Methodological Dilemmas in Research on Prevention and Intervention

Günter Albrecht

1. A General Introduction

The methodological problems involved in research on prevention and intervention as a specific object of research must be investigated and explained with a particular thoroughness and conscientiousness (in a completely literal sense), because, unlike pure academic research, it is directly related to practice, to the application of scientific knowledge. Research on prevention and intervention cannot rely – as is otherwise common practice in science – on its findings initially being subjected to controversial and critical discussion in the scientific community and frequently to replication and modification before a "practitioner" has the idea of taking science at its word and at least partially applying its findings. The researcher in the field of prevention and intervention is generally commissioned to provide scientifically tested statements on the success of the measures that an authority wishes to apply. Depending on the type of cooperation, the scientists are confronted with highly differing tasks and thus also methodological problems. If they think about the design of the prevention and/or intervention in detail, they enter into the general role problem of applied research as well as the problems involving the theory of science, sociology of science, and the ethics of science. This cannot be discussed in complete depth here. I shall focus on the methodological problems, which I consider to be found on four levels.

The *first* problem is how, in any case, to develop proposals or programs (particularly as far as their technological side is concerned) on the basis of the available theoretical and empirical knowledge, as the type and amount of available knowledge is typically in no way sufficient for the derivation of tested and concrete strategies for solving problems. This is particularly a methodological problem with a slant toward the theory of science.

The *second* methodological problem is to use methods that can ensure early statements on the utility or effectiveness of any measure of prevention and intervention that is developed, particularly as such measures frequently generate

major societal costs. In a society of "homo faber," in which it is not just the social sciences that have intentionally or unintentionally conveyed the impression that anything can be done or produced, it is easy to suggest that societal processes can also be manipulated or at least influenced in an unlimited manner so long as one proceeds scientifically. Without wishing to evaluate this state of affairs, it has to be assumed in any case that such an understanding of "policy as experiment" requires a high level of control over its activity (see, in particular, Riecken and Boruch, 1974; also, for a discussion of the methodological aspect of the problem, the early and central article from Campbell, 1969). The possible conclusion is that when policy or reform is already carried out as an experiment for general political, economic, or impending scientific reasons, then the experiments must be designed in such a way that they (as in the experimental sciences) provide tested information on "reality" – insofar as this information is retained in the experiment. In other words, the various social sciences are required to involve their methodological competences in such a way that the implementation of the experiments conceived under their guidance can be followed by stating more or less without doubt whether these experiments do or do not support the continuation and/or generalization of the corresponding measures.

The *third* methodological problem concerns research on prevention and intervention in a much more specific sense, particularly the problems of prevention, but, on a closer inspection, generally also those of intervention. Broadly speaking, prevention can be understood as measures to prevent or impede future impairment or injury to persons. Somewhat more specifically, it is, of course, above all aimed at groups of persons who face a *particular* risk of suffering the corresponding impairment or injury if no prevention is performed. However, this results in a serious dilemma: On the one side, it is assumed to be true that the earliest application of measures of prevention increases the chance of avoiding impairment or injury (naturally provided that these measures of prevention are or can in any way be positive). On the other side, because of the difficult problems of early diagnosis, such an early application of prevention can have the side effect that it extends to a major circle of persons who not only do not require this prevention but are even injured or at least inhibited in their "normal" potential for development by it. Thus, we are confronted with the special problems of early diagnosis and the prediction of developmental courses. As this problem is very closely linked to specific theories and their power, it should not involve methodology in a strict sense. However, it cannot be overlooked that serious methodological problems also evolve from this for the performance of studies on the problems of prevention and intervention, as the unequivocal assignment to experimental and control groups naturally re-

quires that the subjects' level of risk can be classified precisely. This makes it clear that this question concerns an extremely explosive combination of theoretical, methodological, and also ethical problems that require further study.

The *fourth* methodological problem is to arrive at an exact evaluation of "social experiments" by following them scientifically. As, ideally, these experiments can be understood as a methodologically controlled investigation within the framework of rational policy development, it would be of central importance to perform a precise evaluation of the experiments concerned on a great number of dimensions. This is the only way to undertake a scientifically based program and organization development and, in particular, to answer the question whether the comprehensive expansion of such generally limited experiments (or model plans) can be recommended. Although this presentation of tasks may first appear to be trivial, when viewed more closely, it proves to be very difficult to arrive at a consensus on how to cope suitably with this goal.

Recent decades have seen the development of a scientific discipline that has addressed the solution of these problems: evaluation research.

2. *Developmental Trends in the Methodology of Evaluation Research*

Although evaluation research only has a short history, it is extremely difficult to overview it and thus scarcely possible to draw a balance on the current state of research, even when this is not directed toward the *results* of evaluation research but toward its *methods*. On the one hand, evaluation research has first appropriated the methodological canon of empirical social research and experimental psychology; on the other hand, it quickly became clear that the specific tasks of evaluation research required the development of an autonomous methodology that is still developing continuously and adopting new perspectives for each specific subdomain of evaluation research. Yet, evaluation research is far from having achieved a clearly demarcated discipline:

"Evaluation research is an applied, largely (and unfortunately) atheoretic, multidisciplinary activity spanning the social sciences and including education, health, and social work as well. While it shares with all these fields the common goal of assessing or evaluating innovative social programs aimed at improving human welfare, one might despair at discerning any other common threats that unite these disparate activities and weave them into a shared destiny." (Wortman, 1983, p. 224).

This heterogeneous field – as Wortman correctly points out in my opinion – is partially structured by society's need for applicable evaluation methods but:

"This does not mean that there is peace and calm in the choice of appropriate methods. To the contrary, there have been religious crusades, minor vendettas, and numerous scholary bloodbaths that one would expect when different disciplines converge on an activity." (Wortman, 1983, p. 224).

After such strong words, we cannot anticipate that a glance at the published literature will provide an overview of the current state of this methodological discussion. We have to expect uncleared, permanently shifting fronts. Therefore the following will be restricted to the more general problems.

An overview is made more difficult for at least two reasons: First, the rapid growth in government interest in social experiments has also led to an immense expansion in evaluation programs. For example, Hellstern and Wollmann (1983, p. 11) have calculated 2,679 experimental programs (that is, containing evaluation) receiving government support for the period 1970 to 1981 in the Federal Republic of Germany, and they assume that, each year, at least 200 to 300 programs receiving government support are of an experimental nature (for information on the generally enormous sums of money that have to be raised, see Hellstern and Wollmann, 1983, pp. 13–16). If we move from the Federal Republic of Germany to the United States of America, we must anticipate substantially larger dimensions, as evaluations are required by law in a great number of governmental programs. For this reason, the number of evaluations per year in certain fields (e.g., education, occupational training, and also criminal justice) is already well into the thousands. It is apparent why the state of this research cannot be documented. There is also an additional problem: The majority of evaluation programs are not published on completion but only circulate in small numbers among the funding authorities. This makes it hard to obtain access to this material.

This is linked to a third variable that plays a central role: Evaluation research continuously faces the threat of being closely linked to rival political interests as – at least theoretically – its findings can be of great political significance if they are able to decide on the weal and woe of comprehensive programs involving a great number of persons (see, for a discussion on the threat of politics to evaluation research, Weiss, 1970). Naturally, this circumstance does not lead to evaluation research reports being particularly accessible.

In view of this situation, it is no surprise that we know relatively little about the state of evaluation research in the Federal Republic of Germany. The pre-

viously available trend reports either only cover a very small section of the research in each field (e.g., Blass-Wilhelms, 1983a, 1983b, on social therapy in prisons) or they are restricted to reports on approximate numbers (e.g., Lange, 1983; Hellstern and Wollmann, 1983), provide overviews on the topics treated, their changes over the course of time, and so forth, but exclude any systematic analysis of the programs from methodological perspectives. For example, Hellstern and Wollmann's (1983) very comprehensive work with examples covering a great number of political fields is indicative in that it hardly reveals any inclusion of *methodological* issues while the more political or political-science side of the individual programs is discussed in detail. However, this finding can be generalized to the entire German discussion in sociology during the first half of the 1980s, so that I consider Lange (1983, p. 259; translated) to be completely correct when he states: "On the other hand, it is surprising that the methodological reflection on this type of applied social research is almost completely neglected in the German sociological journals and monographs."

