of which reduces and perhaps distorts his sample in ways that are difficult to assess.

Whatever the force of these points separately, together they add up to a potent reminder that human beings are not like metal or oil or gas. Opinions differ on whether their properties and life conditions can be disentangled and
assayed accurately enough to produce completely comparable samples of treated and untreated individuals—or samples similar enough to serve as adequate controls. Opinions do not differ on the need for a comparison group, but only on whether the simple classic model of treated and untreated, taken over without modification from the natural sciences, is the one that will prove most serviceable.

The need to achieve the best of all possible controls is the stronger, since a poor control can be worse than none, if it offers deceptive evidence—positive or negative—about the results of therapy. One kind of poor control is a presumably random sample which is not really random (172). Another deceptive control is the one which assumes that certain psychological or physiological characteristics, which have been regarded as evidence of neuropsychiatric disease solely on the basis of their known occurrence in a patient population, are not also common in the general population. The danger of such an assumption is pointed up in a study comparing the data obtained from four subgroups of a heterogeneous control group and one patient group. This study also shows the danger of assuming that "a sample drawn from a single and relatively homogeneous source will serve to represent a population of 'normals'. . . the psychiatrist, in his screening of the control sample, was impressed by the marked prevalence of so-called pathological indices among the nonpatient groups. As all five groups had a relatively high incidence of lassitude, weakness, restlessness and irritability, it would appear that these reactions are common in our culture and should not in themselves be considered as pathological. All but the career military group had had an equivalent amount of such developmental habits as nail biting and enuresis, which are generally considered predictive of a neurotic adjustment. . . . On the other hand, certain symptoms (projection, rigidity, hypochondriasis, and conversion symptoms), when present to a marked degree, were relatively unique to the patient sample." (36, p. 260–261)

Suggested solutions.—Because it is so difficult to set up an adequate control group of untreated individuals, an occasional suggestion is made to compare one type of treatment with another, rather than with no treatment at all (152). The requirements for comparison are of course the same, whether the comparison is between treated and untreated individuals or between two groups of individuals differently treated. If comparison is to be made between the treatment results of different agencies, practitioners or methods, it must be possible to show that the groups compared do not differ in ways that might affect the outcome—might even affect it more than does the treatment. That is, it must be possible to show that the compared groups represent the same population. "Random" samples do not necessarily solve this problem, since random samples from two different agencies would not be comparable unless
it could be shown that the significant characteristics were present in each to an equal degree—these same significant variables that so far have not been fully identified or described. No amount of randomizing will make it legitimate to compare the results of Dr. X with the results of Dr. Y if their patients differ in respects that might influence outcome significantly. And at present it would be very difficult to demonstrate that they do not.

Nevertheless, in some situations comparisons of two treatments, rather than of treatment and no-treatment, might solve the problem of comparable groups—through random assignment of cases within one clinic or agency to one or another type of treatment. If there is doubt about the relative efficacy of two kinds of casework or two kinds of psychotherapy, this random assignment of applicants to an agency or clinic would give some basis for comparative evaluation. For example, there is an honest difference of opinion about the relative efficacy of "conventional" and "nondirective" psychotherapy. A random assignment of patients, within one clinic, to the two types of therapy, could provide a controlled comparison—if each type of therapy is practiced by a skilled, competent exponent. One attempt at such comparison required that practitioners committed to one method practice another for experimental purposes. Such an arrangement destroys comparability in the skill, experience, and conviction of the practitioner—and there is abundant evidence that these are primary ingredients of successful therapy.

Another obstacle must be met before regarding comparative treatment groups as adequate controls, even within one population—namely, that often a type of treatment is indicated for one case and a different type for another case. The value of a type of treatment is not necessarily its applicability to any case, but rather its efficacy for certain kinds. It is possible to imagine four varieties of treatment each of which was the best possible for one kind of case. If, then, applicants are randomly assigned, the apparent success of each treatment type would depend on the frequency of certain kinds of cases among the applicants. A suggestion that might reduce this obstacle is to make random assignments within diagnostic categories, after the diagnosis has been made, to two types of therapy each of which is considered appropriate to the diagnosis. This method of course would profit by dependable diagnostic classifications, a horse already beaten to a pulp in these pages.

A satisfactory control, then, may be achieved by comparing two types of treatment, providing: (1) the treatment groups are part of the same population; (2) the treatments used are considered by the practitioners to be the correct ones for the individuals treated; (3) the practitioners of the two methods are comparable in ability, experience, and conviction.

Another variation on this theme is the suggestion to study the "inadvertent controls" provided by cases which must be treated by other than
the method of choice—whether because of "reality factors," difference of professional opinion, or whatever. "These can then be used for comparative studies with similar patients to whom the presumed ideally indicated treatment method is applied." (315, p. 251-252) Such cases may provide material for intensive study, as a background to more extensive comparison; they could hardly, however, be numerous enough to permit control of the variables that would have to be accounted for in order to establish convincing evidence of differences in outcome clearly due to treatment.

Zubin has suggested the idea of setting up "standard control groups" which could be used for various studies. "By inverting the procedure of matching controls to treated groups and instead match treated groups to available standard control groups, it may become possible to hasten the process of the evaluation of therapy in all its aspects." He admits that the proposal is "somewhat idealistic" and will have to "compromise with many realistic difficulties arising from our lack of knowledge about the true comparability of various individual patients who are to be matched with the controls." But he argues that without some adequate control group "we are reduced to accepting the judgment of outcome made by three of the most biased persons connected with the therapy—the clinician, the patient, and his family." (343, p. 63)

The requirements for control or comparison groups in evaluative research on psychotherapy are paralleled by the requirements that must be met in any attempt to evaluate the results of other efforts to bring about psycho-social change. They are identical with the requirements for social casework. Much of the research in juvenile delinquency has been criticized for lack of control groups and of the information necessary to establish comparisons or controls. A study purporting to compare two types of training schools, for example, is thrown off base at the outset because the "toughest" boys were regularly assigned to one school and the most tractable and promising to the other. That is, although the two samples were drawn from boys adjudged delinquent, they obviously represent two very different populations of juvenile delinquents. Some efforts are being made at present to compare different treatment methods for juvenile delinquents randomly assigned from the same population. Such comparisons are needed and wanted by the field. Their value will depend greatly on the success with which they meet the problems, mentioned above, of randomizing within diagnostic categories. It will be important also to make sure that the qualifications of the practitioners and the conditions prevailing for the different types of treatment are consistent enough and representative enough to constitute a fair test.

Moot points.—The control group, or its equivalent, is required to clinch a causal relationship—that is, to prove whether a given treatment or service is better or worse than another, or than none at all. Lacking a control
group, or a satisfactory substitute for one, we lack solid evidence that improve-
ment or cure after psychotherapy is the result of the therapy. There may be
studies strongly suggesting that it is, and this kind of indication may be suffi-
cient basis for important administrative and professional decisions. Neverthe-
less, without a control group, there is no proof. On this point there is
consensus among research people. However, since an effort has been made to
indicate opinion drifts, it should be stated that a number of other points in this
section are controversial in varying degrees, and therefore should be labeled as
the views of the author.

It is the view of the author that in evaluating psychotherapy or
social casework, comparison of results secured by different methods and services
offers more promise for developing adequate controls than does comparison of
treated and untreated groups. It may be that in some other kinds of efforts
to bring about change in individuals, the setting up of sufficiently comparable
treated and untreated groups is more feasible. Efforts to treat juvenile de-
linquency, for example—unlike efforts to prevent it—deal with captive subjects
who are not in a position to select, reject, or discontinue treatment. This
pathetic fact strongly modifies the control dilemma, and suggests that it may
be possible to set up untreated control groups in any population that lacks
autonomy.

