TRACES are what evaluators left behind—discoveries, records, tracks—which made marks on the profession of program evaluation. Published here are excerpts from our past (e.g., articles, passages from books, speeches) that show where we have been or affected where we are g(r)o(w)ing. Suggestions for inclusion should be sent to the Editor, along with rationale for their import. Photocopies of the original printed versions are preferred with full bibliographic information. Copying rights will be secured by the Editor.

Editor's Note: We are proud to present here a widely distributed but previously unpublished paper by Donald Campbell. After publication of his milestone paper, "Reforms as Experiments," in 1969 (*American Psychologist,* 24(4), 409–429), Campbell wrote a series of papers elaborating on his ideas for an experimenting society. "Methods for the Experimenting Society," printed here, was the basis for his 1971 "abbreviated and extemporaneous" presentations to the Eastern Psychological Association and the American Psychological Association. In the paper, Campbell proposed that a few social scientists dedicate themselves to being methodologists for the experimenting society. He then recounted many of the problems that must be addressed and discussed possible solutions.

In 1988, the 1971 draft was revised and published in a selection of Campbell's papers entitled *Methodology and Epistemology for Social Science* (edited by E. Samuel Overman and published by The University of Chicago Press). But the original version, here in *EP*, contains more detail on the methods. Certainly we can trace a great deal of our evaluation history to Dr. Campbell. We are grateful for his permission to publish the paper in *EP*.

Methods for the Experimenting Society

DONALD T. CAMPBELL

The experimenting society will be one which will vigorously try out proposed solutions to recurrent problems, which will make hard-headed and multidimensional evaluations of the outcomes, and which will move on to try other alternatives when evaluation shows one reform to have been ineffective or harmful. We do not have such a society today. While all societies are engaged in trying out innovative reforms, none is yet organized to adequately evaluate the outcomes. Programs are instead continued or discontinued on inadequate or other grounds. This is due in part to the inertia of social organizations, in part to political predicaments which oppose evaluation, and in part to the fact that the methodology of evaluation is still inadequate.

Evaluation Practice, Vol. 12, No. 3, 1991, pp. 223–260.
ISSN: 0191-8036

While the experimenting society is nowhere yet an actuality, there are in many countries political forces moving in this direction. Finland, England, Yugoslavia, Poland, Czechoslovakia, Chile, and the United States come to mind as a few among the many nations in which such a society might emerge. In developing an experimenting society, the skills of the social scientist will be of utmost importance. What I am proposing in this paper is that a few of us the world over now self-consciously dedicate ourselves to being "methodologists for the experimenting society," that we now tool up for that future day by thinking through as well as we can the problems that will emerge. This paper is a preliminary overview of the field, a survey of problems we could now begin working on. It is broader in scope, but builds upon "Reforms as Experiments." (Campbell, 1969. See also Campbell, 1970.)

THE EXPERIMENTING SOCIETY

Before getting into the methodological details, a little more needs to be said about the nature of the experimenting society as an ideology for a political system (Haworth, 1960).

It will be an *active society* (Etzioni, 1968) preferring exploratory innovation to inaction. It would be a society which experiments, tries things out, explores possibilities in action (as well as, or even instead of, in thought and simulation). It would borrow from epistemology and the history of science the truism that one cannot know for certain in advance, that a certain amount of trial-and-error is essential. Faced with a choice between innovating a new program or commissioning a thorough study of the problem as a prelude to action, its bias would be toward innovating. It will be committed to *action research*, to action as research rather than research as a postponement of action (Lewin, 1946; Sanford, 1970). It will be an *evolutionary, learning society* (Dunn, 1971).

It will be an *honest society*, committed to *reality testing*, to self-criticism, to avoiding self-deception. It will say it like it is, face up to the facts, be undefensive and open in self-presentation. Gone will be the institutionalized bureaucratic tendency to present only a favorable picture in government reports. For many a civil servant on both sides of the Iron Curtain this freedom to be honest will be one of the strongest attractions of the experimenting society. The motive of honesty in political reform, revolution, and personal heroism has been generally neglected until recently. It is of course now a dominant theme among our young idealists, showing up in their standards for their own interpersonal relations and in their criticism of the cowardly hypocrisy, double-talk, and dishonesty of their elders. It also emerges as a major political force within the Communist countries, as Polanyi (1966) has argued in the case of the Hungarian uprisings of 1956. While that revolution no doubt had complex roots, many of which might accurately be called reactionary or Fascist, Polanyi persuades me in his analysis of the motives of the "Petöfi Circle." These elite communist journalists, well rewarded by their establishment with power and wealth, were motivated by the pain of

continually having to write lies and by the promise of a society in which they could write the truth as they saw it. Honesty was also among the prime motives of the Czechoslovak leadership and followership of 1968, accounting for both its great popularity and its disfunctionally provocative excesses in exposing past lies.[1]

It will be a *nondogmatic society*. While it will state ideal goals and propose wise methods for reaching them, it will not dogmatically defend the value and truth of these goals and methods against disconfirming evidence or criticism.

It will be a *scientific society* in the fullest sense of the word "scientific." The scientific values of honesty, open criticism, experimentation, willingness to change once-advocated theories in the face of experimental and other evidence will be exemplified. This usage should be distinguished from an earlier use of the term scientific in social planning. In this older usage, one scientific theory is judged to be established as true. On the basis of this scientific theory, extrapolations are made to the design of an optimal social organization. This program is then put into effect, but without explicit mechanisms for testing the validity of the theory through the results of implementing it. Such social planning becomes dogmatic, non-experimental, and is not scientific in the sense used here even though the grounds of its dogma are the product of previous science. Such dogmatism has been an obvious danger in implementations of Marxist socialism. It is also a common bias on the part of governmental and industrial planners everywhere. In such planning, there is detailed use of available science but no use of the implemented program as a check on the validity of the plans or of the scientific theories upon which they were based. Thus economists, operations researchers and mathematical decision theorists trustingly extrapolate from past science and conjecture, but in general fail to use the implemented decisions to correct or expand that knowledge.

It will be an *accountable, challengeable, due-process society*. There will be public access to the records on which social decisions are made. Recounts, audits, reanalyses, reinterpretations of results will be possible. Just as in science objectivity is achieved by the competitive criticism of independent scientists, so too the experimenting society will provide social organizational features making competitive criticism possible at the level of social experimentation. There will be sufficient separation of governmental powers so that meaningful legal suits against the government are possible. Citizens not a part of the governmental bureaucracy will have the means to communicate with their fellow citizens disagreements with official analyses and to propose alternative experiments. It will be an *open society* (Popper, 1945).

It will be a *decentralized society* on all feasible aspects. Either through autonomy or deliberate diversifications, different administrative units will try out different ameliorative innovations and will cross-validate those discoveries they borrow from others. The social-system independence will provide something of the replication and verification of successful experiments found in science. Semiautonomy will provide some of the competitive criticism that make for scientific objectivity.

It will be a society committed to *means-idealism* as well as *ends-idealism*. As in modern views of science, the process of experimenting and improving will be expected to continue indefinitely without reaching the asymptote of perfection. In this sense, all

future periods will be mediational, transitional, rather than perfect goal states. Ends cannot be used to justify means, for all we can look forward to are means. The means, the transitional steps, must in themselves be improvements.

It will be a *popularly responsive society*, whose goals and means are determined by collective good and popular preference. Within the limits determined by the common good, it will be a *voluntaristic society*, providing for individual participation and consent at all decision-levels possible. It will be an *equalitarian society*, valuing the well-being and the preferences of each individual equally.

* * *

This too brief sketch has glossed over many problems. A few of these are considered later on in the paper. A few need be considered now. While the experimenting society thus described has many attractive features, it obviously has many dangers and costs. We methodologists for the experimenting society should be sensitive to our own and others' ambivalences about it. We should anticipate its dangers and misuses as well as its promises. We should of course design ways of obviating these where possible. But we should keep open the possibility that we will end up opposing it. We who now specialize in thinking about the experimenting society should be the first to decide that it is unworkable or in the new undesirable, if it is and if this can be ascertained in advance of trying it.

The ideals described as characterizing the experimenting society are for the most part ideals that all of today's major ideologies claim as their own. They are endorsed in both communist and capitalist countries, most clearly in each one's criticism of the other. While neither type of society is as yet achieving these ends, the experimenting society could grow out of either, or out of both. A competition to see which could best and soonest implement it might even be envisaged. In any event, we methodologists for the experimenting society should consider the problems of implementing it under all forms of political-economic organization, and should regard designing all routes as a part of our methodological challenge. Such universalism is made easier both by the common ideals, and by the fact that the ideology of the experimenting society is a method ideology, not a content ideology. That is, it proposes ways of testing and revising theories of optimal political-economic-social organization rather than proposing a specific political and economic system. (Of course this is too simple, for the requirements of the experimenting society exclude many forms of political-economic-social organization, but those excluded are not being advocated by either side of the cold war.)

Once implemented in both capitalist and communist countries, one might expect that social experimentation would tend to produce increased similarity of social organization, much as industrialization has tended to do. This expectation is of course based upon the assumption of some universals in human preferences for the good life and the good society. Once these preferences become the selective criteria for choosing among alterations of existing forms, these universals will presumably shape societies

toward a common optimum. (This needs qualifying in terms of the theory of the conditions of convergence and nonconvergence in iterative processes.)

There are of course forces operating against the development of the experimenting society in both capitalist and communist countries, both in ideology as well as in current practice. Thus the Marxist-Leninist commitment to the necessity of a dictatorship during the transitional phase has served in practice to justify a dogmatism and intolerance of criticism that are inconsistent with the experimenting society. Yet in the total complex body of Marxist theory there are stated perspectives and ideals quite sufficient to justify a truly experimental socialism (Schaff, 1960; Skolimowski, in press; Feldman and Campbell, in preparation).

Within western democratic capitalism, there are a number of favorable features. These include the legal tradition, the successful achievement of changes in government through elections, and the genuine pluralism of decision-making units. The so-called "market mechanisms" of capitalist economic theory can be regarded in ideal form as self-regulatory cybernetic feedback systems implementing the collective aspects of the preferences of individual decision makers. But the ideological justification and effective practice of the accumulation of great inequalities in individual and corporate wealth, and the role of wealth in providing grossly uneven weightings of some persons' preferences over those of others, provide great obstacles which may effectively sabotage program decisions genuinely based on the public good. Whatever one's long-range expectations on this matter, it should be noted that the U.S. government under both the Johnson and Nixon administrations, in programs such as the Office of Economic Opportunity, has made great strides in this direction, providing the social science methodology community with more opportunities to work on developing methods for the experimenting society than they have been ready for.

Within both capitalist and communist countries there are universal aspects of political processes which work against the emergence of the experimenting society and which may in the long run preclude it. Among these are institutional inertia, the general opposition to all change and innovation. More important are the several sources of fear of evaluation. The measurement machinery that is used to reflect upon program effectiveness can also be used to evaluate the efficacy of administrators and of organizational units. Such evaluation, valid or invalid, is understandably feared, and is more to be feared the more elaborate and scientific it appears to be.