This was not so strictly true for psychology in the German-speaking countries, but here as well, a more methodologically oriented anthology containing mostly translations from the Anglo-American countries (Wulf, 1972) was initially only followed by isolated research reports and only very much later by methodologically oriented publications (e.g., Biefang, 1980; Wittmann, 1985; Lösel and Nowack, 1987).

This statement does not hold for the international discussion. Independent of the evaluation research actually *practiced*, this is now revealing a consolidating discussion on methods. This is partially documented in the classical works of Caro (1971), Rossi and Williams (1972), Weiss (1974), Guttentag and Struening (1975), Abt (1977), Rossi, Freeman, and Wright (1979), and Rossi and Freeman (1989); even when marked differences in the understanding of evaluation and the accompanying opinions on what constitutes an appropriate methodology can be recognized.

3. *Evaluation Research and Its Various Conceptions*

With reference to Glass and Ellett (1980), it is possible to differentiate seven alternate conceptions of evaluation: (1) evaluation as applied science; (2) evaluation as systems management; (3) evaluation as decision-making theory; (4) evaluation as the measurement of developmental progress; (5) evaluation as

legislation; (6) evaluation as description; and (7) evaluation as rational empiricism.

Each of these conceptions, which will not be described in much detail here, has its advantages and disadvantages as well as its logical and methodological strengths and weaknesses. *Evaluation as applied science* ultimately centers on the evaluator as experimenter searching for the causes of the effects produced. The issue of the application or even applicability of the knowledge obtained appears to become secondary. The conception of *evaluation as systems management* has the advantage of being able to view the systemic embedment of activity, but, according to its critics, the embedment of the evaluator in bureaucratic procedures generally only leads to a selective stabilization of a system, while its rationality remains unquestioned.

In the conception of *evaluation as decision-making theory*, decisions are conceived as quasi-logically linked to the evaluation. This proves to be untenable empirically, because empirically obtained evaluations – despite their explicitness – are frequently not taken into account when decisions are made. *Evaluation as the measurement of developmental progress* or of the proximity to the goal – a very frequent conception historically, particularly in educational policy – overlooks the fact that goals also have to be evaluated, particularly in comparison to rival goals. If we look at *evaluation as a legal conflict*, in which the adversaries can present their arguments, one ignores the fact that it is generally possible to obtain *unequivocal* decisions in court, which is not the case for evaluation. *Evaluation as description* initially appears to be self-contradictory. Evaluation is to be seen here as a detailed and exact description of a program, its actions and steps, its impact on the clients, and so forth. The methods to be applied here are not systematic experiments but ethnographies. This conception assumes that the results of such an evaluation can be replicated. However, it overlooks the fact that generalizability can be so hard to achieve.

Evaluation as "rational empiricism" – closely linked to the work of Scriven (see, in particular, 1974a, 1974b) – is understood as the application of a multitude of methods (e.g., comparative experiments, need analyses, goal analyses, complex measurement of the findings). The selection of the necessary combination of methods must be completely specific to the issue under investigation without it being possible to find some kind of simple formula. The measurement of the rationality of the program is decisive in the evaluation, whether this involves the costs or the effects, the individual actions, their moral quality, or their necessity and so forth. Values and demands are considered to be testable and justifiable, as ethical principles are generally viewed as being

rationally justifiable – a conception that understandably gives rise to skepticism.

Let us turn away from the clear differences revealed in the perspectives from which evaluation is viewed in these different directions and consider on a more technical level the issues to which evaluation research should provide concrete answers.

Here there appears to be a degree of consensus across the different directions that evaluation research essentially covers *four* fields of questions. However, the single directions differ in the importance they assign to each field. As I am particularly interested in the methodology, I shall not discuss these evaluations, and I shall only briefly sketch the main complexes of issues and consider them from the perspective of research methods.

The *first* complex is related to the planning of programs. It particularly concerns the information on the breadth and characteristics of the target population and the analysis of whether the developed program meets the intended goals and whether the chances of successful implementation are exploited as fully as possible.

The *second* complex revolves about the implementation of the program; that is, it concerns the control of the program realization. This involves whether the program actually makes available the resources, services, and other goods that are intended (program strategy analysis or process evaluation).

The *third* complex asks whether the program is successful in its attempt to achieve its goals or whether the observable effects can be traced back to other processes and causes that have nothing to do with the program itself. Of great importance is the further question whether the program produces unintended side effects (effect analysis).

The *final* complex concerns the economic efficiency of the program. This studies the costs of the services, goods, and measures and the effects obtained with particular emphasis on contrasting the achievements of the program with alternate uses of the resources applied (see, in particular, Rossi, Freeman, and Wright, 1979, pp. 32–45, 241–282; Rossi and Freeman, 1989, pp. 375–415; for cost analysis, Phillips, 1980).

This short list, which must suffice here, makes it clear that depending on the way in which evaluation is accentuated (which should take all aspects into

account in the ideal case), highly differing aspects are treated and thus highly differing methodological instruments are required. If the first complex above all presents theoretical and theory-of-science demands, the second complex requires very precise longitudinal analyses on the course of programs and actions, interactions, everyday routines, and so forth, in which it is very clear that not only the appliers but also the clients of measures of prevention and intervention have to be taken into account. Data should be assessed as continuously as possible over the entire duration of the program. To meet the demands of effect analysis, it is necessary not only to measure the effects as exactly as possible and to be able to assign these to causes as certainly as possible but also, in particular, to explain whether effects are due to the program or not. For the last aspect, cost-benefit analyses in economics naturally provide a model. However, they will not receive much attention in this article.

The methodological requirements therefore differ greatly depending on the type of evaluation or the accentuation of one of the four subdomains of evaluation. The major lines of conflict are particularly found between those researchers who emphasize the relevance of process evaluation and those who concentrate more on effect research. It can clearly be seen that both priorities relate to political, theory-of-science, and "theory-political" preferences within the different groups of researchers, and it is well known that this does not facilitate a rational and open discussion.

Those researchers who have regarded effect analysis to be their particular task considered themselves to be assigned to the theory-of-science position of critical rationalism in order to solve their problem – empirical testing of the theoretically claimed relation between specific measures/stimuli (independent variables) and effects (dependent variables) and the exclusion of rival explanations. Therefore they particularly adopted the methodological model of the experiment as their ideal. Those researchers who paid less attention to the outcome of intervention but were more interested in the process of the program adopted the paradigm of action research (Gruschka, 1976). This resulted in a shift of the goal of evaluation from obtaining knowledge in the sense of the falsification of theories to the development of action alternatives for solving problems as they occur. It was accompanied by a change in the position of the evaluator: The separation of the subjects and objects of evaluation was abolished (or should have been abolished); values were demanded instead of the value neutrality of statements. Decisive quality criteria of evaluation were no longer validity and objectivity but communication, transparency, and relevance

(Gruschka, 1976, pp. 142–151; see also House, 1980, pp. 249–257). This countercurrent was very powerful and in the Federal Republic of Germany, for example, it led to more than one half of the evaluations of educational models being based on these methods (see Lange, 1983, p. 256).

However, this development did not just occur in the Federal Republic of Germany but followed a very similar course in the Anglo-American countries. This generally led to very fixed fronts between which hardly any communication was possible. The supporters of the two main directions worked out long lists of the deficits of the opposing positions (see, e.g., the lists compiled by Smith, 1981) – and both were justified in doing so. This alone raises the suspicion that the discussions were based on false premises or that the authors had phrased their questions incorrectly.

I shall therefore refrain from presenting the arguments in detail here and instead simply state that it is not only necessary but also possible to use the different alternatives in a creative way and, by linking them together, arrive at the application of a methodological principle that has long been emphasized in social research: triangulation through multimethod approaches.

I consider this statement to be justified because it can be demonstrated that evaluation research, particularly when it views itself as impact research, is fundamentally confronted with certain dilemmas that can only be overcome by developing its own heuristic. Therefore in the following I shall consider the problems of applying the strictly experimental paradigm in impact research.