It is the author's view that progress toward achieving adequate con-
trols is more likely to be made by recognizing and accepting than by ignoring
the differences between the materials involved in studying efforts to bring
about change in individuals and those involved in controlled experiments made
in the natural sciences. That is, we shall need to work out equivalents rather
than replicas of the classic treated-vs.-untreated model.

It is the view of the author that, in our present phase, efforts to identify
and define the variables that must be matched will contribute more toward
ultimate adequacy of control groups than will direct work on developing such
groups.

For ultimate long-term evaluation, some form of control or comparison
group will be indispensable. It is the view of the author, however, that cer-
tain judgments and decisions can be made on the basis of short-term evaluation
even without a control group. Suppose, for example, that a study shows 80
percent of the patients in a certain clinic improved after treatment; and shows
the practitioners, patients, and collaterals convinced that the clinic treatment
causd the improvement. The association of treatment and improvement
would not constitute proof of a causal relationship. Nevertheless, a Board
committed to the work of the clinic might consider the evidence strong
enough to continue or increase support, even without a control group. On
the other hand, a study of probation services might reveal a rate of recidivism
which, after careful analysis, indicates even without a control group that
something is wrong. A control group would be necessary to show that some other method would do better; but none may be required to show that the present rate fails to satisfy either practitioner or public, and that some way of doing better must be found.

In both cases mentioned, there is an implicit standard based on "common sense," "values," and assumptions (which may or may not be rooted in information and experience) about what should be expected from the treatment or service under study. Granted that many studies give findings less clearcut than these imaginary examples, it is still possible to measure results against the spelling out of implicit standards and expectations which derive in large measure from professional experience and from the public conscience. Such studies, of course, lack a very desirable ingredient. Yet they can be very useful, providing they state clearly their assumptions and tailor their recommendations to the nature of their data.

Some research people would hold, however, that although pre-evaluative studies do not necessarily involve a control group, no true evaluative study can lack one and be scientifically respectable. This view holds that the best available approximation to a control group is better than none—in contrast to the viewpoint presented here, which holds that a poor but pretentious control group is worse than none, since it tends to breed self-delusion about the limitations of the information secured. Merely to call it a control group can be deceptive if the known biases render it suspect, and if—as so often happens—the limitations stated in a ponderous methodological introduction are ignored in presenting a brisk set of conclusions.

Whatever view is accepted, the compelling need for adequate controls remains and will continue to prompt increasing efforts toward achieving that sine qua non of fully adequate evaluation.
IV. ABOUT THE FINDINGS

What is the meaning of the changes found?

So much effort goes into discovering what the findings are that this part of the work is usually referred to as evaluation. The true evaluative question, however, is: how good are the results that were secured? This question is answered, not by the findings alone, but by the findings plus the interpretation put on them.

Problems of interpretation divide into several segments. One involves the nature, degree, and stability of the changes manifested. Another involves the extent to which the study findings can be generalized. These aspects of interpretation will depend on answers to the research questions already discussed. Another group of interpretation problems, however, concern the extent to which the therapeutic outcomes reported can be regarded as gratifying or disappointing.

If a norm or standard exists, the findings of trustworthy evaluative research are gratifying or disappointing according to whether they meet, exceed, or fall below the norm. So far there are no tested and generally accepted norms for psychotherapy. It is often noted that the figure most frequently reported for improvement hovers around 66 percent and that the proportions vary according to diagnostic categories and certain patient characteristics (74, 146). However, the sources of these figures are so hedged about by qualifications, incomparabilities, and unknowns that they cannot be regarded as established norms. One function of the ultimate evaluation, in fact, would be the fixing of norms that could be trusted and that would provide a sound basis for comparison with the results of no treatment at all.

The extent to which results are or are not satisfactory could be established also by comparison with other methods or with the results of other practitioners. Again, the hope of achieving valid comparisons is a primary motive in efforts to perform evaluative research. Apparently the most satisfactory basis for judging whether the findings of an evaluative study are "good" or "bad" is beyond us until we are able to make dependable comparisons
between methods of treatment, kinds of therapy, performance of different agencies or practitioners, treatment and no-treatment.

There remains the interpretation based on what was expected or on a "common sense" judgment of the findings. This may be less than satisfactory but it is far from useless. The common sense estimate of reported results takes into account a number of different elements, which—like everything else about evaluative research—depend on the purpose for which the study was made. It is likely to consider what goes into the treatment or service under evaluation, as compared with what comes out. The purpose may require only a judgment concerning the proportion helped or cured, regardless of cost. On the other hand, it may require a judgment of the proportion helped in relation to the cost of the treatment.

Most people assume that for any effort to bring about desired change, at least as many should improve with service or treatment as improve without it. If two out of three improve without treatment, then either more must improve with treatment or else each one must improve more. Otherwise it would not be worth while. Some varieties of problem, however, are believed to be all but incurable. For these, definite improvement among 5 percent would be a distinct achievement.

Studies are constantly and reasonably challenged because they have no norm or comparison group as a standard against which to measure results. There is no question that a valid norm, control, or comparison group is better than none. There is great difference of opinion, however, whether an invalid one is better than none. Some honestly hold that half a truth is better than none; others as honestly insist that at times it is better to know one does not know than to delude oneself that a half-truth is a fact.

Those who prefer doing without a possibly deceptive half-truth must content themselves with checking the reported outcomes against expectations and value judgments. In the case of an adoption service, for example, if one out of three cases appears to work out unsatisfactorily, a good many readers will not ask for a control group to prove we must do better. It may be that an equal proportion of "own" homes do not work out well for children. Nevertheless, if a home is to be selected for a child, and if the child is already handicapped by the need for adoption, many will consider it reasonable to aspire to a better adoption outcome. If nine out of ten adoptions work out well, there will be incentive to find out what is wrong with the tenth case, but no further figures would be required to prove that this outcome is gratifying.

Unfortunately, many findings are not this clear-cut, but lie in an area that allows for different interpretations. In such cases, much will depend on the values and convictions of those involved. Those who have experienced cure or improvement of a treated problem, either as practitioner or as recipient
of the practitioner's effort, will not easily be convinced by unfavorable findings that efforts to bring about psycho-social change are worthless—although they may be convinced that the efforts should be improved.

A number of experimenters warn against taking too dim a view of temporary improvement. One group points out that "the importance of temporary improvement should not be underestimated. The fact that a diabetic, brought out of coma by insulin, will relapse if the insulin is discontinued, does not mean that insulin is to be dismissed as affording merely temporary relief." (253, p. 350) Another also invokes the medical analogy, commenting that the chronicity of many personality disorders is well established, and indicates "not that psychiatric therapies are worthless but rather that they are similar to other medical treatments for chronic diseases such as asthma, or diabetes, useful in alleviating acute attacks of a chronic illness whose over-all therapy in terms of maintained general health is yet to be promulgated." (313, p. 145-146)

The difficulty of interpreting findings produced by evaluative research is compounded not only by all the elements mentioned above, but also by the fact that this type of research, even more than other types, provokes strong anxiety in practitioner, administrator, board member, and researcher. The reasons are different for each and are too obvious to require recital. The fact remains that, in addition to all the methodological difficulties, this is among the most difficult kinds of research psychologically. One kind of difficulty has been discussed by Blenkner (25). A number of other kinds have been noted by others, and some seem to have gone undiscussed so far. Most of these discussions are interesting and many are helpful. It seems likely, however, that the psychological difficulties specific to evaluative research will be helped more by the solid research studies of these same able discussants than by their insights. When researcher and practitioner have won through demonstrated achievements the ability to speak of their still unconquered areas as openly, hopefully and undefensively as physicians speak of cancer or the common cold, everyone will be better off—and so will research.