Over and above this general fear of evaluation are fears that are specifically linked to program innovation itself. The *overadvocacy trap* merges from the extreme difficulty within any political system or bureaucracy of getting a new program adopted. The advocacy involved is almost certain to make exaggerated claims as to the degree and certainty of the program's effectiveness. In this ubiquitous condition hardheaded evaluation of the program's effectiveness involves almost certain political jeopardy. The advocacy of different reforms by competitors for positions of power exacerbates this threat. Anticipation of marginal effects due to program focus on the failures of preponderantly effective general services, or due to the trivial magnitude of the program as implemented, provide understandable grounds for opposition to evaluation (Rossi, 1969).

The Social Scientists as Servant of the Experimenting Society

Societies will continue to use preponderantly unscientific political processes to decide upon ameliorative program innovations. Whether it would be good to increase the role of social science in deciding on the content of the programs tried out is not at issue here. The emphasis is rather on the more passive role for the social scientist as an aid in helping society decide whether or not its innovations have achieved desired goals without damaging side effects. The job of the methodologist for the experimenting society is not to say *what is to be done*, but rather to say *what has been done*. The aspect of social science that is being applied is primarily its research methodology rather than its descriptive theory, with the goal of learning more than we do now from the innovations decided upon by the political process. As is elaborated below, even the conclusion drawing and the relative weighing of conflicting indicators, must be left up to the political process.

This emphasis seems to be quite different from the present role as government advisors of most economists, international relations professors, foreign area experts, political scientists, sociologists of poverty and race relations, psychologists of child development and learning, etc. Government asks what to do, and scholars answer with an assurance quite out of keeping with the scientific status of their fields. In the process, the scholar-advisors too fall into the overadvocacy trap and fail to be interested in finding out what happens when their advice is followed. Certainty that one already knows precludes finding out how valid one's theories are. We social scientists could afford more of the modesty of the physical sciences, should more often say that we can't know until we've tried. For the great bulk of social science where we have no possibility of experimental probing of our theories, we should be particularly modest. While the experiments of the experimenting society will never be ideal for testing theory, they will probably be the best we have, and we should be willing to learn from them even when we have not designed them. More importantly, measuring the effects of a complex politically designed ameliorative program involves all of the problems of experimental inference found in measuring the effects of a conceptually pure treatment variable — all and more. The scientific methods developed for the latter are needed for ameliorative program evaluation. With the most minor of exceptions, it can be said for the United States that none of our major ameliorative programs have had adequate evaluations (Schwartz, 1961; Hyman and Wright, 1967; Etzioni, 1968; Rossi, 1969; Campbell, 1969a, Wholey, 1970; Campbell and Erlebacher, 1970; Caro, 1971). There is general agreement among these authors that the prospectively designed experiment offers the best possibilities of evaluation.

The distinction is overdrawn. It reflects my own judgment that in the social sciences, including economics, we are scientific by intention and effort, but not yet by achievement. We have no elegantly successful theories that predict precisely in widely different settings. Nor do we have the capacity to make definitive choices among competing theories. Even if we had, the social settings of ameliorative programs involve so many complexities that the guesses of the experienced administrator and politician are apt to be on the average as wise as those of social scientists. But whatever the source of the implemented guess, we learn only by checking it out. Certainly in the

experimenting society, social scientists will continue to be called upon to help design solutions to social problems, and this is as it should be. Perhaps all I am advocating in emphasizing the role of servant rather than leader, is that social scientists avoid cloaking their recommendations in a specious pseudo-scientific certainty, and instead acknowledge their advice as consisting of but wise conjectures that need to be tested in implementation.

The servant-leader contrast is overdrawn in other senses also. The truism that measurement itself is a change agent is particularly applicable to the experimenting society. Advocating hardheaded evaluation of social programs is a recommendation for certain kinds of political institutions. In considering the methodological challenges of the experimenting society in what follows, appeals to the theory and content of the social sciences will be made, as well as to their methodology.

The plan of the following sections is to take standard methodological problems and to reconsider them as they relate to the experimenting society. This will uncover a host of problems that will have to be solved, and on many of which methodologists for the experimenting society could now be working. There will be fragmentary glimpses of what an experimenting society will be like, and these may give us grounds for rejecting it.

Problems of experimental design is considered first, true experiments, then quasi-experiments. Then problems of measurement: procedures, validity, and bias. In all of these, application to the experimenting society turns out to raise a host of new methodological issues. Following these, more general issues of scientific method are considered, such as criticism and replication. Fina lly, the problem of legislative assimilation of evaluational evidence is considered.

Uses of Experimentation in the Experimenting Society

The experimenting society will attempt experimental evaluations of all program innovations for which such is possible and useful. This essay advocates greatly expanded use of experimental evaluations. Yet it would be wrong to demand that only those innovations be made which can be experimentally evaluated. We must be wary of using a requirement for hardheaded evaluation as a grounds for inaction or postponement. We must recognize that in any science usable experimental laboratories are unevenly distributed and may not be available for any given problem. Thus the shift to new math in elementary education was both tried out and established without experimental evidence for its superiority (because the new math made the available achievement tests irrelevant). It has been plausibly argued that many broad-aim programs may be unevaluable (Weiss and Rein, 1969 , 1970; Campbell, 1970). There will also be instances in which reasonable estimates of the possible informational gain fail to justify the costs of experimental evaluation. None of these conditions should be used to argue against implementing a new program unless superior alternatives are at hand.

Experiments are today most common as pilot programs, testing out on experimental populations and for experimental periods programs or aspects of programs that are candidates for adoption as national policy. The New Jersey negative income tax

experiment (Watts 1969; 1971; Office of Economic Opportunity, 1971; Orr, Hollister, Lefcowitz and Hester, 1971) and the others that are following it are the most vivid examples. This is a most important function and may remain the most frequent occasion of research designs involving the random assignment of social units to experimental and control treatments. But such pilot programs will inevitably suffer in generalizability due to some degree of artificiality, reactive measurements, Hawthorne effects, etc.

Experimental design is also important although more difficult, when an innovation is adopted as national policy. At such a time, the results of the pilot experiment should be cross-validated in full application. There are perhaps four general models as to how this might be done. If the introduction of the new policy will be spread out in time with some regions being affected sooner than others, a policy of staged implementation which selects the earliest and latest regions for maximum comparability can achieve true experimental status for first year effects, with the added bonus of having the effect demonstrated a second time when implemented in the delayed-introduction regions (See Parker, 1963, for an unplanned version of such a design). The period of introduction thus may provide the last good chance for learning something of the effect of a new national program. But at such a time, the curiosity about its effects may be less than zero, either through post-decision dissonance-avoiding mechanisms, or due to a feeling that the information would be of no use now that the decision has been irrevocably made. While the likelihood of immediate use is undoubtedly less postdecisionally, the future use of the information in later reconsiderations and modifications justifies the effort, and also does its value for the science which will be used in designing later innovations and the metascience involved in extrapolations from the pilot experiments to implemented policy. This staged implementation experiment is not ideal, because it is limited to initial reactions to the new policy. The social mechanisms of accommodation, exploitation, and evasion may not have emerged in full strength, or, more hopefully, the full impact may not yet be felt. Generalizing from early implementation to well established custom has the same kind of problems that generalizing from pilot experiment to implemented policy have, though probably to a lesser degree.

A second method available for evaluating new programs affecting all persons is the interrupted time-series quasi-experimental design (Campbell, 1969, pp. 412–419; Ross, Campbell and Glass, 1970). This requires the existence of relevant data series and an abrupt rather than gradual implementation. If similar states or nations exist which are not making the change at the same time, but are similar in other potential change agents, then control-series add inferential strength. For this design as for the staged introduction, only program *changes*, not ongoing programs, can be evaluated.

Many ameliorative programs, particularly compensatory ones, do not affect all citizens and are in short supply. Such programs can be experimentally evaluated while in regular operation, as well as in pilot or in onset. Random assignment from groups of eligibles larger than program capacity makes possible true experiments. Quantification of the eligibility criteria and use of a sharp cutting point makes possible the regression- discontinuity quasi-experimental design (Campbell, 1969, pp. 419–426).

In addition to experiments by government, the experimenting society should make possible the use of scientific experimental techniques by citizen groups indepen-

dent of, or in criticism of government (Gordon, 1971). Haussenstamm's use of experimentally introduced Black Panther bumper stickers to demonstrate discriminatory enforcement of traffic laws provides one example (1970; See also Gordon and Myers, 1970).

Program evaluations abound. Most are anecdotal, or document the delivery of services without estimates of the effects of those services. The more elaborate ones are correlational (Levine, 1969). Truly experimental ones are rare. Yet it must be emphasized that there is a special kinship among deliberate efforts to improve society, the concept of cause, and the experiment. Whereas social scientists often argue that they, like the astronomers, can build a science on correlational evidence alone, and can do without the primitive concept of cause as does physics, when it comes to knowledge relevant to designing a better society, this cannot be so.

This is not to deny that the concept of cause is a logical hodgepodge, involving a number of analytic criteria of no logical relationship to each other, and of no entailed status beyond observation of past correlation. Let us accept the fact that man's deeply ingrained concept of cause is a product of biology, psychology and evolution (Lorenz, 1939; Campbell, 1974) rather than a pure analytic concept. If so, it reflects the adaptive advantage of being able to intervene in the world to deliberately change the relationship of objects. From among all the observable correlations in the environment, man and his predecessors focused upon those few which were, for him, manipulable correlations. From this emerged man's predilection for discovering "causes," rather than mere correlations. In laboratory science, this search is represented in the experiment, with its willful , deliberate intrusion into ongoing processes. Similarly for the ameliorative social scientist: Of all of the correlations observable in the social environment, we are interested in those few which represent manipulable relationships, in which by intervening and changing one variable we can affect another. No amount of passive description of the correlations in the social environment can sort out which are "causal" in this sense. To learn about the manipulability of relationships one must try out manipulation. The scientific, problem-solving, self-healing society must be an experimenting society.

Problems of the True Experiments for the Experimenting Society

Arbitrary control and disguise. True experiments involving randomization are undoubtedly the most efficient and valid. While I have been an advocate of some quasi-experimental designs where randomization is not possible, detailed consideration of specific cases again and again reinforces my belief in the superiority of true experiments. As advisors to the experimenting society we will often recommend such research designs. Yet they present special moral problems which we will have to consider. True experiments are best done where those designing and directing the study have most complete and arbitrary control over the people participating in the study, that is in total institutions such as prisons and armies (Dubos, 1970). One needs optimally to be able to randomly assign persons to experimental treatments and to enforce 100 % participation in these treatments. To avoid reactive arrangements, the participants should be unaware of the experiment, unaware that other people are deliberately being

given different treatments (Campbell, 1969b, pp. 367-377). Speaking from the point of view of humanistic existential socialism, Janousek (1970) has criticized "Reforms as Experiments" on these grounds. He argues that the whole orientation of assigning persons to treatments by randomization betokens an authoritarian paternalistic imposition, treating citizens as passive recipients rather than as co-agents directing their own society, as "subjects" in the psychologist's and monarchist's sense, as "victims" of the experiment rather than as collegial agents of the experiment.