4. *The Experimental Paradigm and Its Limitations*

An Overview of the "Classical" Experiment

I shall initially consider the measurement of effect and the particularly precarious issue of the causal assignment of effect to program or treatment. The methodology of the social sciences generally recommends – completely in line with the experimental disciplines of the natural sciences – the use of purely experimental research designs: In the ideal case, several experimental groups and control groups are used in such a way that they differ not only with regard to the application of the treatment but also in that they are exposed to different conditions through the temporal distribution of the measurements. The goal

is to ensure that intervention effects, measurement effects, maturation effects, and their interactions are clearly discernable (see, for a discussion of many other effects as well, Campbell and Stanley, 1963).

As good as this may sound, this procedure nonetheless has clear limitations. If, for example, we are concerned with the assessment of the long-term effects of the intervention, the need to reduce measurement effects would require the organization of a whole range of experimental and control groups. Even in posttreatment measurements with two measurement intervals, which certainly would not represent the ideal assessment of long-term consequences, this would result in four experimental groups and four control groups (see Figure 1).

	t_1	t_2	t_3	t_4	
R	M_1	X	M_2		Experimental group 1 (Group 1)
R	M_3		M_4		Control group 1 (Group 2)
R		X	M_5		Experimental group 2 (Group 3)
R			M_6		Control group 2 (Group 4)
R	M_7	X		M_8	Experimental group 3 (Group 5)
R	M_9			M_{10}	Control group 3 (Group 6)
R		X		M_{11}	Experimental group 4 (Group 7)
R				M_{12}	Control group 4 (Group 8)

Note:

R = randomization, chance selection; $M_1 - M_{12}$ = Measurement 1 to Measurement 12; X = intervention, treatment, stimulus, etc.; $t_1 - t_4$ = Timepoints 1 to 4.

Fig. 1: A "classical" experimental design with additional controls for sample and time effects.

When it is considered that each of these groups can only demonstrate statistical certainty if they all contain sufficiently large samples, and that, furthermore, certain initial data must be ascertained because of the mortality that can be anticipated over the course of time, then the problem becomes very clear. In addition, the treatment is regarded as being very short-term here, or it is dealt

with as if it were very short, like, for example, a medical intervention. In reality, most measures of intervention and prevention in contrast need to stretch over longer periods of time, that is, they represent a longer process in themselves that cannot be assessed by pre- and posttest; not only regarding effects but also particularly regarding the ways in which they have effects, as, for example, interaction effects between various elements or phases of the treatment only enter the measurement as a net effect and so forth. Even the control of both parts of a two-phase treatment requires an effort of research that is hard to achieve (see Figure 2 below).

It also has to be taken into account that in many cases in which appropriate experiments should be applied to questions of intervention and prevention, there is not a sufficiently large number of subjects for whom the framing conditions (e.g., the institutional setting) are even approximately equal. This particularly raises problems in the control of the treatment, for as soon as the experiment is dependent upon the cooperation of several institutions that are performing measures of intervention or prevention, the integrity of the ordained treatment is hard to maintain – and if this is not achieved, even the most elaborate design can produce no really useful results (see, for the significance of this factor, Sechrest and Redner, 1979).

However, the experiment creates other fundamental problems as well. These particularly concern the transferability or also the generalizability of the findings, that is, the external validity (see, for a general discussion, Campbell and Stanley, 1963; Cook and Campbell, 1979). An experiment that is designed as a laboratory experiment, in which the experimental design runs through the most constant framing conditions possible and the effects, changes, and irregularities that otherwise occur in everyday reality drop away, occurs under *special conditions* that are not representative for reality. The decisive question is therefore whether one can assume that the effects that are possibly determined under experimental conditions will also occur under nonexperimental conditions. This question is naturally not to be answered ex cathedra in this form but it is an empirical question that has to be investigated in each case; it is *the* question per se for the relevance of the experiment (see, for a discussion of the problems involved in generalizing conclusions from experiments to realistic conditions of application, the case studies from Pillemer and Light, 1981, in which great skepticism is advised).

	t_1	t_2	t_3	t_4	t_5	t_6	
R	M_1	X_a	M_2	X_b	M_3		Exp. 1
R	M_4	X_a	M_5	X_b		M_6	Exp. 2
R	M_7	X_a		X_b	M_8		Exp. 3
R	M_9	X_a		X_b		M_{10}	Exp. 4
R	M_{11}	X_a	M_{12}		M_{13}		Exp. 5
R	M_{14}	X_a	M_{15}			M_{16}	Exp. 6
R	M_{17}	X_a		M_{18}			Exp. 7
R	M_{19}	X_a				M_{20}	Exp. 8
R	M_{21}		M_{22}	X_b	M_{23}		Exp. 9
R	M_{24}		M_{25}	X_b		M_{26}	Exp. 10
R	M_{27}			X_b	M_{28}		Exp. 11
R	M_{29}			X_b		M_{30}	Exp. 12
R	M_{31}	X_b	M_{32}	X_a	M_{33}		Exp. 13
R	M_{34}	X_b	M_{35}	X_a		M_{36}	Exp. 14
R	M_{37}	X_b		X_a	M_{38}		Exp. 15
R	M_{39}	X_b		X_a		M_{40}	Exp. 16
R	M_{41}	X_b	M_{42}		M_{43}		Exp. 17
R	M_{44}	X_b	M_{45}			M_{46}	Exp. 18
R	M_{47}	X_b		M_{48}			Exp. 19
R	M_{49}	X_b				M_{50}	Exp. 20
R	M_{51}		M_{52}		M_{53}		C 1
R	M_{54}		M_{55}			M_{56}	C 2
R	M_{57}				M_{58}		C 3
R	M_{59}					M_{60}	C 4

Note

R = randomization; $t_1 - t_6$ = timepoints; X_a = Treatment a; X_b = Treatment b; $M_1 - M_{60}$ = measurements; Exp = experimental group; C = control groups.

Fig. 2: Experimental design with a two-stage treatment and two posttreatment measurements

The Field Experiment

One way to overcome this problem of endangering the external validity by raising the internal validity through the application of the experiment is to use the field experiment. This offers the opportunity of measuring not only the changes in the target variables (i.e., the effects) but also the situational, organizational, and administrative, as well as sociopolitical conditions that the researcher assumes to be relevant. Thus, this does not involve the exclusion or neutralization of all possible variables that do not belong to the treatment but the additional consideration of the greatest possible number of variables that could influence the effect of a measure being tested as well as their interactions and side effects. This is the only way in which it can be explained at the end of the study whether and to what extent the observed effects are due to the treatment variable or to its framing conditions and third variables or the interactions between them. However, this research design also reveals serious problems, because the list of explanatory factors included in the assessment on the basis of theoretical considerations and empirical knowledge is in principle incomplete, as it always has to be assumed that events may occur during the duration of the study, which has to be relatively long particularly in long-term effect analyses, that, though unpredictable, may have had a major impact on the development of the field experiment (see, for an example of a very careful and creative solution to such problems, Dennis, 1990). The frequently recommended use of a daily research journal to register these events is certainly no more than an *emergency* solution, for, on the one hand, it does not provide an appropriate basis for quantitative analyses while, on the other hand, it is generally only possible for the evaluation researcher to register conspicuous changes in the environment of the experiment, while the many smaller but, in total, conceivably very relevant influences on the everyday program, particularly when they concern the "inner life" of the experiment, cannot be assessed in this way.

Naturally, the criticisms directed toward the research design of the laboratory experiment regarding the complex organization of experimental and control groups and the resultant measurement problem also apply here. However, a further complication is that the effects of the external conditions, for example, of specific events, could be, or appear to be, very different for the various experimental and control groups, because they each indicate a completely different proximity to the measurement timepoints. Therefore, it would not be possible to differentiate with a high level of certainty between treatment effects, external irrelevancies, measurement effects, maturation effects, and their interactions.