Were there unexpected consequences?

Consequences of the means employed?—People who seek help in bringing about psycho-social change—or who have help thrust upon them—often find that it affects more areas of their lives than the one in which the problem was supposed to reside. A familiar story tells about how a man enters
psychoanalysis to help his migraine headaches or his ulcer and comes out with a cure and a divorce. Equally familiar is the case of the unmarried person who becomes able to marry apparently as a result of casework or psychotherapy. Another occasional aftermath of such treatment is that grown children leave the parental home; or that the timid little bookkeeper strikes out and gets a new job—or quarrels with the boss and loses his job.

These byproducts of treatment or service may seem desirable, undesirable, neutral, or even both good and bad. An example of the double-valenced type is the “altruistic” person who becomes more self-seeking and demanding after treatment but who also feels happier; or the brilliant and stimulating companion who settles down to a more contented life for himself but offers his friends less entrancing entertainment.

This type of example could be multiplied endlessly for every kind of treatment or service. It is noted here merely as part of the evidence that must be included in a study and considered in final interpretation of the findings.

Desirable byproducts of the means employed to bring about change fall under the currently popular term, “serendipity.” The word, recently revived, was coined long ago in allusion to a tale, “The Three Princes of Serendip,” and means the finding of valuable or agreeable things not sought for. The heroes of the story were always discovering in their travels, by chance or by sagacity, desirable things they had not actually been seeking. Examples of serendipity are also found among unexpected consequences of research procedures, discussed below.

Consequences of research and researcher?—A little old lady who did her courting in the nineties likes to tell about her Grandmother’s efficient chaperoning. Grandmother would just move into the living room where the two young people were sitting on the sofa. “Now you two young folks go right ahead and visit,” she would say, “and don’t pay any attention to me. Just act as if I weren’t here.” There was some difference of opinion in the family about whether Grandmother thought they were acting as if she weren’t there, but there was no doubt in anyone’s mind about whether they really were acting that way. It takes a wary eye to be sure that a research project is not playing a role like Grandmother’s.

The best known example of research affecting the material under investigation is in the field of industrial psychology. The problem under investigation was the relation of lighting to worker productivity in a certain plant. Careful experimentation showed that as the brightness of the lighting increased, the production rate went up. When the lighting was gradually diminished, however, the productivity rates did not decline until the women were working almost in the dark. It was finally concluded that the significant
factor was not the lighting at all, but rather the psychological situation created by the study (275).

Since analogous effects of the research situation can occur in many kinds of social study, it is highly important to keep this possibility in mind throughout the planning, analysis, and interpretation phases. The possibility is especially strong when practitioners, patients or recipients of service are directly and consciously involved in the research. Any distortion introduced by such effects is, of course, maximized by refusal or inability to recognize them. If they are clearly perceived and assessed three possibilities are open: (1) They may turn out to be so slight as to affect the findings very little; (2) they may be dealt with in ways that reduce or eliminate them; (3) they may require drastic change of study plan, which is painful at any time but far more painful and far more expensive later than sooner. If on the other hand they are ignored, they may either distort the study findings, or open the study to cruel criticism—or both.

A number of people engaged in psychiatric research have urged greater attention to the effects of the research situation, as differentiated from the effects of the treatment under study. One comments, "The application of a sphygmomanometer probably changes the blood pressure; questioning a patient about his hallucinations undoubtedly affects the nature of the voices or visions. Some excellent methods have been devised for dealing with this problem; but in many fields of psychiatric investigation there still exists considerable naivete in assuming that the effect of the test situation itself can be neglected." (172, p. 476)

The effect of the research situation upon material is a special problem in relation to the validity of psychological tests of emotional or personality factors, briefly touched upon in chapter III. A psychiatrist engaged in research has commented that "it might be quite a problem to evaluate how much of the change in a patient was due to the testing program, how much to the treatment being studied." (102, p. 47) The testing program includes, of course, reaction to the test situation and to the tester as an individual. This problem becomes especially acute, for those who grant its existence, when such tests are administered repeatedly to the same individuals over a period of time. Some investigators apparently consider the effects of repetition negligible. One, for example, recommends "measuring a patient on a certain set of dynamic traits, preferably by objective test methods, and recording strengths on these variables from day to day over about 100 days ..." with no reference to the possible effects of 100 daily repetitions of the same tests (53, p. 8). On the other hand, an occasional study that employs repetitive testing will report the elimination of certain tests because the effects of repetition have become obvious.

Such effects are likely to be especially discernible in projective tests.
if they are administered often enough for the tested individual to “catch on.” In Roger’s case of “Mrs. Oak,” the analysis of the fourth administration of a Thematic Apperception Test includes the comment, “the general impression of this entire record is that the client ‘sees through’ the purpose of the TAT and in rather good humored fashion goes along. . . .” (276, p. 271) The analysis appears not to be inhibited by the patient’s perception, although the reader is not told what allowance, if any, was made for it. In any case, recognition of the effects of repetition is one step away from the “naivete” of ignoring them (237). A careful investigator offers the reassuring word that “practice effects on some tests can be separated from other fluctuations as a trend factor. More work is necessary on (a) the types of tests which show promise and (b) the assumptions underlying repetitive methods of intraindividual measurements. . . .” (201, p. 392)

Some ways have already been mentioned in which the research situation can affect the material that is to be analyzed, whether psychological tests are employed or not. One is through the patient’s desire to please the therapist, or to convince himself that he is cured. Some critics of psychotherapy argue that psychiatric treatment is a learning situation and that the learning is chiefly verbal—patients simply learn to say the right things (72). Another way is through the therapist’s emotional stake in therapeutic outcome. Still another is through the patient’s reaction to the tester or interviewer, as well as to the research situation. To be convincing and to be sound, a study plan must take all these possibilities into account—at the least recognizing them and at the most attempting to counteract or circumvent them (95, 131, 187).

The research situation can also affect the material under study through the very questions raised by the research staff and the procedures initiated by them for discovering by what means the desired change is to be brought about. Much of the enthusiasm for the byproducts of research derives from the effects on practice of persistent demands for definitions of goals and problems.

Many reports mention the practitioner’s belief that his goals have been sharpened and his methods enriched through the need to make them explicit in answer to research questions. During one study, group therapy sessions were recorded by an observer, trained in social science but not in the practice of therapy, whose sole function was to keep notes for research. In the beginning, this observer was considered a therapeutic liability to be suffered only for the sake of the research results. Before long, however, he tended to become an asset, both because the patients took his presence as a sign that their meetings were interesting and important and because his questions made the doctors more clear about what they were doing, responding to, and expecting. “From the standpoint of therapy, the doctors found that the making and discussing of inferences gave them a better understanding of the
complex situation with which they were dealing. . . .” (259, p. 21) Some of the doctors said they would never again be without such an observer. Yet, as others comment, clinical enrichment can be research contamination (175, 318). What the group therapists were doing under the stimulus of the observer may not have represented their normal practice. This fact may represent a gain for practice and may not change the research results greatly, but it merits recognition.

A number of years ago, reputable public opinion pollsters claimed that the interviewer had no effect on their polling results, because “we have taken scientific precautions against such influence.” The claim was probably honest, based on careful wording of questions and what were then thought to be adequate interviewer training and sampling methods. Eventually, however, the evidence become too strong for the prevailing faith in the “scientific objectivity” of current interviewing, and a highly ingenious program of studies was carried out, resulting in evidence that in many situations the interviewer—however well trained and however conscientious—did indeed influence the content of the interview to a measurable although not necessarily a decisive extent. These studies, by recognizing and assessing the actual influence, laid the ground for counteracting “interviewer effects” to a considerable extent, and making allowance for those effects which could not be controlled (150).