The enforcement of assigned treatments violates the egalitarian and voluntaristic ideals of the experimenting society. The disguised experiment violates these too and in addition the openness, honesty, accountability. However much we may weight the value of scientific information in deciding upon the ethics of deceit and lack of informed consent in harmless experimentation done to test scientific theories (Cook et al., 1971; Campbell, 1969b, pp. 270–277), social experimentation for policy decisions must adhere meticulously to means idealism on these issues, and should include no research procedures that would be excluded as a part of regular governmental procedures. Participation in policy experiments is more akin to participating in democratic political decision making than to participating in the psychology laboratory. These restrictions all have costs in the validity of experimental inference. They are costs which we must live with, and try to compensate for in other ways.

Thus the task of first priority for the methodologists of the experimenting society is to design experimental arrangements that obviate these difficulties. Janousek has suggested that some system of rotation between the roles of experimenter and subject be designed. While this is not obviously feasible, it is worth more detailed consideration. The following suggestions are likewise only initial fumblings toward possible solutions.

Randomized invitations. Where there is a new treatment in short supply, and where there is on hand a relevant data collecting system covering all potential participants (e.g., Fisher, 1971; Heller, 1971; Levenson and McDill, 1966), randomized invitations can be used as a democratic means of allocating this scarce resource. Thus Stapleton (1970; Stapleton and Teitelbaum, in press) offered a random sample of arrested juveniles the services of a project lawyer. Thus one might offer psychotherapy to a random sample of students or life insurance salesmen performing poorly as judged by grade point average or volume of sales. Or those drawing unemployment insurance or showing low earnings on withholding tax records might be invited to a special training program, or income supplement program, etc. The control group are those uninvited, equally eligible but not selected in the lottery. The invitees are of course aware of the experiment. In general it is the deception of those being experimented with that is crucial, so the control population might well be left uninformed or informed only through public announcement that such a lottery was to take place, or they might be informed individually about the experiment.

Rates of acceptance for most such programs will be so low as to create special problems of analysis that need solving. In Stapleton's study, only 60% accepted. In other studies the rate will be much lower. There result as least three groups: Invited-Treated, Invited-Untreated, and Uninvited. Comparisons of the Treated with the Uninvited are biased, because the acceptance process involves a systematic selection

not paralleled in the uninvited group, and this selection alone can produce significant differences. Stapleton illustrates this selection alone can produce significant differences. Stapleton illustrates this likelihood in supplementary analyses of pre-trial data. An unbiased comparison pools all those invited, whether treated or not, and compares them with all of the uninvited. But this obviously dilutes the effects by including the invited-untreated. While a strong effect may still show up in spite of this, many genuine effects will be swamped. Here is a problem worthy of detailed attention by mathematical statisticians and others. One line of approach is to go back to the treated versus uninvited comparison, and devise methods for estimating limits on how much of this difference could be attributed to selection rather than treatment. Another approach available when the measure is continuous (e.g., income) rather than dichotomous (e.g., guilty vs. not guilty) is to focus the comparison on the upper edge of the posttest or gain scores, using either a chi-square or randomization *t*-test (Campbell and Boruch, 1970). Another approach involves approximate efforts to divide the uninvited control group along lines similar to the invitation-selection of the inviteds (Andrew Gordon, personal communications). Thus for a sample of the uninvited, information about the experiment and the fact that they were not chosen could be accompanied by an inquiry as to whether they would have accepted if they had been invited, and whether or not they would be interested if a future opportunity arose. While most *ex post facto matchings* are unusably biased, the existence of the prior randomization might make some such process usable in this setting, something on the order of post hoc blocking. Those background variables which distinguish accepters from nonaccepters in the invited group would be used to purify both the invited (accepters plus nonaccepters) and the uninvited, hopefully thus shrinking the size of the invited-nonaccepters much more than the other groups. Within this purified sample, the comparison would still pool accepters and nonaccepters, but since the latter would be a relatively smaller group, they would dilute the comparison less. (Note that the opportunism involved in selecting variables might produce a systematic bias.) These are unmatured suggestions. Hopefully they convey the challenge to the inventiveness of the methodological community.

Volunteers for experimentation. Another approach is to assemble volunteers who are told about the experiment and agree to participate no matter which treatment or control group they end up in; accepting the scientific necessity of randomization. All participants are pretested, and then randomly assigned to treatments, with or without blocking variables. (Reversing this order and randomizing before pretesting would throw light on the reactivity of the process.) This design maximizes self-conscious effects for the treatment groups and deprivation feelings on the part of the controls. (Seligman, 1969, points out what strong and unusual treatments a so-called control group may be getting. This problem needs elaboration for social experimentation .) But even with all this artificiality, such experiments will be well worth doing if we can do no better.

This design is particularly appropriate for participant-designed experiments if the methodological community can educate participant populations on the importance of this method in learning whether or not the program they have designed is effective and worth recommending to others.

Problems with randomization. In spite of age old traditions for use of lotteries in social decision making (Moore, 1957; Aubert, 1959), and its recent ritualization in draft lotteries, there is great public opposition to randomization. The methodologist for the experimenting society needs to study these objections and the values they reflect. Where they represent valid fears, he must provide methods to obviate the dangers. Where they represent misunderstandings, ways of clarification and justification must be developed. As has been mentioned above, when a new innovation is in short supply, as all are in their pilot testing and most compensatory programs are continually, then randomization is legitimately presented as the most democratic way of distributing the boon among those eligible: all have an even chance, the decision being fairly made by an unbiased lottery, with the additional moral gain from the social value of the information about program effectiveness made possible.

But it is not only the recipients who feel uneasy about randomization procedures, it is also the administrators who are asked to implement them. One way they express this objection is to say that they don't want to "play god" with other peoples lives. Strangely enough, using a random assignment procedure gives them this "playing god" feeling when their usual use of their own imperfect judgment does not. The secularization of our world view has produced a dramatic reversal of causal perceptions here. In past ages, lottery was used when people wanted to go beyond human judgment with its potential for partisan bias and let God decide instead. Those conducting the lottery felt that the decision was out of their hands and in God's. Decisions based upon fallible human judgment, rather than lottery, were instances of "playing god" in those days. Today, no longer believing in an omniscient intervening God, we see using a table of random numbers as imposing our own arbitrary rule upon others. In contrast, personal judgmental decisions based upon partially inadequate and errorful information seem fairer, less arbitrary, less dictatorial, less exalting oneself to a god who plays with others' lives. For the time being I judge this to be magical thinking, and more irrational than the ancient God-trusting lottery (although as a policy I believe one should suspect that in all such feelings lies a hidden kernel of truth).

The physical form of the lottery may be important. A public mixing and drawing of names may produce more appropriate causal perceptions on the parts of both administrators and recipients than does the solitary copying of random numbers from a table onto a name roster. Subjective experiences under such contrasting conditions are worth studying, experimentally or in role-playing simulation.

Abhorrence of the deprivation suffered by an untreated control group is much stronger when that group has been created by a randomization process than when achieved casually, by unexplicit aspects of the political bureaucratic process. This too seems an irrational objection, although the organizational capacity which achieved the randomization process could usually have delivered the treatment to the controls had it wanted to, and this is more uniformly true for randomized controls than for casual comparison groups.

The University of Illinois two years ago randomly selected arts and sciences applicants from among a pool of those highly eligible. This policy was scolded in editorials across the nation, including one in *Science* (1970, 167, 1201) surprisingly enough. In these editorials, letters to the editors, and letters in the university's file of complaints, is the raw material for a fuller study of objections, and a concrete topic for

studies of persuasion and education. Our methods book must include not only arguments and educational approaches, but also a repertoire of compromise positions, and substitute experiments or "indirect experiments" as Zeisel (1970, 1971) calls them. Perhaps a tie-breaking randomization would be allowed in the borderline cut-off range of eligibility. Perhaps a special set of opportunity admissions could be given randomly among those lacking the usual qualifications.

One reason offered against randomization is that it negates differences in eligibility, be these differences in deservingness or neediness. Popular preference in the University of Illinois case was to give the opportunity to those most deserving. In such a situation, if the decision makers will make explicit their grounds of determining eligibility, and provide an ordering of candidates on eligibility, then the powerful quasi-experimental regression-discontinuity design can be used (Campbell, 1969a). Here too the methods of implementation, of establishing in institutional practice procedures providing an acceptable quantification of the net eligibility dimension, need to be developed.

Even when randomization is accepted as policy, there are many methodological details to be worked out (Cook and Diamond, 1971), and procedural guide books to be developed. Consider, for example, a Veterans' Hospital Restoration Center, a half-way house. The managers and many of the feeder personnel recognize that this one facility cannot meet the potential demand, and that if its efficacy were to be established, tripling or quadrupling the facilities would be desirable. Potentially, the scarcity of the facility should be usable to justify random assignment of patients to the restoration center or to direct discharge into their home community, and thus through long term follow up to determine whether readmissions and unemployment were less for those receiving this special form of gentled discharge. The higher administration is sympathetic to such a hard evaluation, if feasible administrative procedures can be worked out, and if these can be sold to those decision makers involved at each of the assignment decision steps. But there are many procedural difficulties for which the alternative modes of handling each has its own statistical bias tendencies or imposed disappointments. The methodology in this area has yet to be accumulated and formalized. Some examples: If one were to assemble a large group of eager applicants for the restoration center, and then randomly reject half, one has introduced not only the experimental treatment of the restoration center, but also the treatment of rejection and disappointed expectations. For these reasons of reactive arrangements it might seem better to randomize among supervisor-designated eligibles, and then randomly invite half to participate, leaving the non-invited unaware of their lost opportunity. But this produces three awkwardly comparable treatment groups: Uninvited, invited-treated, and invited-declined (consisting in part of those who preferred release directly into the community.) The present procedure involves screening interviews by the Restoration Center staff, who reject on compatibility grounds, etc. If this feature is to be kept, randomization would ideally follow this step. At the present time, the treatment units use lower standards of health for assignment to the Restoration Center than for direct discharge. Can those rejected on one lottery be placed on a waiting list or be eligible for chances for a later assignment, or does this vitiate the randomization design? The random allocation system must take into account the feature that three to five persons must be assigned each week, keeping the beds filled, and that no back log pool of persons waiting on the regular wards for

assignment is acceptable. Randomization procedures with detailed administrative instructions must be developed before experimentation can take place in such settings.

Differential loss of contact with controls. In typical social ameliorative and educational experiments the research program must generate the posttest measures, and the contacts which obtain these, from both experimental and control groups. In this case it is typical that the rate of successful contact and cooperation is much higher with the experimentals than with the controls. For example Ikeda, Yinger, and Laycock (1970) are conducting a "true experiment" with an Upward Bound type program. For questionnaire follow-ups, 21% of cases are lost from the experimental group and 37% from the control group. In some comparisons the differential is even worse. Developing methods to control this potential source of bias represent an important challenge to technical methodologists. One approach is to degrade the quality of the experimental group to a degree comparable to that of the control group, as by employing in a mailed follow-up as out-of-date addresses for the experimental group as for the controls, even though better addresses for the experimentals are available. Similarly, by having the follow-up done by an independent agency to which the experimentals owe no gratitude, the refusal rates of those actually contacted might become as large as for the controls.