The Problem of Achieving Random Assignment

To complete the measure of problems, a further fundamental difficulty for experimental designs is revealed: A central prerequisite of experimental designs is that the subjects are assigned to the experimental and control groups in strict randomness, as this is the only way to avoid mistaking selection effects for intervention effects in the study. This randomization in the construction of experimental and control groups often raises very great problems (see also Dennis and Boruch, 1989; Dennis, 1990, p. 350). The difficulties extend from moral-ethical and legal ones to simple technical difficulties that can frequently never, sometimes approximately, and in infrequent cases completely be overcome. In my opinion, the central ethical issues, which can only be mentioned here, involve the problem of whether it is permissible to deny persons a more effective treatment or one that is expected to be more effective and assign them, for example, to a control group, although they may possibly require this treatment more urgently because of their problems than other persons who, however, are randomly assigned to the experimental group with treatment. The frequent argumentation in the field that no truly reliable statements can be made about the anticipated treatment effects so that this moral issue does not even arise in this form only leads us deeper into the fundamental moral and ethical issues, for if we really know so little about the effects that can be anticipated, then experiments of this kind should possibly not even be carried out and so forth.

Without presenting the reasons in detail, we can nonetheless assume that the randomization of assignments to experimental and control groups is a very difficult problem to solve in all cases involving decisions that could be highly relevant to the development of persons. This raises so many aspects that I prefer to close with this preliminary balance.

The Quasi-Experimental Research Design

The situation described above has led methodologists to try for approximately two decades to find solutions to those cases that permit neither a laboratory nor a field design. A series of *quasi-experimental* research designs have been developed that offer more or less useful emergency solutions depending on the initial conditions – in particular, the quality of data, the possibilities of forming control groups, and the type of data (e.g., cross-sectional or longitudinal). Because of the great number of designs that have been developed, I

can only mention a few here briefly and sketch their strengths and weaknesses, as my primary concern is only with naming the methodological problems.

Campbell and Stanley (1963, pp. 34–64) have observed that in a great number of cases in which the researcher is unable to control completely when a certain stimulus or treatment is exposed to whom and is also unable to achieve a strict randomization of the distribution of study populations to experimental and control groups, he or she has certain possibilities of drawing (careful) conclusions from the data obtained as long as certain prerequisites are met. Some typical designs can be mentioned that are available to research as quasi-experiments.

One example is the time-series design:

$$M_1\ M_2\ M_3\ X\ M_4\ M_5\ M_6$$

in which, in the ideal case, "measurement effects," "maturation effects," "regression effects," and "mortality effects" can be tested although external historical influences can only be calculated with difficulty. Naturally, the power of the design is improved when multiple time series are present:

$$M_1\ M_2\ M_3\ X\ M_4\ M_5\ M_6$$

$$M_1\ M_2\ M_3\quad M_4\ M_5\ M_6$$

that only differ according to the presence or absence of a treatment. Further possibilities of the time-series design are opened up when designs with equivalent time samples or with equivalent materials are available.

If the following simple designs are compared with the experiment, it is revealed that although they are unequally simpler and thus probably more frequently to be found in reality, they are also unequally less powerful for arriving at conclusions.

Mention should particularly be made of the pretest-posttest design with nonequivalent groups:

$$\begin{array}{llll} R & M & (X) & \\ R & & X & M; \end{array}$$

and the pretest-posttest control group design with nonequivalent groups:

$$
\begin{array}{cccccc}
R & M & X & & \\
R & & & X & M \\
\multicolumn{5}{c}{- - - - - - - - -} \\
R & M & & & \\
R & & & & M
\end{array}
$$

In the last design, two "control groups" are added that are subjected to no treatment, so that the interaction between treatment and pretest can be controlled. However, it can be observed that, in each case, randomization is only achieved *within* the two upper and within the two lower groups, so that, in some circumstances, further control groups need to be introduced in order to avoid the possible error of assigning a special property of the control group to the experimental stimulus.

My very brief and extremely shortened presentation of a few designs in quasi-experimental research (see Cook and Campbell, 1979) should not suggest that I do not consider them to be very relevant. In contrast, I assume that the further development of quasi-experimental research in combination with other methods will bring about a decisive advance in empirical social research.

As the "orthodoxy" of evaluation research generally far more strongly prefers *experimental* research because of its undeniable advantages, and it judges research on the extent to which it satisfies the quality criteria that apply for these methods, in the following, I shall test in somewhat more detail to what extent evaluation research within the framework of research on prevention and intervention is in any way able to meet these criteria.

5. *The Contradictory Methodological Demands Placed on Evaluation Research in the Sense of the Validity Criteria of Empirical Social Research*

If a first look at the experimental design reveals its fundamental problems, the problem is increased in a detailed investigation of applicability.

The central methodological problem of program evaluation and effect research is to ascertain as far as possible through the design of the study whether the research findings can meet the criteria of (1) *statistical validity*, (2) *internal validity*, (3) *construct validity*, and (4) *external validity*. The comprehensive

catalogue of threats to the different types of validity (see, e.g., Cook and Campbell, 1979) does not just reveal how strongly the validity of research findings can be impaired by insufficient methodological precautions. A somewhat more detailed inspection can also show that an optimization of the research design in order to protect it from one source that threatens validity can quite enormously raise the risk of other impairments to validity.

Threats to Statistical Validity

From the catalogue of seven decisive sources of threats to statistical validity (low statistical power, violated assumptions of statistical tests, fishing for significance and the error rate problem in multiple comparisons, reliability of measures, reliability of treatment implementation, random irrelevancies in the experimental setting, and random heterogeneity among the subjects), some should be particularly dangerous for evaluation research. I shall sketch these briefly here. I shall leave aside the factor "low statistical power," as I shall consider this elsewhere. First of all, the problem of "fishing" and the error rate in multiple comparisons appears to be explosive. The problem is that a relation is incorrectly assumed when multiple comparisons of means are possible and no knowledge on random significance is given. As precisely in comparably atheoretical research – which evaluation research frequently represents – the largest possible catalogues of conceivable effect dimensions are compiled and corresponding data are collected, it is not improbable that among the many comparisons a few significant relationships will occur by chance. The problem of the reliability of the treatment implementation is also of central importance as it can vary from person to person, from setting to setting, and from timepoint to timepoint even when the program calls for strict constancy. Naturally, this particularly applies to field experiments with a longer duration and different settings (see Boruch and Gomez, 1977). Finally, random irrelevancies in the experimental setting and random heterogeneity among the subjects are sources of error that are hard to control because the former would require either a strict laboratory experiment design or a field experiment design with systematic assessment of error (a very unrealistic assumption). In the latter case, the only opportunity leading to better solutions would be the selection of homogeneous groups that nonetheless would have damaging effects on the external validity in turn; or an additional data collection carried out during the study phase on the relations between characteristics of the population and dependent variables that nevertheless would only really be possible when one already knew the findings.

Threats to Internal Validity

The *internal validity*, the fundamental prerequisite for any meaningful application of the results of a study, is potentially threatened by a series of errors that are very difficult to control. This is particularly true when the studies indicate very specific features precisely because of the particular requirements placed on the evaluation research. In the evaluation of measures of prevention and intervention, one of the validity criteria in the evaluation of the program has to be seen as the program not just achieving short-term and quickly extinguished effects but at least confirming middle-range, and as far as possible, also long-range positive effects. Therefore, evaluation is required to assess the effect not only immediately after the measure but also in apparently meaningful temporal intervals that follow. In some circumstances, a long duration of the study has to be anticipated solely for this reason. Furthermore, in some circumstances, the measure of intervention or prevention itself is conceived of as taking a long period of time. This circumstance also results in the entire process of evaluation, in particular that of measurement, having to extend over a longer period of time. If we examine a few examples from the comprehensive catalogue of threats to internal validity, the problems lurking here become clear.