Such experiments are likely to be undertaken by those more interested in improving than in defending current research methods. Yet their end result is likely to make the methods more defensible.
V. AFTERWORD: SOME PRACTICAL IMPLICATIONS

WHERE DO WE STAND?

This publication has reviewed some questions that must be answered for fully satisfactory evaluative research on efforts to bring about social or psychological change in individuals. To the best of our knowledge, no study has ever fully answered all of these questions, and it will be many years before all of them could be answered satisfactorily.

Merely stating the questions defines a dilemma between what is wanted right now and what can be delivered right now. One view of this dilemma is suggested by three terms applicable to three kinds of evaluative research:

1. "Ultimate evaluation" refers to the kind that everyone wants most. The practitioner, the public, the administrative official, the supporting contributor, all want right now evidence of the degree to which the practice or service under examination helps the people it serves. With regard to psychotherapy, they want to know the effectiveness of psychotherapy in general or of a particular school of psychotherapy. Similarly, those working with juvenile delinquents want to know, for example, the effectiveness of probation services and the relative effectiveness of different methods of probation. Or they may want the answer to analogous questions about training schools or about measures for preventing juvenile delinquency.

If the preceding discussion has conveyed its intended meaning, then it is clear that research cannot produce here and now the "ultimate evaluation" of efforts to bring about psycho-social change in individuals. It is clear also that the evaluative questions may need to be reformulated and sharpened if "ultimate" answers are to be secured.

2. "Pre-evaluative research" refers to the kind of studies that will be necessary to answer the questions that must be met before fully satis-
factory evaluative studies can be made. Pre-evaluative research will be needed on most of the questions listed before ultimate evaluation will be feasible—questions about what change is to be produced, in whom, by what means, by whom, etc. Such research will contribute to practice as well as to ultimate evaluation. It will contribute also to reformulating our ideas about what is desired from ultimate evaluation. As diagnostic classifications and treatment goals and methods are more sharply defined, for example, the focus of the evaluative question is likely to be sharpened so that we may no longer be asking, how effective is psychotherapy or social casework in general but rather, how effective is such-and-such a kind of treatment in producing such-and-such changes in such-and-such kinds of people (258).

3. “Short-term evaluation” means research that can be accomplished within relatively few years. Such research is possible and useful here and now. It is possible without extensive pre-evaluative research to give properly qualified answers to properly qualified questions about the effectiveness of treatment or service by a specific agency or individual, with a specified population. The requirements of proper qualifications have been discussed at length. In brief, properly qualified answers would state clearly the limitations of the methods employed, observe the rules of evidence, make no generalizations beyond the limits of the data. Short-term evaluation can often be done in a way that meets research requirements, fills immediate need, and at the same time contributes to pre-evaluative research. It cannot, however, give the answers that many people want most. These require ultimate evaluation, which in turn demands many pre-evaluative studies.

Apparently the kind of evaluative research under discussion here is at an interesting cross-road where it seems necessary to proceed in both directions at the same time. Fortunately, there are enough travelers to deploy forces along both routes. It is necessary to continue with pre-evaluative research in the effort to come nearer the long term goal of ultimate evaluation—recognizing that this goal may have changed its outlines somewhat before we finally reach it. It is also necessary to do whatever can be done with more approximate and less complete efforts at short-term evaluation, as background to immediate steps and decisions. Either type of research, evaluative or pre-evaluative, can contribute to the other type—if and only if it observes the rules of evidence, explicitness, and restraint that are binding on research of any type and at any level.

Some practical implications of the points brought out in the preceding discussion can be summarized under a number of “do's” and “don'ts” for evaluative research. Since most of these have been discussed rather fully in the report, only a few call for extended comment here. The marked unevenness in space given to each one, then, does not reflect an estimate of their
relative importance, but merely differences in the amount of discussion that seems called for at this point.

**SOME RESEARCH "DON'TS"**

*Don't undertake evaluative research if the purpose can be served by some other kind.* It is expensive, time-consuming, difficult, and not always successful. If the purpose is to contribute to professional knowledge and understanding, a "pre-evaluative" study is likely to be more directly rewarding. If an evaluative answer is urgently needed, the answer can often be secured—or approximated—by quicker, more feasible and less costly types of research, such as fact-finding or survey studies. Accordingly, short term evaluation should be undertaken only if thorough investigation of the purpose shows no ingenious method of circumventing an outright evaluative study. For example, a proposal was made to evaluate an ambitious program of individual treatment theoretically under way at a training school for boys. A simple survey revealed, however, that the current staff lacked the qualifications necessary to carry out the program as formulated, and in addition labored under a time schedule which precluded giving the boys any but the most superficial, perfunctory, and unindividualized attention. In this instance, analysis of actual operations, as compared with stated objectives, served the evaluative purpose.

*Don't undertake evaluative research unless adequate resources are available.* Adequate resources include money, staff, and time, with assurance of continuity since interruptions can be wasteful and also harmful to final results.

The question is often raised, can adequate research be done in an organization that does not have a full-blown research department? The most straightforward answer seems to be that full-blown research requires full-blown research people. It is usually a mistake to think that satisfactory research can be done by agency staff with a part-time research consultant. The demands are too heavy. Any substantial research project requires full-time research staff plus full collaboration from practitioners.

Agency administrators who undertake research are often unprepared for the amount of practitioners' time required by a research project; but the more able the research staff, the more consultation they are likely to want from the practitioners. The canny administrator will count such time as part of his research budget and will not attempt to add a research project, research department or research worker without making due allowance.
either by increasing the number of practitioners or decreasing the number of their cases during the time they are involved in research activities. For example, the report of a study of short-term cases says that preparation of schedules and instructions took many months of discussion and tryout by the Planning Committee—including highly trained caseworkers; and that once the study was under way the administration allowed participating staff members twenty to thirty minutes after each interview to fill out schedules, cutting down the caseloads accordingly for the individuals involved during the time of the study (180).

Administrators are probably less surprised than they used to be at the length of time that elapses before a research project is completed. It may, for example, take six months to track down the subjects for a modest follow-up study (195). If electric recording is used, the requirements both in time and money approach the fabulous. The Rogers group in Chicago report that the transcription of one forty-interview case ("Mrs. Oak") filled over 300 single-spaced typed pages (112). The same investigators found that it took over 500 man hours to collect and transcribe the data from one typical thirty-interview case and the matched control individual—not counting analysis of data (277).

Don't count on using existing agency records as the sole source of data for an evaluative study. Case records make such interesting and instructive reading that it is hard to believe they would not furnish a useful basis for evaluative research. Yet again and again investigators find that they do not. The needed items of information are seldom if ever included in every record. When present, they are seldom comparable in explicitness, detail, and documentation. Exorbitant amounts of time may be spent trying to discover or deduce the most elementary facts about a case. If relatively recent records are used, there may be serious problems in making them available for analysis—especially if closed cases are frequently re-opened by reapplication for service.

All this is highly regrettable since no one doubts that there is gold in all the mountains of case records piled on agency shelves. Yet, on the basis of experience to date, most researchers prefer if possible to work out recording forms and procedures in advance. If this is not possible, it usually becomes necessary to supplement existing records with other sources of information. These are merely the more superficial problems arising from attempts to base research on existing records. Problems of reliability have already been noted, in chapter III.