Hypotheses of selection bias in the differential loss of control cases must be specific to be plausible. Those hypotheses which produce differences in the same direction as would the treatment are most invalidating. The presence of selection bias may be demonstrable on pretest measures or pretreatment correlates, and such analyses may produce useful estimates both of direction and maximum possible magnitude of the selection effect (e.g., estimates based on the assumption of a .90 correlation between some pretreatment measure and posttest). Through such explorations one might encounter reasonable ways of estimating maximum selection effects so that large differences might be fairly attributed to the treatment. Conservative tests of significance to accompany such comparison might be developed.

Generalized surveys of the characteristics of lost cases may also be helpful, although local situations will have local laws. (In Ikeda's study, the source neighborhoods were such that children with more adequate families would be most likely to move away, and there is the probability that some of the experimental families stayed because of the program.)

In the Oberlin study, as in many others, the controls received some minimal treatment just to keep contact for follow-up and reduce differential attrition. Some experimenters call such treatments placebos, but this connotes a deception which is often not present, and should not be in policy-relevant experiments. Worth studying is the use of a graduated series of treatments not including zero treatment, but from which a zero treatment effect might reasonably be extrapolated. Again, I am not trying to offer solutions, but to illustrate the need for methodological attention.

Quasi-experimental designs. While most agree that a true experiment is best if you can get it, evaluation methodologists of today have too much faith in the power of multivariate statistical adjustments applied to correlational data and casual comparison groups. If my past advocacy of certain quasi-experimental designs has contributed to this complacency, I hope I have effectively attoned by coauthoring the most damning

criticism of the major Head Start evaluation (Campbell and Erlebacher, 1970). This study involved neither randomization nor pretests, and attempted through matching and covariance to "equate" former Head Start and non Head Start groups then in the first three grades of school, interpreting the residual differences after this adjustment process as effects of Head Start. The study neglected the fact that matching, covariance, partialing, and similar adjustment procedures regularly underadjust, since there are unique components and error in the independent variables. In the specific situation, with the quasi-controls being selected from among a generally superior population, these underadjustments are biased in the direction of making Head Start look harmful were it in fact to have no effect at all. (That covariance adjustment produces the same magnitude of regression artifacts as does matching is something that Lord, 1960, has known for some time, but has been missed by most texts, including Campbell and Stanley, 1963.)

In our judgment, the study should not have been done, nor could any study have been done at that time, except by setting up new experimental Head Start programs for that purpose. Because its impressive numbers and statistical analyses lend the prestige of science to a misleading analysis, it was in our judgment much worse than nothing. In so far as I have sampled them, however, most evaluation methodologists agree instead with the rebuttals by Cicirelli (1970) and Evans and Schiller (1970) that the study, for all its admitted imperfections was much better than nothing. More research and clarification is needed. One means of clarification would be for such studies to regularly include *a split-control adjustment demonstration*. From among the untreated, two natural population units would be selected which before adjustment differed on the achievement tests, etc., due to socioeconomic differences etc., but which did not differ on the treatment, both lacking it. Applying the covariance and matching processes to the comparison of these two control populations should produce adjusted differences of zero if the proponents of the method are correct (for they assume that had there been no Head Start effect, the adjustment procedures would have equated the Head Start and non-Head Start groups). If those, such as I, who hold to the dogma of chronic underadjustment are correct, then the split-control adjustment demonstration should dramatically confirm that fact, and render Head Start–no Head Start post hoc comparison untenable. Matching on social units such as schools is much less biased than matching on individuals. There may eventually be discovered acceptable sequences in which within matched schools, matching on subject variables is unbiased. And perhaps reasonable analogs to Lord's (1960) and Porter's (1967) reliability adjusted covariance can be developed. But in the meantime Erlebacher and I (1970, p. 224) recommend that the methodological community rule out all *ex post facto* studies, that is, studies lacking either a pretest similar in factorial structure to the posttest or random assignment to treatments.

More methodogical work is needed on the very common evaluation design using a nonrandomly selected comparison group, but with factorially similar pretests and posttests for both experimental and comparison groups. Again, matching or covariance on pretest scores is unacceptable, producing a systematic bias in the direction of underadjustment (Lord, 1960; Campbell and Erlebacher, 1970). Similarly, any control variable correlating more highly with the pretest than the posttest produces some degree of the same bias, and conversely for variables correlating higher with the posttest

(Campbell and Clayton, 1961; Campbell, 1971). But variables correlating equally with pretest and posttest should be unbiased as far as pretest-posttest comparisons are concerned. They would still underadjust in an absolute sense, but this underadjustment would show up on the pretest, and thus need not mislead the analysis. (*Ex post facto* studies lack this means of estimating the degree of underadjustment.) By assuming similarity of within group and between group processes, one might use the computed correlation of the adjustment variables with pretest and with posttest within experimental and within control to ascertain equality of correlation. (Such variables as parental income and education might meet this criterion.) But having estimates of the differential correlation and the pretest-posttest correlation should enable one to correct for the differential regression, perhaps even making possible the use of the pretest in an adjusted adjustment process (Campbell and Boruch, in preparation). In any event, this design has high priority for statistical development.

The interrupted time series design, with or without augmentation with a control series, will be one of the most important designs for the experimenting society as has been noted above. While great strides have been made in tests of significance (Box and Tiao, 1965; Glass, 1968; Glass, Tiao and Maguire, 1971; Glass, 1971; Kepka, 1971), more work is needed, particularly on using control series. At the present time, such designs are limited to data series already collected. In the measurement section to follow, we focus attention on the problem of creating new and more relevant series.

The regression-discontinuity design has also been mentioned above. Appropriate tests of significance are a problem for it also. Sween (1971) presents double-extrapolation tests for assumptions of linear, cubic and quadratic trends, and procedures for choosing among them. Misidentification can produce pseudo effects, and it would be very desirable to develop a test independent of assumptions about the nature of the curve. Appropriate tests are needed also for the "fuzzy" condition, where a sharp cutting point has not been used (Campbell, 1969a, pp. 423–424).

A number of ameliorative programs, overwhelmed by the vastness of the problems of urban poverty and health, but desiring to see what effect a massive remedial effort could bring, have set up treatment areas with sharp boundaries. There are at least two such in Chicago, one a 16 block area, another a square mile area. Reputedly they have been quite strict in limiting their extra services and efforts to within such boundaries. Transferring the logic of the Interrupted Time-Series or Regression Discontinuity designs to the spatial dimension, one should be able to use the abruptness and artibrariness of the treatment boundary for quasi-experimental analysis. The prospects are made more difficult by the fact that, reputedly, the 16 block experiment is averse to being evaluated, in spite of the very intense concentration of effort, including a staff of several hundred persons. We can envisage block-by-block plottings of residential stability, eviction proceedings, landlord repair investments, refusal to be interviewed, interview contents, reactions to experimentally introduced salesmen and petition circulators, home addresses of offenders in police records, etc. It may well turn out that the staff has found it humanly impossible to be perfectly strict about limiting services to those with in the area. Even in this case, there are probably analysis techniques that would pick up a strong effect.

New types of quasi-experimental designs should be invented, or rediscovered and generalized from past studies of specific situations for which ingenious analyses have

been devised. Each of these should be carefully scrutinized to make sure that it avoids the recurrent problem of packaging underadjusted selection differences as though they were treatment effects. Failure to recognize this problem has been the major error of recent evaluation research.

Measures of the Quality of Life for Social Experimentation

The Experimenting Society will use data banks and social indicators to evaluate the quality of life. Within this currently much discussed realm of possibilities, the experimental orientation provides selective emphasis: Rather than focusing on an overall composite goodness-of-life index, the focus is upon specific data series relevant to evaluating specific social reforms. Even for a specific social experiment, multiple indicators, all recognized as partially imperfect and only partially relevant, are advocated. The likelihood that indicators will change in validity once they become foci of social decisionmaking is emphasized.

Archival indicators. At their best, specific government records on life, health, taxes, unemployment, crime, and the like, will be important parts of the score keeping of the Experimenting Society. At their best, they will provide fine-grained time series sensitive to reform efforts, "hard data" less subject to the interview biases of fear, courtesy, and expectation which are apt to provide pseudo-successes in survey-research evaluations. But many archives are incompetently kept for indicator purposes, or are corruptible by the administrative bookkeepers whose administrations are being evaluated, or are responsive to reform by shifts in their irrelevant rather than their relevant components (Campbell, 1969a, pp. 414–417). Webb et al. (1966) have provided a general survey of such methods and the criteria for their evaluation, but more experience is needed in social-experimental situations.

Dead records that no one ever expected to be spotlighted may have quite different quality from those under the glare of public decision-making scrutiny. The records may be better kept, or more defensively and biasedly kept. We should extend our preoccupation with validity into this situation. We should inform ourselves of the historical impact of government audit laws upon the bookkeeping practices of banks, and on the history of election reforms. We should even experiment. For example, using the authority of freedom of information laws, we should regularly visit half of the government agencies with regional headquarters in Chicago, copying relevant records. One year later, quality and accessibility of such records should be compared with those of the same agencies never before visited, and with the office of the same agencies in other regional headquarters. A comparable experiment could be done with police precinct stations, coroner's offices, hospitals, etc. While such a study may seem foolish, it is equally foolish to assume a perfectly conscientious bureaucracy errorlessly and promptly recording "hard" data, which has automatically classified itself. Even death itself is biasedly reportable, by the time it gets classified as suicide, murder, manslaughter, medical and natural causes.

While there may, temporarily and partially at least, be free access to non personal aspects of government records, there are, correctly, legal restraints on giving one citizen (even though a methodologist for the experimenting society) personal information

about another citizen. With the collaboration of legal specialists, we must inform ourselves about legal technicalities and the legally acceptable safeguards. We must be prepared to educate the record keepers of the agencies we contact about these laws. We must collaborate in preparing revisions for laws which needlessly prevent social reality-testing. We must be continuously alert to specific organizational inventions to solve such problems (e.g., Schwartz and Orleans, 1967; Sawyer and Schechter, 1968; Boruch, 1971; Fischer, 1971).

As methodologists for the experimenting society, we need to develop the organizational, sociological and political theory of bias-free record keeping, and make recommendations as to the kind of bureaucratic structures, the kinds of separations of authority, the mutual check and duplicated record procedures, which make it possible for a nation to publish votes, census figures, unemployment indices, cost of living indices, and the like, which defeat or weaken the government in power. We need historical studies of corruption and influence attempts, not only in this country but also in countries where the record keeping has been both more corrupted and less. We need to be alert to sociopolitical inventions which will maximize such objectivity. For example, it has been suggested (Campbell, 1969a, pp. 415–416) that there be independent data collection agencies monitoring social experiments.