1. First of all, there is *history*. A great number of events can occur between the pretest and the last posttest, particularly if the study lasts a long time. These may have nothing to do with the treatment but have more or less strong effects on the dependent variables. A quantitative control of these sources of error is very hard to achieve (see, however, the appropriate attempts in Dennis, 1990, pp. 354–364), though nonetheless it could be achieved in a substitutive way through other types of study (e.g., qualitative, more ethnographic procedures).
2. Closely related to the above-mentioned problem is the threat of *maturation*. Between the measurements, particularly when the intervals are relatively large, changes could occur that have nothing to do with the treatment but have to do with the maturation of the subjects. This can be very important, particularly with subjects in certain age groups (e.g., children and adolescents) because comparatively large developmental steps could be made in relatively short periods of time during the study interval.
3. In the attempt to assess empirically the developmental processes of subjects who are exposed to certain measures, it will rarely be possible to avoid the repeated use of the same measurement instruments. This leads to repeated measurement effects that can trigger not only satiation in the subjects but

also changes in response behavior that are hard to estimate (see, e.g., Krauth, 1981; Petermann, 1978; Zielke, 1982). A dilemma also becomes clear here: It is certain that pure pretest-posttest designs are unsuitable for an appropriate measurement of the courses of intervention (see Bastine, 1970; Hartig, 1975), but by attempting to overcome this deficit, one once again faces the threat of making another methodological error.

4. I have repeatedly pointed out the selection effect and its very decisive threat to the valid evaluation of programs, particularly with reference to the assignment of changes in the dependent variables to treatments. However, it also threatens to slip in through the back door in a truly experimental approach with an originally random assignment of subjects, and this in the form of the varying mortality rates, as, under certain circumstances, they can completely destroy an originally successful random assignment in studies and treatments with a long duration. This is particularly a deadly threat to validity when the reasons for mortality are closely related to independent and/or dependent variables.

5. In summary, it should be pointed out that interaction effects can occur between several of the above-mentioned threats to validity (e.g., interactions between selection and other threats such as maturation, history, and instrumentation). Indications for these can scarcely be obtained through standardized assessments, so that this requires the application of a completely different type of study in order to experience the corresponding sensitivity and to search purposively in the quantitative evaluation for possibilities of testing the suspicion that corresponding impairments to validity have occurred and, if necessary, being particularly careful in the selection of the interpretations in this sense.

6. The plausibility of the assignment of specific observed effects in the dependent variables to specific treatments naturally requires a clear explanation of the direction of causal relations. This question cannot always be answered in a purely logical manner, and even then when the study design appears to indicate clearly that the characteristics of one variable changed previously in time to those of another, this does not exclude the occurrence of repercussions of the variables that are initially seen as dependent on the independent variables and so forth. If the statistical models used to analyze the data do not take such possibilities into account, the study can arrive at completely untenable conclusions.

7. Repeatedly mentioned in the literature are the problems of diffusion or imitation of a treatment or program, of compensatory equalization, compensatory rivalry by subjects receiving less desirable treatments, and the loss of motivation in members of control groups. These factors may also become

more virulent over the course of a long program because information on treatment differences is more widely known and so forth. This is a further example showing that the attempt to increase validity can lead to new problems of validity. In these cases, it also holds that the corresponding threats to validity can only be recognized suitably and thus partially controlled when data are available that refer to an intensive knowledge of the program and the "inner life" of the organizations in which processes of intervention and prevention take place.

Threats to Construct Validity

From the 10 threats to construct validity listed in Cook and Campbell (1979), I consider some to be particularly explosive in view of the properties of evaluation research. This particularly applies to the inadequate preoperational explication of the constructs. If the relevant constructs are not explained unequivocally, the study cannot be valid because partial assessments and even nonassessments of the central constructs must be feared, as operationalizations without an explanation of the theoretical constructs naturally cannot succeed. As already mentioned elsewhere, it is particularly evaluation research that frequently suffers from time pressure, as the funder wishes to ascertain information as rapidly as possible. Evaluation research additionally has the central problem that funders, program formulators, and program appliers frequently make unclear or even contradictory statements about their goals, are unable to obtain a consensus on the priorities of goals, and so forth. The same generally applies to the formulation of treatments. Here as well, the researcher is frequently at pains to undertake the appropriate explanations; not least because practitioners sometimes prefer to choose vague statements in order to avoid systematic controls of their work.

The sources of error from the application of only one operationalization and only one method are well-known; their dubiousness has already been strongly emphasized in other contexts above. The same applies to the subjects' suspicions regarding the experimental conditions (hypothesis guessing), the subjects' needs for social desirability (evaluation apprehension), and the expectations of the experimenter, so that I do not wish to deal with them further here. More critical, while frequently overlooked, is in my opinion the threat to construct validity of neglecting the level of the constructs. This has the effect that the researcher confounds constructs in general with specific levels of the same, thus, for example, out of the circumstance that an independent variable A on the level A_1 has no impact on the dependent variable B, it is concluded that A

generally has no effect on B, that is, also not on the higher level of expression A_2, and so forth. As we know that there are threshold effects and also that curvilinear relations are not so infrequent, it should be ascertained that all possible levels of the independent variables are present in the study. Nonetheless, the criticism also applies here that ethical and legal arguments may well speak decisively against these – methodologically meaningful – requirements.

Impairments of construct validity through interactions between various treatments and through interactions between testing and treatment should also play a significant role if the programs to be evaluated extend over longer time intervals and, for example, it is not ascertained in a field experiment that the subjects do not become the object of other interventions independent from the measures proposed in the program, or when the measurement of the treatment elicits reactions in the subjects that would not have occurred through the treatment alone – and could be mistakenly traced back to the treatment.

One very central problem remains to be mentioned that appears to be particularly important as far as the relation of evaluation research to practice and to science is concerned. Frequently, the design of the study does not sufficiently ascertain whether a generalization can be made over constructs. Through the researcher's too limited orientation toward a small number of constructs that are held to be important, on which data are collected, the relation to neighboring or rival constructs remains unclear as no operationalizations are made or data collected in order to explain this. This leads to a neglect of opportunities for generalization that could be significant for practical action and also for the formulation of theory.

Threats to External Validity

Ascertaining the external validity is absolutely indispensible for experiments or studies that are expressly performed to test the effects of measures on a small scale to determine whether they should be extended to larger populations. Though a purely scientific study still remains meaningful when its findings cannot be transferred to other populations or when this is uncertain, as the knowledge about the study population is nonetheless improved, an evaluation study can be completely without value if the generalizability, that is, the external validity is not given. An exception would only be the case if the study had had the task of testing the effectiveness of the study under ideal conditions, that is, the question whether it can in any way obtain effects. However, this is rarely the task in the social sciences, in contrast to medical research.

The threats to external validity – interactions between selection and treatment, interactions between setting and treatment, and interactions between history and treatment – have been dealt with elsewhere, so I shall not go into detail here. Although they cannot be eradicated completely by research planning and major material effort – for example, replication in various time intervals and various settings – they can be estimated and thereby approximately controlled. The long study intervals and the time pressure involved in evaluation tasks nonetheless almost certainly exclude the solution of controls through replication in practice. Here as well – thus at a central point – a scientifically satisfactory solution that will principally cope with the problems arising still needs to be found.

6. *Conclusions*

If we look at the relevant demands on evaluation research that face it from a variety of sides, the central demands are for (1) internal, statistical, and construct validity, because otherwise all statements become problematic; (2) generalizability; (3) applicability (i.e., replicability, transparency, ethical acceptability, etc.); and (4) "scientificalness" or the expansion of scientific knowledge. This results – as in economic policy – in a "magic square" of goals that cannot be achieved simultaneously (see Figure 3).

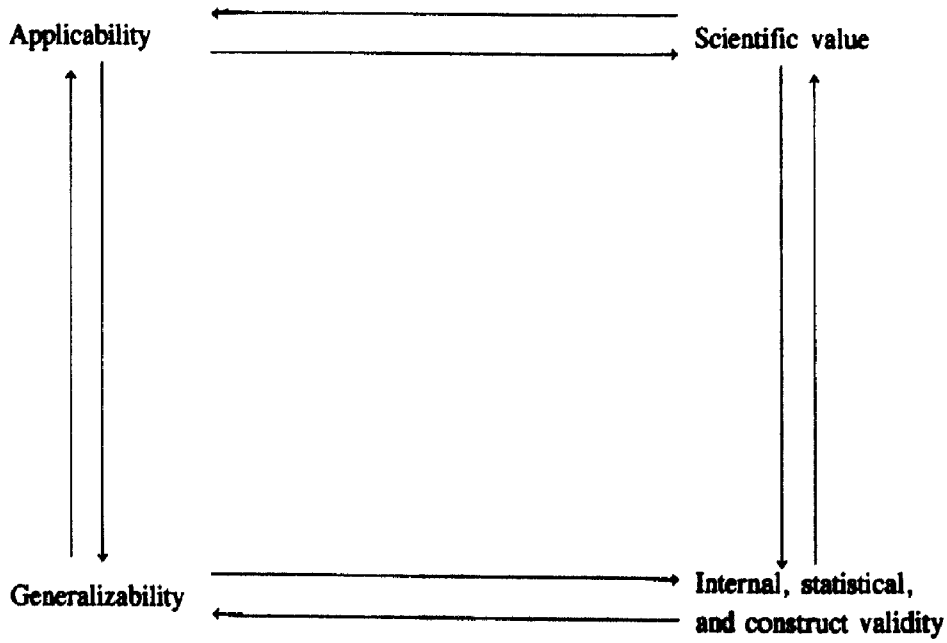

Fig. 3: Magic square of the goals of evaluation research.