Don't indulge in lopsided research. It does not pay to lavish time and money on being extremely precise in one feature if this is out of proportion with the exactness of the rest. For example, it profits little to go to great
lengths in insuring the quality of sample and reliability, if criteria are fuzzy and definitions ambiguous. This type of imbalance often tempts the researcher to report as if the whole study were impeccable because of the good sample and reliability—forgetting that "no study can be better than its criteria." Part of the secret of appropriateness and harmony in design and pretensions is the recognition that research offers not one model but many models and that the plan must depend on the purpose of the research.

**Don't be afraid of unpretentious research.** Better be simple, clear, and forthright about limitations than to employ techniques more ambitious than the data warrant. The value of frank opinion material is not to be minimized in connection with short-term evaluative studies (59). If therapist, patient, collaterals, and record analyst agree that certain types of clearly specified change have taken place the evidence is not to be belittled, even though it cannot accurately be described as "objective." This point is brought out by Brewster Smith in discussing the evaluation of the exchange of persons program—with a reminder of the close relation between purpose and method. "When evaluation is primarily for the benefit of the programme's own administrators, skilled judgment may be substituted for proof at various points in the ideal pattern of evaluation, with great saving in cost and feasibility. The ideal requirements remain a useful reminder of the points at which judgment is being substituted for evidence; they indicate where cautious interpretation is likely to be in order." (302, p. 391)

**Don't be confused by loosely used terms**—such as reliability, objectivity, statistical significance. Such terms represent important research elements. But if consumers—and researchers too—were more clear about what these words really mean, they would be less likely to assume that reliability insures validity, that counting insures objectivity, and that statistical significance insures significance of content (52, 294, 337).

**Don't be misled by fantasies of a neat, precise, utterly objective social science** modeled after a naive conception of the natural sciences. For one thing, the materials studied do not lend themselves to identical techniques. For another, the precision of the natural sciences—although far beyond that of the social sciences—is less absolute than we sometimes assume.

**Some research "Do's"**

**Do bring the researcher in early enough and fully enough.** A chronic menace to sound and useful research is tardiness in enlisting the research director. It is not enough for him to be in on the ground floor. He
must help to dig the ground and lay the foundation. This means that he
must help to investigate the need for the proposed study, to formulate the
purpose, to determine whether the purpose as formulated can be served by the
type of study proposed, or by any feasible research. All this represents the
foundation that must be solid before he begins to work out the study plan.

Do include "intellectually hospitable research specialists and
practitioners on the research team. This requirement is often taken for
granted but its full meaning is seldom recognized in advance. Successful in-
terdisciplinary research requires (1) selection of individuals qualified by train-
ing, experience, and temperament for this kind of research, (2) allowance for
sufficient practitioner time, (3) readiness to cope with the classic problems
of interdisciplinary research which competence and experience can mitigate
and cope with but cannot obviate.

The need to allow for practitioner time, if research is undertaken in
a practice agency, has just been discussed under adequacy of resources. If
practitioners are employed full-time for research, this particular problem does
not arise. Its absence, however, does not diminish what has been referred
to as the "classic problems" of interdisciplinary research and the need to
select individuals qualified to cope with them. These are discussed later
(p. 88).

Do appreciate the rewards to be gained through pre-evaluative
research. There is no denying that if the researcher is left free, he is likely to
choose pre-evaluative rather than evaluative research for his own activity. A
number of instances have been mentioned, and many more could be cited, of
researchers turning from evaluative to pre-evaluative projects because they
became convinced (a) that the most satisfactory kind of evaluation could be
done only after an extensive and intensive tooling-up period and (b) that,
other things being equal (though usually they are not), pre-evaluative re-
search offers a more direct contribution to better professional practice and to
better understanding of people. Many of the researchers interviewe~ and the
theoretical articles reviewed during our survey emphasized the need to know
more about just what we are doing before we try to say just how well we are
doing it; and not one favored trying to find out "how well" before doing
more work on "what." This means, on the one hand, attempts to perceive
and describe the significant factors in the problems treated, the individuals
treated, the methods used, the therapist as an individual, the treatment process.
On the other hand it means that an effort will be made to describe "change"
rather than "improvement" or "deterioration." That is, to tell what change
occurs before trying to tell how desirable it is (23, 116, 241, 277, 333).

A good deal more pre-evaluative research has been done in psycho-
therapy than in social casework. Repeated reference has been made to re-
search on diagnostic categories, on treatment process, on patient and therapist variables related to treatment outcome, etc. But as more serious and more large scale research efforts are getting underway in social casework, the number of pre-evaluative projects seems to increase.

Research in other areas, such as juvenile delinquency, would also profit greatly by more emphasis on pre-evaluative research. Efforts to review evaluative research in juvenile delinquency have suffered from inability to trust or to compare the results of studies that fail to meet elementary research requirements, and at least to some extent the failure is due to lack of sufficient pre-evaluative research. In this area, as in many others, the most promising current efforts seem to be veering toward supplying the pre-evaluative gaps—such as the lack of adequate classification for the many kinds of behavior problems lumped under the term “juvenile delinquency.”

There is room for a great deal more pre-evaluative research in juvenile delinquency. It may be suspected that a simple descriptive survey of the treatment of juvenile delinquents throughout the country would be more effective, less costly and less long in completion than abortive attempts at evaluation. If the adult public, nervous and angry over “the juvenile delinquency problem,” were faced with a factual account of exactly what happens to young people adjudged delinquent, of the facilities available, and the qualifications of working staffs, the focus of attention might shift—and the results might help to reduce juvenile delinquency. Thus, relatively simple pre-evaluative research could also serve an evaluative purpose.

This is not to imply that pre-evaluative research always offers an immediate evaluative byproduct. Such a claim would be unnecessary, for its direct products are valuable enough in themselves. They have been referred to so frequently throughout this report that they need only be mentioned here. Improved diagnostic classifications, explicit statements of goals, improved descriptions, and definitions of therapeutic methods are needed and wanted for practice as well as for evaluative research. Examination and analysis of practice are useful not only to the administrator but also to the practitioner. A number of studies report that practitioners say they have gained new angles and insights through the very process of answering the endless questions of the researchers about what they do, why they do it, what signs or clues they respond to, etc. (130, 259). Testimonials to the helpfulness of a “research look” at practice are a familiar part of research reports. Ralph Kolodny has written in some detail of the practical gains an agency reaps through the research process, listing a number of concrete effects “which even the procedure of simply ‘thinking in research terms’ can have upon the day-to-day practice of a group work agency.” (183)

Do appreciate the value of coordinated efforts. The questions that
press for answer are far too vast and complex to yield to the efforts of a single research project or organization. Until recently the kind of research under discussion here has tended to be piecemeal and to exist as if in a vacuum. Increasing efforts are evident to digest and build on what has been learned from previous research, to test out research instruments by using them in new settings or in followup studies, to test out findings by duplicating studies or by repeating followup studies after a further period of years. This kind of interrelation between research projects represents coordination through time. Somewhat less frequent but no less desirable is coordination through space—between agencies or individuals, working on parts of one project or simultaneously undertaking similar projects in different places. The Rogers group in Chicago is a notable example of coordinated research within one structure (146, 215). A few research projects in social casework represent coordination through space.

The promise of new types of research coordination is apparent in the current plans and directions of several national organizations, such as the United Community Funds and Councils of America, the Child Welfare League of America, and the Family Service Association of America. Such organizations, each with a research department in its national office as well as in a number of local agencies, and with a great number of local affiliates from which to derive research materials, are in a position to pioneer—for example in pre-evaluative research, drawing on material from constituent agencies some of which may not have their own research departments.