Voluntary verbal measures. To measure the goodness of life, it seems necessary to go beyond the commonly available, automatically archived, counters of life's outcomes (e.g., income, employment status, vital statistics) and make efforts to assess both general happiness and specific enjoyments, satisfactions with both the work day an recreational aspects of life. In thus moving into the area of personal utilities, we must be careful to include the altruistic as well as the self-centered, measuring not only pride in own achievement but also pride in family, community, and national achievement. We must attempt to measure satisfaction with family and neighbor interactions, as well as consumer satisfactions. Thus we must avoid falling into an assumption that personal achievement is the only or most relevant yardstick. Avoiding the paternalism of planning people's lives in a way good for them whether they like it or not, we should assess preferences, satisfactions, and resentments with regard to the specific program and problem area, both in the respondent's own case and at the general policy level. We should make contact with past and current efforts to develop standardized scales or voting forms suitable for a wide range of ameliorative efforts.

We are the benefactors of forty years of insightful research on validity problems in verbal voluntary self-description measures administered by way of interview or questionnaire. We have hundreds of studies of interviewer bias, response sets, faking, effect of anonymity, etc. To these studies need to be added research on the effect of telling respondents that their decisions are going to contribute to governmental evaluation of ameliorative programs, both past and anticipated. Will such awareness lead to less or more complaint? Should such respondents be anonymous, as voters, or named, as select jurors? Testing the new interview measures of the quality of life such as the Survey Research Center is developing should include experimental variations of such contexts of responding. While pleasure or pain, weal or woe, may be phenomenally absolute, they are demonstrably context dependent (Brickman and Campbell,

1971). Two principle contexts or implicit comparison bases may be most important: comparison with one's own past status, and comparison with others' present status. We can elicit and emphasize one or another of these comparative contexts in a preliminary part of an interview, and determine the effect upon the well-being subsequently reported. (Particularly discrepant set effects might be expected for the urban black community.)

Using Welfare Recipients' Judgments in Evaluating Changes in Welfare Delivery Systems

Problems of measurement in the experimenting society are illustrated in more detail by considering possible ways of collecting recipients' judgments to evaluate a major change in a state welfare program. What follows draws from a report (Gordon and Campbell, 1970–1971) prepared for the Illinois Institute of Social Policy.

One of the first and most important commitments of the Experimental Service Program is that reports from the systems customers and potential customers be given a major role in judging program effectiveness. Furthermore, such reporting is seen not only as a means for measuring the effects of new programs, but also as an effective and important part of the new program itself.

This point is important enough to take space to restate the program's arguments: The natural inclination of the staff in any public office is to adjust to the demands of their fellow staff members. Clients, customers, the public, interfere as strangers with the ingroup social system that develops among the staff. There arises an unconscious tendency to define the agency's mission narrowly so that there are as few clients as possible. In case of doubt, an applicant is sent to some other agency, or is told there is no agency for his needs. The customers accepted are given the type of service most convenient for the staff member, not most needed, etc.

These same pressures exits in business enterprises. Many of us have felt like unwanted intruders in some fancy department stores. But in the business world there is a powerful feed-back system from us customers. Stores that have no customers go out of business. No such feed-back exists at the present time in most government service agencies. It is the program's goal to create such feed-back, and to achieve a situation in which agencies compete for satisfied customers, and go out into the street to dig up new business, people who are entitled to services they may not know exist.

The proposed mechanisms for doing this are (1) Units-of-service-delivered accounting; asking all State service agencies to report each month who they have served an how many units of service they have delivered. (2) Putting government service agencies in to direct competition with private agencies with whom the government will contract for delivery of specified amounts of service. (3) Initiating customer satisfaction reports and unsatisfied need statements from needy potential customers. (4) Using all three of these productivity criteria as a basis for decisions on budgeting, agency expansion, reorganization, elimination, etc.

At the present time, governmental decision making on the adequacy of programs is based on the judgment of those employed to run the program, or those still higher up and still more removed from the actual situation. There is no machinery available at

present to let those served give their judgments. We might expect therefore that the plans of the experimental service program to provide such a voice would be welcomed by the community. Unfortunately, many of the means by which these judgments might be obtained are similar to search procedures now rejected by the community. We must understand these objections and attempt to correct the evils to which they point.

The people of the community, and no doubt people in other urban communities, are fed up with being researched by public opinion polls or questionnaires. Such research is seen as exploitative, as providing jobs for white middle-class researchers using money that would be better spent on helping poor folks. Such research is also seen as helping the researcher through the articles and theses that get written. In contrast, surveys are only a burden to the community respondents. As to "making the community's needs known" this has been done again and again and nothing helpful ever come of it. Rarely, if ever, are the results reported back to the community.

Furthermore, community public opinion surveys, like the case-work interviews, usually contain humiliating invasions of privacy. The focus seems always to be on what is wrong with the respondent rather than what is wrong with government services. Even in terms of the respondent's needs, it is assumed that he doesn't know his own needs, a specialist must pry them out and tell him.

While one cannot hope to meet all of these objections, one must keep them in the forefront and only counter them when there are strong reasons, understandable by the community. As a start, here are four statements of intent:

1. The community opinion surveys should focus upon agency performance and the quality of service available in the community rather than asking personal questions. They should provide a convenient means of expressing complaints, and of mobilizing them so that they get heard in government headquarters.

2. In order to be of service to the people, the surveys should cover topics which the people want to sound off on, even when these involve problems not under the control of the State.

3. The people should have complete and immediate access to all the results. They should not be exploited by being excluded from information created by their own cooperation. The local communities should be allowed to follow shifts in their own local opinion just as, at the National level, we follow Gallup Poll results and election returns. The State should cooperate in this regard with all groups claiming to represent the community or segments of it. It is to be anticipated that the government will thus be providing strong ammunition that will be used in pressing community claims against the government. This is not to be avoided, but instead will be the clearest evidence that the opinion research activity has not been exploitative, but has instead provided a real service to the community.

4. The people responding will be told explicitly whom they are talking to. The sponsorship of any opinion interview or questionnaire will not be disguised or attributed merely to some survey agency such as NORC, Gallup, or Harris Poll. It is commonly believed in both commercial and political polling that the respondents

knowing who the sponsor is will bias the results (presumably in the direction of being more favorable to the sponsor than is accurate). If so, this is a bias we must live with if we are truly to level with the public. Every statement from one person to another occurs in a conversation, and is inadequately interpretable without the context of that conversation. The public opinion survey respondent is entrapped and betrayed if he does not know what conversation he is in or whom he is talking to.

These recommendations do not meet all of the objections. The research activity will continue to be dominated at the administrative level by middle class persons, although with a greatly improved representation of Blacks. Because it must cover control areas as well as experimental areas, because it must eventually go statewide, and to avoid challenges of bias, the information collection activity cannot become a monopoly of any one community nor even give preferential employment to residents of any one area, although staffing procedures should certainly give high priority to employing those now unemployed and under-employed, and to creating jobs in areas where jobs are scarce. As to the challenge that adequate information on community needs is already at hand, we must attempt to educate the community to the fact that welfare funds can be spent in quite different ways, and that in order to choose the best way we need to repeat the same questions again and again.

How should the surveys of opinion be conducted? Somehow the judgments and preferences of the people should get assembled so as to become a part of government decision-making processes. Something more official than newspaper reports of Gallup Poll surveys would seem desirable. The more important the role of such community opinion becomes, the more it must be able to stand up to challenges and the more it must have safeguards against inadvertent or deliberate bias.

As a background for recommending specific procedures, it seems well to lay out the range of possible techniques, and to inventory the recurrent problems or sources of bias. (In what follows, a list of issues and problems, A, B, C, D, etc., are interwoven with a list of ways of conducting surveys. These issues and problems can be regarded as extensions of the Webb et al. (1966) list of 13 validity issues for measurements in general.)

Election-type procedures. A survey of community opinion, judgments, and complaints could be set up as a ballot, on which people "voted" in a neighborhood (precinct) polling place. Such a procedure would be (A.) very *expensive*, but not necessarily more expensive than door-to-door interviewing. It would be (B.) *representative-in-opportunity* to participate, but (C.) *unrepresentative-in-participation*, in that the indifferent, incompetent, too busy, and totally alienated would be underrepresented. It might be more biased in this regard than some public opinion surveys. If there were (D.) *stability-in-representation bias*, then shifts in votes from time to time would be interpretable as a shift in public opinion. It would be possible for a shift to occur due just to a shift in who votes, without a shift in opinion, although this is not likely, and the number of voters would usually in itself be a symptom of community discontent. However, a record of (E.) *campaigns to get out the vote* would be essential in interpretation, as would also campaigns to get people to vote in specific ways.

The specific voting procedures now in use are the result of a long historical process in which features were added to correct previous problems. These problems occurred because voting was important in governmental decision making. If the community judgments we collect become important in governmental decision making, as we intend they should, similar problems will emerge.

(F.) *Anonymity* is insisted upon in modern voting procedures, although absent in the town-meeting democracy of our past. It is instituted to avoid (G.) *fear-of-retaliation* which would lead people to vote as others wanted them to rather than as they actually felt. These others feared usually, but not always, the government in power. To prevent (H.) *people filling out other* people's ballots, the voting is done alone and in the polling place (except for those few qualifying for assistance in voting, and this assistance must be given by a family member, or by two judges of opposing parties. For such assistance, special affidavits are filed.) Anonymity creates problems as well as solves them. It makes (I.) *ballot-box-stuffing* easier, by persons who vote twice, by persons who come in from other areas to vote, etc. To curb these biases there are prepared (J.) *registration lists of eligible voters*, and (K.) *signed records of who did vote*. To prevent ballot-box stuffing by *substituting false ballots* for the true ones, (M.) *poll watchers* from competing political parties are required. To make possible meaningful challenges to the validity of the process, there is (N.) *public accessibility of the records*, public verifiability for a period of time long enough to allow for challenge, perhaps a year, and the possibility of recounts. These procedures are probably adequate, if vigorously implemented, to insure honest elections. That we do not always have such is due to lack of vigilance in removing dead and gone persons from the registration lists, lack of verification of the identity of the voters, poll watchers all of whom are in actuality from the same party or political machine, etc. It probably would not help to add still further precautions. Much less adequately handled in our system is genuine (O.) *public representation in the design of the ballot*. The public power to decide is limited by the alternatives that get on the ballot to be voted on. These, whether candidates or issues, are often decided upon by quite unrepresentative procedures and omit what would have been the most popular public choice.

We have discussed voting primarily to raise issues for other procedures, rather than as a procedure to be seriously considered. But perhaps it should be. While we must not regard *any* procedure as ideal, and must be on the alert for misleading results from each, it might be judged that precinct voting on a "Community Problems Inventory" would be the best procedure if it were not for cost. These costs would be greatly reduced if such a survey could be added as an additional ballot at elections scheduled for other purposes. This would slow up such voting, and require additional booths and space just for that purpose, but would probably not involve more than a 20% increase in cost. As in other voting, sample ballots in local newspaper, etc. could ready the voter for the polling place decisions. Much of the ballot would be highly structured questions that could be electronically tallied, an adequate procedure as long as the ballots remained available for more traditional verifications. (Space for written comments and suggestions for future inventories would also be included, but probably only analyzed for samples, and impressionistically.) Even though we are not seriously considering this procedure, we should at least find out if it is legally and financially feasible.