Finally, brief mention should be made of one issue demonstrating that evaluation research is faced with an enormously complex task while simultaneously making it clear that it can potentially serve as a source of legitimization for the continuation of a traditional, but not necessarily beneficial, societal practice.

An insistence on the understandable demand that specific social innovations must be evaluated with randomized field experiments if they are to be applied generally frequently raises the problem that it is exceedingly difficult to obtain a sufficient number of cases that can be subjected to such an experimental procedure (see, e.g., Dennis, 1988; Dennis, 1990, pp. 358–359). The majority of studies that originally believed that this could be done were finally unable to achieve it. This is decisively significant. Lipsey (1988) has been able to show that the rate of Type II errors (i.e., not confirming a statistical significance when, in fact, a truly significant difference is present) above all depends on the size of the sample. Fifty-five percent of more than 1,400 studies for which meta-analyses revealed an effect size of .25 or more were unable to confirm a statistically significant difference because of insufficient power. The average sample size of the studies with a demonstrated Type II error was $N = 40$ per relation analyzed. Such a sample size would (with a power of .90) only permit a reliable determination of effect sizes of .75 or above. A great number of social experiments or reform programs that fundamentally prove to be worthwhile are thus mistakenly judged to be failures because they do not assess enough subjects. This may be because there simply are no more subjects available, or because insufficient funding was provided to recruit enough subjects – possibly precisely because the uncertainty of success had been used to justify an economical approach. This makes it clear that some kinds of economy can be very expensive. The purely theoretical, statistical argument that an overlarge sample size can lead to an inappropriate power and thus to other problems of interpretation (see Bortz, 1984, pp. 489–492) unfortunately never becomes a problem in this context.

If I try to draw a balance on my discussion of the chances of ascertaining the statistical, internal, external, and construct validity by optimizing the methodological measures *within* the classical concept of evaluation research that is committed to the experiment, I have to conclude that this strategy cannot completely achieve its goal, as the elimination of one source of invalidity gives rise to another and so on.

As a consequence, I consider that the further development of the methodology of evaluation research must clearly take the direction of *methodological pluralism*. Nonetheless, this only indicates the rough direction and does not

specify with which methodological approaches and with which expectations this task can be tackled.

The systematic clarification of the possibilities of development that suggest themselves and their critical evaluation should be given particular attention in the future, as most of the previously published work (with a small number of exceptions; see, e.g., Cronbach et al., 1981; Cronbach, 1982; Rossi, Freeman, and Wright, 1979) either only works out individual aspects or presents single-case studies that are difficult to evaluate in terms of generalizability.

I shall select some methodological problems that appear to be central and sketch the direction in which the search for new ideas should proceed.

1. *Randomization.* I have explained in detail why this procedure is so important for evaluation research (see, in addition, Campbell and Boruch, 1975, who demonstrate why quasiexperimental designs frequently underestimate effects of the treatment, that is, mistakenly claim that an effect is nonexistent) and why it is demanded so emphatically (see Gilbert, Light, and Mosteller, 1975; Lumsdaine and Bennett, 1975; Riecken, 1974; Edwards and Guttentag, 1975). On the other hand, I have also described the many problems that are involved in randomization. That there are nonetheless possibilities for true experimental design has been documented in many places (see, among others, Kennedy, 1981; Federal Judicial Center Advisory Committee on Experimentation in the Law, 1981; Baunach, 1980; and, for a discussion of experimental possibilities within criminology, Bönitz, 1984). The methodological considerations regarding this state of affairs should take the direction of testing what power and what statistical possibilities are available when randomization cannot be achieved for compelling reasons. The frequently maintained opinion that the chances of gaining tested statements is very slight without randomization is certainly exaggerated. Boruch and Rindskopf (1977) as well as Grizzle and Witte (1980) have demonstrated a number of possibilities of using a complete arsenal of useful procedures in these cases as well.
2. *Program evaluation or process evaluation.* My review of the threats to validity has repeatedly demonstrated that one of the highly critical values of the study is the assessment of the program and its realization. One has to avoid taking the official program as a true guide to action and making the goals it contains the measure of analysis (Deutscher, 1976). For obtaining effects, it is the daily routine of the program, the reality of the program implementation, that is significant. The actual treatment varies from person to person, from setting to setting, and from timepoint to timepoint.

With the usual techniques of standardized social research it is difficult to assess; although it may be possible through systematic, quantifying observation. However, this procedure also has its limitations through the fixation of the observer on his or her role, and so forth. The process character of the treatment can only be assessed appropriately if additional data can be obtained from a quasi-ethnographic perspective (see, among others, Britan, 1981) that permits the capture of the qualitative side of intervention. The research techniques of qualitative social research could provide completely major statements on process evaluation (see, e.g., Kriesberg, 1980), and qualitative research could bring important aspects to the fore that would complement the quantitative study (see, on the integration of quantitative and qualitative evaluation research, among others, Ianni and Orr, 1979; Guba, 1978; Knapp, 1979) and would clearly point beyond the orthodoxy of program evaluation (Cronbach, 1981; Cronbach et al., 1982; Fitz-Gibbon, Taylor, and Morris, 1978; Posavac and Carey, 1980). A further, very important and, in my opinion, even central effect could be obtained by combining process and impact evaluation. Remarkably, this is very frequently neglected (see, regarding this possibility, Hollister, Kemper, and Wooldridge, 1979; Hughes, Cordray, and Spiker, 1981). As the discussion develops, it becomes clear that the massive conflicts and the hardened fronts are in a process of dissolution. What better confirmation can be found for this than that one of the most prominent supporters of the experimental and quasi-experimental approach, D.T. Campbell, has conceded even the single-case study a thoroughly significant role in the framework of evaluation research (Campbell, 1979)? However, this also makes it clear that the methodology of evaluation research has entered a crisis of orientation. Almost anything goes, but does it also work? An evaluative stocktaking appears to be urgently necessary.

3. *Quasi-experimental designs.* Quasi-experimental designs, which have been described repeatedly above, offer a variety of study possibilities of highly differing methodological quality that urgently require further development, as they provide many unexploited opportunities (see, once again, Cook and Campbell, 1979; and particularly regarding the issue of the statistical analysis of experimental and, above all, quasi-experimental research findings, Judd and Kenny, 1981). For several reasons, interrupted time-series designs deserve particular attention; on the one hand, because they permit statements on very interesting questions in many cases without the need for great effort in collecting data; and, on the other hand, they provide opportunities for very differentiated and powerful statistical analysis that have only recently been recognized by social scientists (see, among others, Cook, Dintzer, and Mark, 1980; McCleary and Hay, 1980; Judd and

Kenny, 1981; Renn and Mariak, 1984; Trochim, 1984). Developments toward methodological pluralism and the multimethod approach can additionally be seen here; for example, in the combination of time-series analysis with interviews (Marsh, 1981). For other forms of quasi-experimental research design, a statistical heuristic is also developing that helps to gain a maximum of information from the data obtained in this way (see, e.g., Cain, 1975; Judd and Kenny, 1981; Novick, 1981; Rindskopf, 1981). A series of individual examples makes it clear that (1) the potentials of quasi-experimental research are in no way fully exploited; and (2) the canonization of these methods has yet to be performed.