Do appreciate the value of the research prerequisites: systematic study and exploration. Pre-evaluative research itself has important prerequisites, namely, exploration and clarification of terms, processes and concepts, based on review of cases. This kind of exploration can be begun by small agencies and by individuals working alone, with great value for research and for practice. The results of their work can then be utilized and tested in more rigorous research undertakings.

Genevieve Carter has discussed in detail the fact that "concept clarification is one of the important outcomes of all social work research," pointing out that while it can be the objective of a large research project it can also be undertaken by individuals working alone (49, p. 300). She cites as examples an article by Fritz Schmidl inquiring into what is meant by "supportive treatment," and one by Lionel Lane examining the meaning of the "aggressive approach" in casework (189, 288). Each author proceeds to evolve an "operational definition" of the term under discussion by analyzing concrete examples to determine their elements. Carter points out the beneficent cycle represented by such study, since "research clarifies concepts and... clarification of concepts makes research possible."
Systematic case review for another purpose also offers great value for both research and practice, and also lies within the scope of the small agency or the individual working alone. It is possible, for example, to single out for analysis one type of case—say, cases concerned with one type of problem, or cases in which two family members are treated by two different practitioners—and by systematic examination to identify elements and characteristics not previously recognized or comprehended. A survey of research on the short-term case cites a number of "case reviews" which, without elaborate research techniques, provide information not perceptible on the basis of experience with a varied case load (298). A number of interesting analyses have started with the superficially homogeneous category, short-term case. One of these took for its point of departure a large study of short-term cases, comparing the national figures with those of the author's agency, and considering the various types of brief service case in more detail and with more commentary than the quantitative survey permitted. This detailed consideration, based on cases known to the author and her agency and drawing also from the larger study, built up a case for not viewing with alarm the large number of short-term cases, for modifying the methods of categorizing and reporting such cases, and for profiting by sharpened diagnostic differentiations and by greater differentiation in the functions of intake (309).

To study one particular kind of case has much the value of a one-man exhibit of paintings by an artist whose work was seen previously only as part of large and heterogeneous showings. Characteristics and interrelations emerge that were not recognized before. These may be characteristics of cases or characteristics of treatment. For, as has been remarked above, actual practice does not always conform precisely to the administrator's or the practitioner's conception of what is being done, and sometimes what seems to be a case characteristic turns out to be a result of the way a certain kind of person is likely to be treated. A systematic case review is apt to hold a number of surprises.

Many kinds of systematic study are possible, depending on the problems of most immediate interest to the agency. Although the examples mentioned come from social casework, the same kind of study can be fruitful for any agency or service attempting to bring about psycho-social change in individuals. For example, many of the pre-evaluative questions that haunt researchers in the field of juvenile delinquency can be approached by a modest exploratory case review, laying the basis for further steps toward getting an answer.

In other words, concepts can be clarified, definitions can be made explicit, characteristics of case types can be brought out without elaborate research procedures. One does not need an ambitious project in order to begin evolving needed research tools. Systematic study, without elaborate
methods or pretensions, can contribute to the general reservoir that must be built up before "ultimate evaluation" can be achieved. To do well what lies within available resources will contribute far more to the agency and to the field than to do badly what requires time, money, and staff beyond the available resources. It goes without saying that in such investigations the conclusions and interpretation must be limited by the nature of the investigation. Such studies will not answer the pre-evaluative questions, but they will help evolve the tools required for answering them. It should be added that, on the whole, proper limitations are more likely to be observed in this type of unpretentious study than in one that aspires to a scope and definitiveness beyond its actual capacities.

INTERDISCIPLINARY RESEARCH

It has been assumed throughout this publication that satisfactory research on efforts to bring about social or psychological change in individuals will require the viewpoints of both practitioner and researcher. The necessity that this type of research be interdisciplinary has come to be taken for granted by most people involved in it. The experience of doing interdisciplinary research, however, is definitely not taken for granted by those who have it. On the contrary, interdisciplinary research seems to resemble love in the fact that it is vastly written about and yet when it happens to a person it feels new, unexpected, uncharted. It fills him with a desire to tell others all about it, and very often he does—without quite realizing that this is what all those pages and pages he has read were about, and that his testimony will probably be as fruitless for his listeners as the accounts of others were for him. Again and again, with a sense of discovery, both researchers and lovers try to explain what it means, what it demands, and what it feels like. But for all that, no one ever seems able to prepare anyone else or to help him avoid the pitfalls so often and so eloquently described.

Interdisciplinary research, unlike love, has standard early phases that are usually wasteful and often painful, and that seem avoidable to those who have lived through them. The many pages written on the subject represent an effort to help others avoid these phases. Yet these efforts, on the whole, seem more successful in producing hearty agreement from those who have lived through it than in forestalling interdisciplinary growing pains for those who have not (8, 103, 209, 258). Perhaps it should just be taken for granted, then, that people will have to go on learning by experience rather than precept some outstanding facts about interdisciplinary research, such as:
Different professions or disciplines use different languages and it is necessary for each to learn what the other means by the words he says: what he means by the unfamiliar words, and even more important, what he means by the familiar words he uses in an unfamiliar sense. For example, as Cunningham has pointed out, the word “case” means different things to doctor, lawyer, caseworker, watchmaker, distiller (68). Similarly, members of different professions—or different schools within the same profession—may find they mean rather different things by such words as “projection,” “empathy,” “community,” “generic,” “functional.”

Different professions have different conceptual constructs and frames of reference. Although at the beginning the unfamiliar one is apt to look merely distorted and askew, after some time it may begin to seem a valuable addition to one’s own conceptual apparatus. For the researcher, the best way to become familiar with the practitioner’s concepts and frame of reference is to become familiar with the practice, through reading the literature, reading records, interviewing staff and—if possible—through observation. “The social research scientist,” warns Pollak, “will do well to use his tools of measurement only on the basis of full understanding of . . . practice.” (237) It is impossible to overemphasize the need for the researcher to become familiar with the material he is to investigate before he begins to plan research. Fortunately, this need is increasingly recognized.

Different professions have different attitudes to the apparatus of research. Often the practitioner is convinced that pre-structured schedules or note-taking during an interview will distort material and interfere with rapport, to the detriment of the product. Often the researcher views these misgivings as mere disciplinary myopia. Sometimes the practitioner tries it out and becomes convinced that the apparatus he resisted is useful and helpful, not only to research but also to practice (238). Sometimes, on the contrary, the researcher secures evidence that the apparatus does in fact hamper full and undistorted communication; or that it inhibits the practitioner enough to prevent his best performance. The truth that members of interdisciplinary teams ultimately hail as revelation is that neither side is always and inevitably right. Cases can be cited to support either one, and each situation must be worked out on its own merits. The great gain comes when the team members become emancipated enough to realize this and therefore to concentrate on the situation rather than on defending the doctrines involved.

Collaboration is a two-way street. There is a tendency to structure a team—implicitly or explicitly—as a hierarchy of disciplines, and to assume that intellectual illumination can only flow downward. It is proverbial, for example, that teams composed of social workers and social scientists tend to
assume that the social scientist is there to shed light and the social worker to receive it—even though the social worker also believes the social scientist is forever denied certain basic insights. It usually takes a long time before both recognize that any collaboration is a two-way process; and only when the social scientist begins to recognize that there is light for him to receive as well as to shed does collaboration become fully fruitful.

A recognition of the other discipline's basic value enhances readiness to adapt to its preferred and most effective method of functioning. An outstanding example of such adaptation between research psychologists and practicing psychiatrists is a method worked out at the Menninger Foundation. The researchers recognized that the clinician found great difficulty in rating a patient on an absolute scale, with regard to general progress or to specific variables such as manifest anxiety, ego strength, etc.; but that he apparently found much less difficulty in saying which of two patients showed more or less of the element under consideration—i.e. anxiety, ego strength, etc. Accordingly, they worked out a system of paired comparisons with which the clinician was comfortable and able to give his most reliable judgments; and which the researchers could manipulate to give the kind of rank order they needed among the patients under study (204).