Residential interview samples. By this heading we mean to designate the now standard public opinion survey procedures, where a representative (often random) sample of the community is selected to be interviewed in their homes. For many reasons this is often the social scientists' first recommendation, and was recommended in the form of a Base Line Study for the Experimental Service Program. This was rejected by the Community Advisory Board for several reasons, at least one of which is intrinsic to the method. One objection was that personal questions were to be asked. This could have been cured by eliminating those questions and retaining the many excellent questions calling for judgments of community agencies. Another important objection we can label (P.) *intrusiveness.* Interviewers intrude like bill collectors, police, and public welfare workers in a poor neighborhood. Their very presence is a burden, an invasion of privacy no matter what questions are asked. On these grounds, door-to-door interviewing is to be avoided if other adequate means are available and unless the values to be achieved are real and can be made clear to the community.

Interviewing is also expensive. With typical rates for interviewing and analyzing results running near $100 per respondent, with 1,000 respondents being a reasonable number to provide sampling stability, with interviews required in equal numbers for control areas, and with repeat surveys needed at least each year, the annual budget for the program could run $200,000. (We must bear in mind that public opinion surveying has achieved enough experience to do realistic budgeting, unlike most of the other procedures we suggest. Also, much of the $100 per interview cost comes from coding free responses. Were the interview to use highly structured questions amenable to mechanical tallying, as do many of our other suggestions, costs could be greatly reduced, perhaps as low as $15.00 per interview.) Representative samples can be drawn up, but interviewers will find 25% not at home. Call backs can reduce this but are expensive. Another 25% may be expected to refuse. Nonetheless, overall, opinion surveys are likely to be more representative than any other procedure, since representativeness is increased through preventing procrastination, and through not requiring the effort of reading and writing. We meet here, however, the new problem of (Q.) *perceived representativeness.* The public is apt to distrust a survey in which they were not asked, are apt to not believe that randomization and stratification can provide representativeness, or have actually been done. That is, it may be assumed that only those known to be favorable were asked. Thus procedures which give everyone a chance to respond may be perceived as more validly representative than those which use a small proportion of the community randomly selected (even though a statistician might decide the opposite if there were a small turnout on the all-invited procedure). Anonymity is not provided in that the interviewer knows or can know who the respondent is, and name, address, and telephone numbers are often requested to make interview verification easier. Depending on whom the interviewer is, fear of retaliation may operate. (The interviewers should be powerless outsiders rather than local officials with power over the respondents.)

Interviewers can easily guide respondents to preferred answers. Ballot-box-stuffing in the form of false interview records can easily be done, especially if respondents' names are not recorded, but also when they are. Polling agencies do abbreviated repeat interviews on a partial spot-check basis to check on whether some

interviewing was actually done. This is to check against a lazy interviewer inventing interviews. It does not check adequately on a systematic biasing of responses in the recording process. Human memory and the vagueness of conversations make this almost impossible to check. If immediately after a bonafide interview, the interviewer were to read back his summary of respondent opinions on free-response items, the average respondent would probably feel the recording inaccurate on one third of the items. Weeks later the identification of deliberate misquoting would be hopeless. (If door-to-door interviewing employed in part structured ballot, one verification device would be to leave a copy of this with the respondent, which, if he kept it, could be used in the repeat interview spot checks.) To avoid falsification at the survey administration level, the sample verification should be done by some truly independent agency. It should also include verification of the random sampling procedures. Under the requirement of public verifiability, the names of respondents should probably be available for independent verification efforts by protest groups challenging the study. (In the New Jersey Negative Income Tax study not even a congressional committee or the General Accounting Office has been allowed to do such verification, and for good reasons.) Another approach to verification would be to have different agencies each do half of the surveying, with agreement in results being the verification. (To some extent, the opinion polling agencies keep each other honest by publishing competing results on the same topics.)

Mailed questionnaires are an important means, and one we recommend in spite of the many weaknesses. Its main weakness is low and selective return rate. Election type voting and door-to-door interviewing probably get returns from some 50% or more of those invited. Mailed questionnaires occasionally got better return rates than that, but figures as low as 10% also occur and 25% might be a better overall estimate. At this return rate, and for electronic scoring of structured questions, costs as low as $1.50 per respondent might be estimated. Such low costs would make possible (R.) *high frequency of measurement*, providing the "fine-grained time series" so desirable for the time series designs, particularly for a program which is going to be introducing different innovations at different times of the year. Thus four to twelve waves of measurement per year would be possible. This might be combined with the goal of perceived representativeness so that once per year each person was questioned, the total address list being randomly divided into twelve sub-lists, one for each month. Anonymity is readily achieved if wanted, although it makes ballot-box stuffing easier, by community groups, through collecting and filling out unused ballots or by printing their own, and by the surveying agency, and makes verifiability impossible. Even if something like polling-place voting records could be achieved by having each person also return a signed post-card stating that he had sent in his anonymous survey, ballot substitution in the data analysis process and people filling out other people's ballots could easily occur.

On this latter point, consider two extremes. If the person to whom the ballot is mailed asks some neighbor to help him fill it out, or even turns it over to someone to fill out for him, the processes of democratic representation are still probably well served. The opinion reported is a genuine community opinion. But with three-quarters

of the ballots mailed out lying around unused, it would be easy for an alert community organization to collect them and fill them out in accordance with the organization program. This would still represent community opinion of a few leaders. Probably the tendency to allow this to be done would be much less with signed ballots. But how different would this be in practice from a community organization preparing a sample questionnaire with detailed recommendations on how to respond? This latter would seem to be a perfectly proper electioneering procedure. It could play havoc with a time series, particularly if the persons interpreting the results were unaware of the campaign. It would be hoped that if a meaningful channel for community opinion were to be established, that community organizations would see its value and work to preserve its usefulness by preserving its credibility.

(S.) *Comprehensibility* is a problem in any such survey. Bureaucratic agencies and social science academics generate forms and questionnaires that are incomprehensible to the bulk of U.S. citizens, as the Harris study has shown (1970). Familiarity with questionnaires and tests is a middle-class trait. Questionnaires use an arbitrary short hand, which makes for efficient responding once understood, but is baffling when first encountered. While more readable and clear questionnaires can be achieved, this tends to make them longer and more formidable for this reason. Thus comprehensibility is always a disadvantage for questionnaires in contrast to individual interview procedures.

Questionnaires can be "structured," with fixed multiple-choice answers, or "free-response," requiring written answers, and introducing the problem of (T.) *writing fluency*. It is hard for the middle-class person whose job requires daily practice in writing, to realize what an obstacle this can be for a person who may have occasion to write only once or so a month. Infrequency of use of writing also affects the size of the script, and middle-class questionnaire constructors regularly make the blanks too small for those who do not write often.

The problem of fluency of writing argues for use of structured rather than free response questions, where possible. But there are many contents where the structured format is very awkward. As discussed in more detail in the full report, inquiring as to what problems a citizen has had, and what agencies he has gone to for help, is utterly unfeasible in a structured format. On such topics, individual interviews are much to be preferred.

What mailing lists are to be used? For some purposes, these could be names of present recipients of state services. But to cover those who should be recipients and are not now, it must be broader than this. For the best sampling procedures, a list of residence addresses (with or without names) would be needed, mail could then be addressed to "Occupant, Apartment 2b, 6565 South Woodlawn, Chicago 60637." (The fact that one apartment may house several unrelated families is a problem.) Sampling by blocks, instead of persons, and with cooperation from postmen, one might use a non-specific "Postal Patron, Local," without a specific street address.

Other problems. Going through these several possible ways of getting community opinion have raised many problems, but not all. As soon as people came to believe that their opinions are influencing the distribution of funds with more funds going to

more needy areas, we run the risk of (U.) *exaggerated complaint*, in which people portray their situation as bad just to be sure they get their share or more of government attention. The direction of this bias might be expected to be opposite to fear of retaliation on most topics. The recognition of the usefulness of exaggerated complaint would probably be greatest in those communities with active local protest organizations, and this could produce spurious regional differences in apparent need. No easy way of controlling this bias seems available. Conditions of anonymity might make respondents more irresponsible in this. Campaigns by local organizations to encourage this type of response might be looked for. Familiarity with the survey procedures might reduce fear of retaliation and increase sophistication, producing results which made it look like things were getting worse even while they were getting better. Basically, all community opinion survey procedures must be used with caution in comparing regional needs. They will be much more interpretable as a basis for deciding which local problem area deserves most attention, and once set up and familiar, with comparing reactions before and after an experimental treatment.

Another danger that has emerged in our discussions with the experimental service program is the potential role of public problem surveys in creating (V.) *frustration due to raised expectations*. Thus if we start surveying dissatisfaction on a given topic too long before the program is going to be able to do anything about it, we may create false expectations of immediate improvement and anger when nothing happens. This problem would be particularly bad in the control areas where no special program improvements were to be taking place. In part this is a matter of how introduction and questions are worded. One of the special goals in pretesting instruments should be to check on this. Wordings that are misleading in the expectations they generate should be avoided. On the other hand, if the asking of the questions mobilizes public demand that something be done about problems, this is an outcome intrinsic to the goal of providing a meaningful and useful voice for the community in governmental decision making.

The policy of giving respondents full access to tabulations made from their answers will undoubtedly produce some changes in the answers given at first. We can name this problem after its most likely form, (W.) *the bandwagon effect*. Judging from college classroom studies, telling people how their group as a whole has voted leads some persons to change their vote so as to agree with the majority, thus increasing the size of the majority on a second vote. From a simple-minded application of conformity principles, a larger majority should be more impressive than a smaller one, and thus one might expect repeated voting with feedback to push all majority positions up toward 100%. This implication has never been checked, as far as we know, and almost certainly is wrong as far as one can judge from practical experience. Making votes public should have this effect, even though attenuated by the two to four year delays between votes. Perhaps the one-party areas are kept so through the help of this mechanism, but for most districts such trends are non-existent. Since the advent some 40 years ago of published public opinion polls predicting election outcomes, there has been fear of such an effect, but no clear illustrations are available, and trends in the opposite direction occur as frequently. If there are such effects, they probably diminish after the first few waves. Since our general orientation is toward relative measurement, rather than absolute, once

the process has settled down, repeated surveys accompanied by prompt feedback should produce interpretable trends for the evaluation of other impacts. (By delaying feedback for a random half of the reporting units, an experiment on the effect could be built into the early stages of the introduction of a measurement program.)