All this should make it clear that the evaluation of programs of prevention and intervention represents a thoroughly demanding task that requires the assessment of complex social and individual processes with appropriate methods. The conditions under which this has to occur balk at the simplifying assumptions of the classical tradition of research but, through the development of an independent heuristic, they have to be harnessed to the extent that a rational discussion on the basis of evaluation research will become possible.

On the other hand, it is important not to overlook the fact that the social relevance of evaluation research does not just depend on the "methodological" quality measured with the canon of strict scientific rules: The utilization of the knowledge gained from evaluation research depends on a great number of other variables that only partially depend on the structure of evaluation research (see Albrecht, 1982; Lösel, and Nowack, 1987, pp. 80–83). Interestingly, a closer inspection of the necessary conditions for the use of the knowledge gained (e.g., consideration of the cognitive style of the policy makers, speed of obtaining information, selection of research designs and methods of data analysis that can also be understood by laypersons) reveals that they tend to clash with the strategies for ascertaining the precision and truth or validity of the findings. Therefore, it has to be seen that there is not only a conflict of aims between the various strategies to ascertain the validity of findings but that conflicts between the goals of maximizing the validity and maximizing the probability of utilization are also unavoidable.

While sneaking a look at the probability of utilization should naturally not tempt the conscientious social researcher to abandon scientific standards in an unacceptable way, Lösel and Nowack (1987, p. 83; see also Wottawa and Thierau, 1990, p. 166) have drawn attention to yet another issue: Their criticism is that there is no fundamental reason why the utilization of scientific knowledge should have to be positive. Deficits in predictability and the control

of planning permit the variety that a society needs in order to adjust flexibly to changing circumstances. I therefore support their conclusion "that with regard to the techniques of evaluation, there is a need for the most imaginative, differentiated instruments and for flexibility as well as an understanding of the problems involved in their application" (Lösel and Nowack, 1987, p. 83, translated). Dogmatism is (also) no solution here.

References

Abt, C.C. (Ed.)(1976). The evaluation of social programs. Beverly Hills: Sage.

Albrecht, G. (1982). Muß angewandte Soziologie konforme Soziologie sein? Zum Verhältnis von Theorie und angewandter Soziologie im Bereich des abweichenden Verhaltens und der sozialen Kontrolle. In U. Beck (Ed.), Soziologie und Praxis (pp. 161–204). Göttingen: Schwartz.

Bastine, R. (1975). Forschungsmethoden in der klinischen Psychologie. In W.J. Schraml and U. Baumann (Eds.), Klinische Psychologie: Vol. I, (3rd ed.). Bern: Huber.

Baunach, P.J. (1980). Random assignment in criminal justice research: Some ethical and legal issues. Crimonology, 17, 435–444.

Bennett, C.A., and Lumsdaine, A.A. (1975). Social program evaluation: Definitions and issues. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 1–38). New York/San Francisco: Academic Press.

Biefang, S. (1980). Evaluationsforschung in der Psychiatrie. Stuttgart: Enke.

Blass-Wilhelms, W. (1983). Evaluation im Strafvollzug. Überblick und Kritik vorliegender Studien. In H. Kury (Ed.), Methodische Probleme der Behandlungsforschung – insbesondere in der Sozialtherapie (pp. 81–119). Köln: Carl Heymanns Verlag.

Bönitz, D. (1984). Experimentelle Forschungsmöglichkeiten in der Krimonologie. In H. Kury (Ed.), Methodologische Probleme in der kriminologischen Forschungspraxis (pp. 287–306). Köln: Carl Heymanns Verlag.

Bortz, J. (1984). Lehrbuch der empirischen Forschung für Sozialwissenschaftler. Berlin: Springer.

Boruch, R.F., and Gomez, M. (1977). Sensitivity, bias and theory in impact evaluation. Professional Psychology, 8, 411–434.

Boruch, R.F., and Rindskopf, D. (1977). On randomized experiments, approximations to experiments, and data analysis. In L. Rutman (Ed.), Evaluation of research methods: A basic guide (pp. 143–176). Beverly Hills: Sage.

Britan, G.M. (1981). Contextual evaluation: An ethnographic approach to program assessment. In R.F. Connor (Ed.), Methodological advances in evaluation research (pp. 47–60). Beverly Hills/London: Sage.

Cain, G.G. (1975). Regression and selection models to improve non-experimental comparisons. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 297–317). New York/San Francisco: Academic Press.

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

Campbell, D.T. (1969). Reforms as experiments. American Psychologist, 24, 409–429.

Campbell, D.T., and Boruch, R.F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 195–295). New York/San Francisco: Academic Press.

Campbell, D.T. (1979). "Degrees of freedom" and the case study. In T.D. Cook and C.S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 49–67). Beverly Hills/London: Sage.

Caro, F.G. (Ed.) (1971). Readings in evaluation research. New York: Russell Sage Foundation.

Cook, T.D., and Campbell, D.T. (1979). Quasi-experimentation. Design and analysis issues for field settings. Boston: Houghton Mifflin Comp.

Cook, T.D., Leviton, L.C., and Shadish, W.R. (1985). Program evaluation. In G. Lindzey and E. Aronson (Eds.), The handbook of social psychology, (3rd ed.), (pp. 699–777). New York: Random House.

Cook, T.D., Dintzer, L., and Mark, M.M. (1980). The causal analysis of concomitant time series. In L. Bickman (Ed.), Applied social psychology annual: Vol. 1 (pp. 93–135). Beverly Hills/London: Sage.

Cronbach, L.J. (1981). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Cronbach, L.J. (1982). Designing of educational and social programs. San Francisco: Jossey-Bass.

Dennis, M.L. (1988). Implementing randomized field experiments: An analysis of criminal and civil justice research. Diss. Northwestern University.

Dennis, M.L. (1990). Assessing the validity of randomized field experiments. An example from drug abuse treatment research. Evaluation Review, 14, 347–373.

Dennis, M.L., and Boruch, R.F. (1989). Randomized experiments for planning and testing projects in developing countries: Threshold conditions. Evaluation Review, 13, 292–309.

Deutscher, I. (1976). Toward avoiding the goal-trap in evaluation research. In C.C. Abt (Ed.), The evaluation of social programs (pp. 249–268). Beverly Hills/London: Sage.

Edwards, W., and Guttentag, M. (1975). Experiments and evaluations: A reexamination. In C.A. Bennett and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 409–463). New York/San Francisco: Academic Press.

Federal Judicial Center Advisory Committee on Experimentation in the Law (1981). Experimentation in the law. Report of the Federal Judicial Center Advisory Committee on experimentation in the law. Washington, D.C.: FJC.

Fitz-Gibbon, C.T., and Morris, L.L. (1978). How to design a program evaluation. Beverly Hills/London: Sage.

Gilbert, J.P., Light, R.J., and Mosteller, F. (1975). Assessing social innovations: An empirical base for policy. In C.A. Bennett, and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 39–193). New York/ San Francisco: Academic Press.

Grizzle, G.A., and Witte, A.D. (1980). Criminal justice evaluation techniques: Methods other than random assignment. In M.W. Klein, and Teilmann, K.S. (Eds.), Handbook of criminal justice evaluation (pp. 259–302). Beverly Hills/London: Sage.

Gruschka, A. (Ed.) (1976). Ein Schulversuch wird überprüft. Das Evaluationsdesign für die Kollegstufe NW als Konzept handlungsorientierter Begleitforschung. Kronberg: Athenäum.

Guba, E.G. (1978). Toward a methodology of naturalistic inquiry in educational evaluation, C.S.E. Monograph Series in Evaluation, No. 8. Los Angeles: Center for the Study of Evaluation.

Guttentag, M., and Struening, E.L. (Eds.) (1975). Handbook of evaluation research (Vols. 1–2). Beverly Hills/London: Sage.

Hartig, M. (1975). Probleme und Methoden der Psychotherapieforschung. München: Urban und Schwarzenberg.

Hellstern, G.-M., and Wollmann, H. (1983). Bilanz-Reformexperimente, wissenschaftliche Begleitung und politische Realität. In G.-M. Hellstern, and H. Wollmann (Eds.), Experimentelle Politik – Reformstrohfeuer oder Lernstrategie? (pp. 2–77). Opladen: Westdeutscher Verlag.