The method of paired comparisons may, of course, become extremely arduous if large numbers are involved. The example is mentioned, not for the specific method—which may or may not be fruitful in a given situation—but rather as an instance of a constructive approach to problems of interdisciplinary accommodation.

Another exercise in interdisciplinary accommodation occurred during a study that included intensive interviewing by highly-trained caseworkers. The research members of the team favored a "standardized interview" in which certain points must always be covered, although the order and manner of introducing them were not pre-determined. The caseworkers felt that such hard and fast restrictions would prevent free use of their best skills. Accordingly, they were asked to conduct some preliminary interviews without restrictions, merely covering the general topic as seemed best to them for the subject under investigation. When the records of these untrammeled interviews were examined, it was found that most of the points originally included in the standard outline had been covered by all of the interviewers, and all of the desired points had been covered by some. Thus it became clear that the requested outline represented points a competent caseworker would be almost certain to cover, and that the "standardization" merely insured against omitting one or another accidently here and there. Seen in this light, the research stipulation became less burdensome so that at the end of the study the caseworkers, with one exception, declared that the outline had been no burden or restraint at all.
Flexibility and ingenuity in adapting or presenting research tools is fostered by clear comprehension of and regard for the special aptitudes of the practitioner. The enemy of such flexibility and ingenuity is the assumption that social science has ready-made answers and all research materials must be stretched or lopped to fit the Procrustean bed of current research techniques—with no thought for the possible value of the hands or feet that might be tossed aside in the process.

A reciprocal recognition of value helps the practitioner to countenance the researcher's need to search for separate elements—however interwoven—in a complex whole. Within reasonable limits, the practitioner's fear that the whole will be slighted in favor of its incomplete parts is valid and is likely to be shared by any researcher worth his salt. When carried to extremes, however, it is a familiar and serious problem in interdisciplinary research. When this fear is tempered by regard for the special values of the researcher's field, the analytic research approach has much to offer to practice—according to the testimony of many practitioners who have engaged in interdisciplinary research (114).

Different professions incline toward different perceptions of the basic structure of reality—some tending to perceive it as more atomistic, static, dynamic, complex, discoverable, etc. than others. Increasing exposure to the unfamiliar viewpoint may increase one's estimate of its utility in efforts to approximate the elusive essence of reality. Here, however, the individual's basic world view is so vitally involved that resistance to exposure is especially durable and differences that seem strictly methodological may trigger strongly emotional reactions. Meehl has remarked that "It is customary to apply honorific adjectives to the method preferred, and to refer pejoratively to the other method. For instance, the statistical method is often called operational, communicable, verifiable, public, objective, reliable, behavioral, testable, rigorous, scientific, precise, careful, trustworthy, experimental, quantitative, down-to-earth, hardheaded, empirical, mathematical, and sound. Those who dislike the method consider it mechanical, atomistic, additive, cut and dried, artificial, unreal, arbitrary, incomplete, dead, pedantic, fractionated, trivial, forced, static, superficial, rigid, sterile, academic, over-simplified, pseudoscientific, and blind. The clinical method, on the other hand, is labeled by its proponents as dynamic, global, meaningful, holistic, subtle, sympathetic, configural, patterned, organized, rich, deep, genuine, sensitive, sophisticated, real, living, concrete, natural, true to life, and understanding. The critics of the clinical method are likely to view it as mystical, transcendental, metaphysical, supermundane, vague, hazy, subjective, unscientific, unreliable, crude, private, unverifiable, qualitative, primitive, prescientific, sloppy, uncontrolled, careless, verbalistic, intuitive, and muddleheaded. There are also some words (e.g.,
Characteristic differences between practitioners and researchers are paralleled by differences among researchers. It is significant that the comment just quoted concerned, not differences between practitioners and social scientists, but differences among social scientists. For the characteristic interdisciplinary divergences in viewpoint are paralleled (within a narrower range of variation) by differences among social scientists. Some are trained in and inclined toward strictly statistical methods. Others have more experience and more confidence in a clinical approach. Probably no researcher of reasonable maturity leans wholly on one or the other type of analysis. To be at either extreme of the methodological continuum from statistical to clinical suggests either lack of experience or considerable rigidity. However, most individuals have an inclination in one direction or the other. This inclination is likely to be reflected in the individual's selection of his field, since different branches of social science and different specialties within each branch differ in the degree of their commitment to the clinical or the statistical approach. Choice of field, however, reflects many other elements—psychological, social, or accidental—and social scientists under any label vary greatly in their methodological leanings (188, 268).

Interdisciplinary problems in a research team of behavioral scientists have been thoughtfully analyzed by Simmons and Davis (300). They point out that most behavioral scientists lean either toward a "clinical" or a "quantitative" approach to research, and that where an individual stands in this respect depends on his disciplinary orientation, his research experience, his knowledge of the materials under study, and his personal temperament. Recognizing that none of their team members was "purely" clinical or "purely" quantitative, they add the sound comment that no interdisciplinary research project can or should be purely one or the other. Nevertheless, though "our explication of procedures has helped allay fears of clinicians and quantifiers that each wanted to push out the other and made for increasing recognition that the two approaches are complementary and necessary for the kinds of research in which we are engaged . . . the differing points of view . . . persist . . . as barriers to communication and consensus which will have to be overcome as they arise in continuing attempts at collaboration." (300, p. 301)

Perhaps it will be necessary for each new interdisciplinary team member to learn from scratch such principles of collaboration as those outlined above. On the other hand, ways may still be devised for speeding up the process. One possibility is that seasoned and experienced members of the novice's own profession, individuals who command his respect and credence
and talk his language, might help him to accelerate his own seasoning. This would have to be done by actual work on some project, for the lesson taught by the many articles on the subject written so far seems to be that the principles of interdisciplinary collaboration are conveyed far less effectively by precept and abstract discussion than by work in a field situation. A teaching situation modeled on a field situation might be effective if the individual inexperienced in interdisciplinary research were working side by side with an accepted representative of his own persuasion who had mastered the art of fruitful interdisciplinary collaboration. The status and prestige of the senior member might help to eliminate the emotional blocks to perception which grow out of a novice's extreme veneration of his own gospel and his fear of betraying it to "lesser breeds without the Law." Such seasoning would be especially useful for the social scientist trained in statistical methods and for the practitioner with no research training.

The choice of research staff for interdisciplinary collaboration, then, requires attention to far more than technical training and competence. It requires attention also to the theoretical orientation of the individual and to the amount of experience he has had with the kinds of material involved in efforts to bring about psycho-social change in individuals. It requires above all assessment of the capacity for meeting new challenges with realism, ingenuity and freedom from doctrinaire rigidity. If the research director has all these qualifications in high degree, he can often work effectively with a team less experienced in interdisciplinary research, providing the team members do have full technical competence plus the "intellectual hospitality" that enables one to listen, to perceive, to communicate, and to modify previous positions in the light of new information.

As Simmons and Davis (300) also point out, all members of an interdisciplinary team must meet the large demands for patience imposed by the constant need to explain and even to defend what seems obvious, by the constant need for group discussions and decisions, the painstaking labor of coordinating and standardizing procedures, and the slow tempo of group, as compared with individual, activity.