When combined with a likely fear of retaliation for criticism, the bandwagon effect might be of a selective nature. Seeing from the results that others have been brave enough to complain with apparent impunity might increase the number of respondents willing to express complaints. Complaint frequencies less than majorities should have some of this effect. No parallel process would occur for expressions of satisfaction, since these would have been under no inhibition. A related effect might result from seeing the results from neighboring regions, stimulating competitive complaining.

Our main concern must be those methodological problems which will cause shifts in measures which might be misinterpreted as program effects. The repeated answering of the same questions may well produce (X.) *boredom due to repetition.* This would most likely appear as a reduction in the number of respondents, and therefore, on some instruments, a reduction in the number of complaints, on others a reduction in the number of reports of satisfactory service.

More General Aspects of Scientific Method

In discussing how to increase the validity of government sponsored research, it must not be forgotten that the objectivity of scientific theories comes in large part from social processes transcending individual experiments. Among these processes, efforts to replicate and vigorous criticism of experiments and theories, play very important roles (Popper, 1963; Polanyi, 1966, 1967). An advocacy model is appropriate for science as well as law (Levine, 1970; Shaver and Staines, 1971). Because of costs and political pressures, it will be very difficult to achieve these in government supported policy research. We must therefore seek norms and organizational arrangements which maximize replication and criticism. The following are preliminary suggestions along this line.

Simultaneous replication. Large social experiments and evaluations (such as the New Jersey Negative Income Tax experiment and the Westinghouse-Ohio University Head Start evaluation) might well in the future be divided between two contractors, who would each carry out studies of half magnitude. These simultaneous replications would have the same problem statement, but would make independent decisions as to implementation at all points where the policy goals and scientific considerations left open alternatives. This procedure would increase costs through the duplication of planning costs and through higher per unit administrative costs. The total budget therefore might run 25% to 50% higher. But the greater degree of certainty when the studies were in agreement, and the greater likelihood of avoiding errorful or deliberately biased reporting, would be worth this extra cost.

Critical reanalysis of data. While offering no protection against biased basic data, nor against withheld data, this procedure is already being employed with the effect

of challenging original research interpretations, as in both the Head Start evaluation and the Negative Income Tax experiment.

It is suggested that the Freedom of Information Act be modified and interpreted as authorizing access to basic data by an interested person, both in the case of research conducted directly by a government agency and for research delegated to private agencies. Since in most such studies, only a portion of the data are utilized, the selection of which offers opportunity for bias, access should be given to all collected data, except for the names of specific respondents. Preparation of all data for public access should be included as an explicit contract requirement. Possibly such access should be delayed until the contractor and funding agency have prepared a report, or until one year from the time of data collection, whichever comes first. (The latter requirement seems needed for those studies never reported.) The right to publish such reanalyses, with due acknowledgements, should also be made explicit. (The Office of Economic Opportunity already has such policies.)

Encouragement of internal criticism. Establishing norms which foster minority reports and alternative analyses by members of the research staff would greatly increase the freedom from bias and honesty of policy relevant research. The Freedom of Information Act should be explicitly interpreted or modified so as to give each research assistant on a project the right, eventually, to publish alternative interpretations. Where an assistant believes that relevant data are being neglected or misinterpreted it should be a duty rather than a disloyalty to report his own best interpretation. It might also be feasible to require that each professional research associate be invited to insert a minority statement as a part of the basic project report.

While official permission for dissenting reports would be of some help, it would not solve the jeopardy of loss of job or future contracts. A consideration of the conditions that have in the past resulted in such reports suggests some general principles. Researchers whose quality and publications are such that they can get university appointments can in this way acquire employment permitting such report writing. This suggests two general principles. The conditions for objectivity are greater when the research is entrusted to persons with university-level qualifications, whether the research be done within the government agency or by private contracting agency. Research staff on leave from university positions perhaps have the greatest freedom to be honest. Universities, centers, and foundations would increase the amount of such freedom by establishing a number of rotating professorships to be filled for one or two year terms by persons with governmental research experience.

Within government itself, some such procedures might be instituted. Traditionally, systems of checks and balances, of deliberately designed mutual monitoring, have been used to increase the efficiency and honesty of government. Recently this has been illustrated in that form of agency organization which places program evaluation responsibilities in a separate division from program implementation (Williams and Evans, 1969). It is also illustrated in the reanalyses of data on the negative income tax experiment by the General Accounting Office. Applying this general orientation to the problem at hand, two institutional programs can be suggested. In the Bureau of the Budget or the General Accounting Office, there might be created an *Institute for Critical Reanalysis*, in which 20 to 50 salaried positions would be open to transferees

from other government agencies, with the expected activity being dissenting reports. Another device might be one-year termination salaries for professionally qualified research staffs. Under such a system, a researcher who was fired would be guaranteed one year's additional salary. Perhaps he could also have the right to resign and receive one year's termination pay if he had been an employee for a certain number of years. University tenure to some staff members in policy research institutes would also help, especially if it were accompanied by extended leaves of absence for participating more directly in social experimentation.

These inadequately developed suggestions obviously need much more careful study. However, if they seem extreme and expensive this properly communicates my feeling of the very great importance of the problems of objectivity, and the very great value to the nation of any mechanisms which would make objectivity more likely.

Science of the politics of social experimentation. Among the methodologists for the experimenting society must be political scientists who know the political process well, and who can invent political solutions to the problems that block the effective and honest use of social experimentation. The problem of the overadvocacy trap can be used as an example. In "Reforms as Experiments" (p. 410) I made the following suggestion:

> One simple shift in political posture which would reduce the problem is the shift from the advocacy of a specific reform to the advocacy of the seriousness of the problem, and hence to the advocacy of persistence in alternative reform efforts should the first one fail. The political stance would become: 'This is a serious problem. We propose to initiate Policy A on an experimental basis. If after five years there has been no significant improvement, we will shift to Policy B.' By making explicit that a given problem solution was only one of several that the administrator or party could in good conscience advocate, and by having ready a plausible alternative, the administrator could afford honest evaluation of outcomes. Negative results, a failure of the first program, would not jeopardize his job for his job would be to keep after the problem until something was found that worked.
>
> Coupled with this should be a general moratorium on ad hominum evaluative research, that is, on research designed to evaluate specific administrators rather than alternative policies. If we worry about the invasion-of-privacy problem in the data banks and social indicators of the future, the touchiest point is the privacy of administrators. If we threaten this, the measurement system will surely be sabotaged in the innumerable ways possible. While this may sound unduly pessimistic, the recurrent anecdotes of administrators attempting to squelch unwanted research findings convince me of its accuracy. But we should be able to evaluate those alternative policies that a given administrator has the option of implementing.

It is now apparent that this is in this form an inadequate solution. What wisdom it has lies more in pointing to the problem than in offering a practical cure. Given that advocates of programs are always in competition with other persons advocating differing solutions, the posture of tentativeness seems doomed to failure, even if the understanding of these matters were to be greatly increased in the publics and administrative hierarchies that choose among the competitors. And if adopted as a general posture, it too could be used by a political machine to explain away a recurrent ineffectiveness due in fact to corruption. Shaver and Staines (1971) fear that tentativeness on the part of the reform leader is incompatible with the determination that is required to implement a new program.

Here is a problem area in which research could now be done. One could start with actual advocacies of program innovation within a bureaucracy, within a legislature, or in appeals to voters. These could be examined in detail to see if indeed they involved the risk of the overadvocacy trap and how the advocacy process might have been modified to make experimental evaluation more acceptable. Role playing simulations and interviews with the political actors on the feasibility of suggested modifications would be useful. For the Shaver and Staines objection, the literature of double-blind studies in pharmaceutical research may be relevant. The major reason for keeping the doctor in the dark as to whether or not he was giving the patient a placebo has been fear of the doctor's bias as a recorder of the patient's health. But there must be studies which tease out the genuine therapeutic effect on the patient of the doctor's faith in the medicine he prescribes. These would be models for conceptualizing the importance of the administrator's confidence in the new program in fulfilling his role as leader. The whole issue of faith, zeal, and convincing leadership is of course also relevant to the greater effectiveness when programs are in demonstration stage or are newly implemented, in contrast with long established implementation.

For the problem of fear of evaluation, the advice against ad hominum research still seems very wise, and useful independently of solutions to the overadvocacy trap. To the bureaucratic record keeping of the schools should be added a "student's annual report" on the goodness of the school program as the pupil sees it. These should avoid evaluating teachers, and focus on program alternatives which any teacher could implement. Similarly, the "teacher's annual report for program evaluation" should avoid evaluating either pupils or principals. In the social work area, the recipient's program evaluations should avoid evaluating their social workers, and the social worker's annual report for the program evaluation should avoid evaluating either their supervisors or their clients. This abstinence on ad hominum evaluation is a political necessity if such procedures are to be adopted at all. But is also justified on other grounds. The expensive machinery of experimental evaluation should only be used where the findings can be generalized to other settings and can add to our knowledge of how to manage a good society. It should not be wasted on such a petty topic as the quality of a specific person. (While I speak dogmatically on this, note that I do not speak from adequate experience. This should again be read as illustrating but one of the many problems for which an applied science of politics needs to be developed.)

The conflict between action and measurement. Almost universally in social ameliorative programs there is a conflict between the action personnel or therapists, and the research staff. This should be regarded both as a practical problem and as philosophy of science issue. It may be that much of the hostility comes from the threat felt by the therapists. But much of it legitimately points to the rearrangements in the action program required to make an experimental evaluation possible. Most of these changes are appropriately regarded as distherapeutic. There may be a fundamental social science indeterminacy issue here. In a broader framework, the problem becomes one of the compatibility between psychological health and continuous measurement.

Qualitative versus quantitative methods. The issue of "scientistic" versus "humanistic" methods is still a lively metascientific issue for the social sciences, and

one of crucial importance for the experimenting society. While the recommendations of the present paper are clearly on the scientific, quantitative side, this is not intended as denying values of the qualitative approaches or of their criticisms of pretentions to science. In any event, the issues should be kept open and studied for their particular relevance for social experimentation.

My own view is that scientific knowing is not an alternative to ordinary perception and common sense knowing, but instead always depends upon it, even though at best it goes beyond it. In physics, the use of laboratory equipment, and the communication of what was done all depend upon a common body of understandings about ordinary objects. Where physics contradicts common sense at one point, it does so only by trusting common sense at one hundred other points.

This is even truer for the social sciences. Our ability to interpret any vital statistic or percentage tally of questionnaire answers depends up on our unquantified common sense understanding of the data collection processes. Occasionally we may be able to use experimentation and quantification to go beyond common sense knowing, but this is only because we build upon it, not because we substitute for it.

Framework and paradigm. As Kuhn (1962) has pointed out to our generation, the scientific method works within an encompassing framework or paradigm, and changes of framework are based on grounds not included in the scientific method of the earlier paradigm. This shift of rules he calls a scientific revolution. Changes according to the rules of scientific method of any one paradigm period he calls normal science. He has argued for abrupt revolutionary discontinuity rather than the gradualism and continuity usually assumed for science. Needless to say he has generated much vigorous criticism (e.g., Lakatos and Musgrave, 1970) and I am myself quite ambivalent on the issue (Campbell, 1969c).