Hirst, E. (1981). Combining archival search with telelphone surveys: Evaluating energy information centers. In R.F. Conner (Ed.), Methodological advances in evaluation research (pp. 127–139). Beverly Hills/London: Sage.

Hollister, R.G. Jr., Kemper, P., and Wooldridge, J. (1979). Linking process and impact analysis: The case of supported work. In T.D. Cook and C.S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 140–158). Beverly Hills/ London: Sage.

House, E.R. (1980). Evaluating with validity. Beverly Hills/ London: Sage.

Hughes, S.L., Cordray, D.S., and Spiker, A. (1981). Combining process with impact evaluation: A long-term home care program for the elderly. In R.F. Conner (Ed.), Methodological advances in evaluation research (pp. 109–125). Beverly Hills/London: Sage.

Ianni, F.A.J., and Orr, M.T. (1979). Toward a rapprochement of quantitative and qualitative methodologies. In T.D. Cook and C.S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 87–98). Beverly Hills/London: Sage.

Judd, C.M., and Kenny, D.A. (1981). Estimating the effects of social interventions. Cambridge: University Press.

Kennedy, M.M. (1981). The role of experiments in improving education. In C.B. Aslanian (Ed.), Improving educational evaluation methods. Impact Policy (pp. 67–77). Beverly Hills/London: Sage.

Knapp, M.S. (1979). Ethnographic contributions to evaluation research: The experimental schools program evaluation and some alternatives. In T.D. Cook, and C.S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research (pp. 118–139). Beverly Hills/London: Sage.

Krauth, J. (1981). Statistische Methoden zur Veränderungsmessung. In U. Baumann, H. Berbalk, and G. Seidenstücker (Eds.), Klinische Psychologie. Trends in Forschung und Praxis: Vol. 4 (pp. 98–131). Bern/Stuttgart/Wien: Huber.

Kriesberg, B. (1980). Utility of process evaluation: Crime and delinquency programs. In M.W. Klein and K.S. Teilmann (Eds.), Handbook of criminal justice evaluation (pp. 217–236). Beverly Hills/London: Sage.

Lange, Elmar (1983). Zur Entwicklung und Methodik der Evaluationsforschung in der Bundesrepublik Deutschland. Zeitschrift für Soziologie, 12, 253–270.

Lösel, F., and Nowack, W. (1987). Evaluationsforschung. In Schultz-Gambard, J. (Ed.). Angewandte Sozialpsychologie. Konzepte, Ergebnisse, Perspektiven (pp. 57–87). München/Weinheim: Psychologie Verlags Union.

Lumsdaine, A.A. and Bennett, C.A. (1975). Assessing alternative conceptions of evaluation. In C.A. Bennett, and A.A. Lumsdaine (Eds.), Evaluation and experiment. Some critical issues in assessing social programs (pp. 525–553). New York/San Francisco: Academic Press.

McCleary, R., and Hay, R.A. (1980). Applied time series analysis for the social sciences. Beverly Hills/London: Sage.

Marsh, J.C. (1981). Combining time series with interviews: Evaluating the effects of a sexual assault law. In R.F. Conner (Ed.), Methodological advances in evaluation research (pp. 93–108). Beverly Hills/London: Sage.

Novick, M.R. (1981). Data analysis in the absence of randomization, In R.F. Boruch, P.M. Wortman, D.S. Cordray and Associates, Reanalyzing program evaluations (pp. 144–162). San Francisco: Jossey-Bass.

Patton, M.Q. (1980). Qualitative evaluation methods. Beverly Hills/London: Sage.

Perloff, R., Perloff, E., and Sussna, E. (1976), Program evaluation. Annual Review of Psychology, 27, 569–594.

Petermann, F. (1978). Veränderungsmessung. Stuttgart: W. Kohlhammer

Phillips, L. (1980). Cost analysis. In M.W. Klein and K.S. Teilmann (Eds.), Handbook of criminal justice evaluation (pp. 459–472). Beverly Hills/London: Sage.

Pillemer, D.B., and Light, R.J. (1981). Using the results of randomized experiments to construct social programs. In R.F. Boruch, P.M. Wortman, D.S. Cordray and Associates, Reanalyzing program evaluations (pp. 225–236). San Francisco: Jossey-Bass.

Posavac, E.J., and Carey, R.G. (1980). Program evaluation methods and case studies. Englewood Cliffs, N.J.: Prentice-Hall.

Reichardt, C.S., and Cook, T.D. (1979). Beyond qualitative versus quantitative methods. In T.D. Cook and C.S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills/London: Sage.

Renn, H., and Mariak, V. (19849. Interventionsanalyse kriminologischer Zeitreihen. In H. Kury (Ed.), Methodologische Probleme in der kriminologischen Forschungspraxis (pp. 153–219). Köln: Carl Heymanns Verlag.

Riecken, H.W., and Boruch, R.F. (Eds.)(1974). Social experimentation. New York: Academic Press.

Rindskopf, D.M. (1981). Structural equation models in analysis of nonexperimental data. In

Boruch, R.F., Wortman, P.M., Cordray, D.S., et.al. (Eds.). Reanalyzing program evaluation (pp. 163–193). San Francisco: Jossey Bass.

Rossi, P.H., Freeman, H.E., and Wright, S.R. (1979). Evaluation. A systematic approach. Beverly Hills/London: Sage.

Rossi, P.H., and Freeman, H.E. (1989). Evaluation. A systematic approach, 4th ed. Newbury Park etc.: Sage.

Rossi, P.H., and Williams, W. (1972). Evaluating social programs. New York: Academic Press.

Scriven, M. (1974a). The concept of evaluation. In Apple, M.W., Subkoviak, M.J., and Lufler, H.S. (Eds.). Educational evaluation: analysis and responsibility (pp. 55–82). Berkeley, Cal.: McCutchan.

Scriven, M. (1974b). Evaluation perspectives and procedures. In Popham, W.J. (Ed.). Evaluation in education: Current applications (pp. 1–93). Berkeley, Cal.: McCutchan.

Sechrest, L., and Redner, R. (1979). Strength and integrity of treatment in evaluation studies. Washington, D.C.: National Criminal Justice Reference Service. National Institute of Law Enforcement and Criminal Justice. Law Enforcement Assistance Administration.

Smith, N.L. (1981). Developing evaluation methods. In Smith, N.L. (Ed.). Metaphors for evaluation. Sources of new methods (pp. 17–49). Beverly Hills/London: Sage.

Smith, N.L. (1981). Metaphors for evaluation. In Smith, N.L. (Ed.). Metaphors for evaluation. Sources of new methods (pp. 51–65). Beverly Hills/London: Sage.

Trend, M.G. (1979). On the reconciliation of qualitative and quantitative analysis: A case study. In Cook, T.D., and Reichardt, C.S. (Eds.). Qualitative and quantitative methods in evaluation research (pp. 68–86). Beverly Hills/London: Sage.

Trochim, W.M.K. (1984). Research design for program evaluation. The regression-discontinuity approach. Beverly Hills: Sage.

Weiss, C.H. (1970). The politization of evaluation research. Journal of Social Issues, 26, 57–68.

Weiss, C.H. (1974). Evaluierungsforschung. Methoden zur Einschätzung von sozialen Reform-programmen. Opladen: Westdeutscher Verlag.

Weiss, C.H., and Bucuvalas, M.J. (1977). The challenge of social research to decision making. In Weiss, C.H. (Ed.). Using social research in public policy making (pp. 213–233). Lexington, Mass.: Lexington Books.

Wittmann, W.W. (1985). Evaluationsforschung: Aufgaben, Probleme und Anwendungen. Berlin: Springer.

Wortman, Paul M. (1983). Evaluation research: A methodological perspective. Annual Review of Psychology, 34, 223–260.

Wottawa, H., and Thierau, H. (1990). Evaluation. Berlin/Stuttgart/Toronto: Verlag Hans Huber.

Zielke, M. (1982). Probleme und Ergebnisse der Veränderungsmessung. In Zielke, M. (Ed.). Diagnostik in der Psychotherapie (pp. 41–59). Stuttgart: Kohlhammer.