CLAIMS AND EXPECTATIONS

One frequently cited aim of psychotherapy—and also of social casework—is to help individuals attain the "need-free perceptions" that are part of mental health (155). That is, to help them achieve a sturdy realism capable of perceiving, without distortion or evasion, the situations and problems that confront them.
Evaluative research of the kind under discussion here urgently requires need-free perceptions on the part of those who carry out research, those who request it, and those who use its results. Such research at times has been plagued by unrealistic expectations on the part of research consumers and also of research producers. As the magnitude and complexity of the problems become evident, these expectations often give way to a sense of let-down on one side and a considerable defensiveness on the other. One means toward realism in the research producer is familiarity with the material to be researched. One means toward realism in the research consumer is understanding of the research problems involved. It is necessary to recognize on the one hand the difficulty and distance of the ultimate evaluation goal, and on the other hand the richness of the rewards to be achieved in approaching it.

A healthy realism is required not only concerning research goals and potentials but also concerning the purposes of those who use research and those who produce it. The administrators and boards who requisition a study must be clear whether their primary objective is short- or long-term evaluation. If the primary purpose is to advance professional knowledge, then pre-evaluative research is the best investment. If the primary aim is administrative, then there may be sound reason for short-term evaluation—but the primacy of this aim should be recognized and avowed.

On the other hand, the producer of research needs a healthy realism concerning the nature and values of what for convenience has been dubbed "administrative research." Research designed to help administrators serve people better hardly deserves the frequent implication that it is inferior to other types of research, even though it may be less gratifying to the researcher.

There is need to be on guard against a number of confusions, including the confusion of realizable research values with the status values that have grown up around certain types of research, and the confusion of need for a certain type of research with the need to make "an attractive package" that will get financial support. Need-free perception does not demand ignoring any of these values, but it does require recognizing which is which.

It would seem, then, that a major objective in research, as in the treatments, services, and programs research is asked to evaluate, must be honest, enlightened and outspoken realism. The necessary basis for research realism is understanding of the materials to be investigated, the questions to be answered, the limitations to be recognized, and the rules of evidence to be respected in interpreting research results.
APPENDIX

"Two Out of Three Improve, With or Without Treatment"

Lack of adequate control information has weakened both sides in the controversy stirred by the publications of Eysenck (91), drawing heavily on material brought out by Denker (74, 194, 285). Eysenck notes that, according to the best figures available on the results of psychotherapy, about two out of three among the treated show improvement or cure and about two out of three among the untreated show spontaneous remission of symptoms. He points out that the figures thus "fail to support the hypothesis that psychotherapy facilitates recovery from neurotic disorder." He does not claim (as do some who quote him) that he has disproved the effectiveness of psychotherapy. However he does state that "even the much more modest conclusions that the figures fail to show any favorable effects of psychotherapy should give pause to those who would wish to give an important part in the training of clinical psychologists to a skill the existence and effectiveness of which is still unsupported by any scientifically acceptable evidence." (91, p. 323)

Few serious researchers would challenge Eysenck's statement that so far we have no solid statistical evidence proving the efficacy of psychotherapy. At the same time, many would challenge a claim that the figures Eysenck cites demonstrate identical results for treated and untreated individuals. As various commentators have noted, there is no evidence that all members of the presumably untreated sample really received no psychotherapy, and there is no evidence on the nature and degree of their problems or the level of their recovery, as compared with a sample of individuals who received psychiatric treatment (201, 280). There is no evidence, then, that the treated and untreated groups were comparable in type of problem, severity of problem, or degree of recovery. These are all points on which, as has been mentioned, comparability must be established if the results of one treatment method are to be compared with those of another, or of no treatment.

95
Questions about the comparability of Eysenck's "treated" and "untreated" groups are reinforced by findings produced with the "own-control" method used by the Rogers group. As pointed out in the section on controls (p. 62), these findings suggest that those who improve without treatment represent a group different from those who improve with treatment.

Doubts about Eysenck's control group are further sharpened by the fact that the "two-out-of-three" figure is more solidly established among the treated than among the untreated. In reports of therapy this figure is, as Hunt says, the most frequent—although as Levitt points out, the range is wider than is sometimes assumed (194). There are far fewer reports of the untreated, however, and those that exist are often perceptibly biased—for example, by being drawn from individuals who started and then discontinued treatment. Moreover, however questionable the samples of the treated, there is still no question that they were treated. In Denker's "untreated" group there is doubt—aside from all the other questions raised—whether in fact all of them lacked treatment. Thus, the two-out-of-three remission rate seems to have gained currency on grounds even more shaky than those that underlie the two-out-of-three treatment rate.

References

The list of references given below represents the publications that have been drawn on most directly for this paper. It does not include all the references used, nor was it possible even to review all the voluminous literature on evaluation of psychotherapy. The selection of studies and commentaries is related to specific points in the discussion and includes a considerable range of views and of methods. However, others not included may have equal or in some cases greater merit and significance.


97

Provided by the Maternal and Child Health Library, Georgetown University


25. Blenkner, Margaret: Obstacles to Evaluative Research in Casework. SOC. CASEWORK, 1950, 31, 54-60; 97-103 (February-March).


34. Brenman, Margaret, [Chairman], Kubie, Lawrence S., Rogers, Carl C., and Others: Research in Psychotherapy. Round Table, 1947. AMER. J. ORTHOPSYCHIAT. 1948, 18, 92-118 (January).


54. Cattell, Raymond B., and Luborsky, Lester B.: P-Technique Demonstrated as a New Clinical Method for Determining Personality and Sympt-


72. de Grazia, Sebastian: ERRORS OF
73. Denker, P. G.: The Prognosis of Insured Neurotics. NEW YORK STATE J. MED., 1937, 37, 238.


91. Eysenck, Hans J.: The Effects of Psy-


108. Gill, Merton M., and Brenman, Margaret: RESEARCH IN PSYCHOTHERAPY. [In] Brenman, Margaret [Chairman]: Research in Psychotherapy. AMER. J. ORTHO-


121. Grummon, Donald L., and John, Eve S.: CHANGES OVER CLIENT-CENTERED THERAPY EVALUATED ON PSYCHOANALYTICALLY BASED THEMATIC APPRECIATION TEST SCALES. [In] Rogers, Carl R., and Dymond, Rosalind F. [editors]: Psychotherapy and Personality Change. Chicago: Uni-


138. Hollingshead, August B., and Redlick, Frederick C.: Social Stratification...


170. Klein, George S., Holzman, Philip S., and Laskin, Diana: The Perception


204. Luborsky, Lester B., and Sargent, Helen D.: The Psychotherapy Research Project of the Menninger Foundation: Sample Use of Method. BULL.


221. Menninger, Karl: Psychological Aspects


253. Parloff, Morris B., Kelman, Herbert C., and Frank, Jerome D.: Comfort, Effectiveness, and Self-Awareness as Criteria of Improvement in Psycho-


266. Raskin, Nathaniel J.: An Objective Approach to the Study of Psychotherapy. AMER. SCIENTIST, 1949, 37, 410-413; 420.


269. Rennie, T. A. C.: Follow-Up Study of 500 Patients with Schizophrenia Admitted to Hospital from 1913 to
1923. ARCH. NEUROL. & PSYCHIAT., 1939, 42, 877–891 (November).


285. Saslow, George and Peters, Ann DeHuff: A Follow-Up Study of "Un-


288. Schmidl, Fritz: A Study of Techniques Used in Supportive Treatment. SOC. CASEWORK, 1951, 32, 413–419 (December).


309. Thomas, Dorothy V.: The Relationship Between Diagnostic Service and Short-Contact Cases. SOC. CASEWORK, 1951, 32, 74-81 (February).


335. Witmer, Helen and Students: The Later Social Adjustment of Problem Children: A Report of Thirteen