At this point, the philosophy and history of science, plus the sociology and psychology of science, have great relevance for the methodology of the experimenting society. What is proposed here is normal science. It presupposes a stable society with governmental stability both in general and in record-keeping particular. It presupposes the meaningful comparison of present, past, and future. The experimenting here proposed is within the framework of such a society, any given experiment being of such a small magnitude as to not fundamentally change that society. For any given step, it is limited to small changes rather than fundamental framework changes. The gradualist will argue that fundamental changes can be made in this way, that by small steps, each validated as improvements, we can move to any optimal society. The revolutionist will agree with Kuhn that the framework is the problem, not the details, that radical changes going beyond the framework are needed, and that these cannot be scientific by the criteria of the existing framework.

Kuhn's challenge to our traditional accretionary model of science is leading to a clarification of basic issues directly relevant to the experimenting society.

Value study. As has already been touched upon in discussion of indicators, the issue of values, of which criteria, will become of preeminent importance in the experimenting society. In the present section, it is proposed to further emphasize this by separating it from the question of indicators (i.e., the question of how well certain

values have been implemented), and to focus analysis more directly upon values as a prior question.

Value critique. This is one relevant activity and one to which today's young social scientists are particularly attracted. Applying this interest to the social indicators already in use and "obviously" relevant, we should be able to generate not only hesitancy in reifying these as goals, but also a long list of additional values to consider.

Critique focused not on criteria but on past ameliorative efforts, such as slum clearance, high rise tenements, aid to dependent children, automation, and the like, can generate an explicit list of noxious values, possible pernicious side-effects to be measured just in case they might be among the effects of a specific program.

Critique should be focused upon the process of measurement and experimentation itself, upon its undesirable side effects, upon ways of deciding upon ameliorative programs and upon measurement procedures without the authoritarian imposition of the boon upon the recipients. Beginnings along this line have been discussed above, especially in citing Janousek (1970). Others challenge the measurement of man and the intangible values of his life as dehumanizing. Even though many of the young social thinkers who would provide such critiques are themselves fundamentally persuaded to an antiscientific humanism, they present challenges which the methodologists of the experimenting society must not overlook, but must face up to with sympathy and ingenuity.

Value analysis. We need direct values-of-life interviews with citizens, focused not only in the abstract, but also on each institutional aspect of their own life history (schools, church, job, shopping facilities, recreational facilities, etc.). But we must not count on free-response public opinion surveys to elicit the full range of recognizable desiderata. Creative value critiques by professional social scientists and humanists must be available to provide a maximally wide range of reminders as to the possible desiderata of the good life.

Methods determining value priorities and of value pooling for collective decision making represent developed social science skills relevant to this quest. We need to make contact with the logic-mathematical puzzles provided by theorists such as Arrow. We need subjective-utility theory from psychoeconomists, and the scaling of social values in the psychometric tradition. We need to study the context-independence of such scaling, of the feasibility of zero point scaling, and the stability of such indices under complex choice contexts. We need to study the transferability of such indices under complex choice contexts. We need to study the transferability of such complex-preference judgments between their lived-in instantiation and verbal-hypothetical presentation. All this should be valuable even if it be recognized as of limited transferability to specific program evaluations. The Demobilization Point system at the end of World War II (Guttman, 1946; Stouffer, 1949) may represent a high point in such psychometric relevance. The detailed relation between the social science input and the practical implementation in that instance deserves more detailed study.

Data Pooling and Decision Making

As described here, the experimenting society will have many experimental comparisons of program alternatives. For each experiment, there will be multiple indicators, each recognized as imperfect. There will be numerous indicators of desired outcomes, and numerous indicators of potential undesirable side effects. The indicators will vary in the degree of success they show and in their own reliability, validity, and potential bias. The experimental design employed will meet to ambiguous degree the nine threats to interval validity and the six threats to external validity.

How is this bewildering mass of data to be digested and integrated into the social decision-making process? Here too the inventive attention of methodologists for the experimenting society is needed.

A social science priesthood. One approach which I reject at present would be to have an expert advisory board integrate all of the data and produce the decisions. Computerized decision processes might be devised, using detailed knowledge of public value preferences. (Such procedures would be valuable even if not used in this way.)

This approach is rejected because it tends to obscure from public view the equivocality of science, and because many of the issues that would divide the council of advisors would be ones of vital interest to legislators and public also. The social science priesthood goes against the open evaluation and nondogmatic qualities of the experimenting society. But the information-reduction needs that make it attractive must be solved somehow.

An auxiliary legislature of social scientists. The concept of an expert panel of social scientists might be expanded into an auxiliary legislature of social evaluation experts, made more representative by the fact that each expert participant would be designated by one legislator as his vicar. Each member would then process advisory decisions in a legislative manner. The actual legislator would then have available his vicar's vote, the auxiliary legislature's collective recommendation, and knowledge of the cruial items of controversy which developed. Such auxiliary legislators might be full time participants for legislative terms (at the discretion of their sponsors), preparing recommendations on all appropriate issues.

Expanding the legislator's capacity for multi-dimensional decision making. We know experimentally that judges in situations such as legislators become overdependent on a few indicators and neglect other dimensions they themselves recognize as relevant. Methodologists need to work on social inventions to support the better digestion of indicators and value priorities in this process. Hammond (1971) is providing decision makers with instantaneous computer support which tells him how heavily he is weighing each ingredient, and allows him to change his weights in a computer implemented multiple-regression type statistical composite. The computer interaction procedures are such that such assistance could be utilized by legislators.

Again, these are wild and unmatured suggestions, hopefully conveying a challenge to potential methodologists for the experimenting society.

CONCLUDING COMMENT

As we develop in detail the procedures, possibilities, and problems of the experimenting society, we will be acquainting ourselves with what it would be like as well as this can be done in advance. As this portrait emerges in greater clarity, it will be our duty to continually ask ourselves if we really want to advocate this monster of measurement and experimentation. We must share the developing picture with the most articulate and hostile critics of such a society and consider in detail their warnings.

If is not a future we want, who should know better or sooner than we, the ambivalent methodologists of the experimenting society.

NOTE

1. This sounds more unsympathetic than I feel. While I share the belief that more moderation in press statements, in the program of retrials, and in other corrections of past falsehoods probably would have avoided the external intervention while allowing the continuation of the more tangible innovations, I also sympathize with the spirit that rejected such compromise. In October 1968 , I attended with 50 others the International Conference on Social Psychology at Prague. The numerous Czechoslovak people we met, while profoundly distressed by the occupation, were still glowing with the excitement of their January to August experiment, and were eager to talk. They freely conceded that any economic reforms that might have been initiated could not have had any effect in that short time, and perhaps would not have even in the long run. But the honesty reforms had immediate effects and were profoundly enjoyed. These were most obvious perhaps in newspapers, television and radio, which became exciting as specific reporters were really allowed to describe things to the best of their personal belief and knowledge. It is hard to exaggerate the contrast which this represented over the prior times in which carefully worded, multiply vetoed statements said only what was wanted said, and even disguised policy changes under ritual jargon. As it affected personal lives, the honesty reform took the shape of dismantling the thought-police apparatus, cessation of persecutions because of beliefs or statements, re-establishment of falsely maligned reputations, and full exposure of past lies of the state. Honesty was a great part of Dubcek's amazing personal popularity. Here was a weak compromise candidate whose past history gave no more promise of greatness than did Harry Truman's, for example. He has been great in how he carried out a role, rather than being one who got a role through greatness. Prior to Dubcek, every official appearing on television read carefully from a multiply censored, cautiously expressed, and often dishonest text. Dubcek instead spoke freely without notes, describing things as he saw them, naively expressing honest emotions in work, facial expression and tone of voice.

As one came to know how decisively (and flagrantly) they had been challenging, changing, and criticizing the past regime, and exposing its cruelty and dishonesty, one naturally wondered if they would not have been wiser to have moved more slowly and less provocatively, and thus have avoided the occupation and retained moderate gains. But this was our question as outsiders. The issue was never raised by them in anything like these terms, even though they were uniformly deeply pessimistic over what was in store for them. When one raised the question, they said: "Perhaps so, but Dubcek had no choice, he was the most conservative and moderate of the liberal majority." The real answer, one felt, was that their joy and pride in their brief, outspoken, liberal and optimistic period made it something they would not want to have missed, that it had a great value that compromise would have spoiled, that it was an experience worth the price of the risk they ran, even as seen then after the gamble had been lost.

Their reform as they saw it was within communism or socialism, not at all changing the ownership of the means of production, and quite interpretable as compatible with Marx and Lenin. (A quotation from Marx was widely displayed which damned large states for using their power to impose their will on small states.) Their slogan was "Communism with a face," a humane communism, a democratic

communism. Their main target was inhumane bureaucracy. Their reforms were primarily in the location of authority: workers' councils for factories, local authorities as decision makers and censors, etc. Coming as it did from within the party apparatus as a majority position, and inspired by their most gifted Marxist writers, the movement served to create a new enthusiasm for worn-out idealogy. A revitalized exportable evangelical communism could easily have resulted, appealing to the disenchanted in both capitalist and communist countries.

REFERENCES

Aubert, V. (1959). Chance in social affairs. *Inquiry, 2*, 1–24.

Boruch, R. F. (1971). Maintaining confidentiality of data in education research: A systemic analysis. *American Psychologist, 26*, 413-430.

Box, G. E. P. and Tiao, G. C. (1965). A change in level of a nonstationary time series. Biometrika, 52, 181–192.

Brickman, P. and Campbell, D.T. (1971). Hedonic relativism and planning the good society. In: M. H. Appley (Ed.), *Adaptation-level Theory: A Symposium*. New York: Academic Press.

Cain, G.G. and Hollister, R.G. (1969). The methodology of evaluating social action programs. Institute for Research on Poverty: Discussion paper. The University of Wisconsin, Madison. Campbell, D.T. (1969a). Reforms as experiments. *American Psychologist, 24*(4), 409–429 (April).

Campbell, D.T. (1969b). Prospective: artifact and control. In: R. Rosenthal and R. Rosnow (Eds.), *Artifact in Behavioral Research*. New York: Academic Press.

Campbell, D.T. (1969c). Objectivity and the social locus of scientific knowledge. Presidential address to the Division of Social and Personality Psychology, America Psychological Association. Duplicated (September 2).

Campbell, D.T. (1971). Methods for the experimenting society. Research proposal to the Russell Sage Foundation (January).

Campbell, D.T. (1970). Considering the case against experimental evaluations of social innovations. *Administrative Science Quarterly, 15*(1), 110–113 (March).

Campbell, D.T. (1971). Temporal changes in treatment-effect correlations: A quasi-experimental model for institutional records and longitudinal studies. In: G. V Glass (Ed.), *The Promise and Perils of Education Information System s*. (Proceedings of the 1970 Invitational Conference on Testing Problems.) Princeton, NJ: Educational Testing Service.

Campbell, D.T. (1974). Evolutionary epistemology. In: P.A. Schilpp (Ed.), *The Philosophy of Karl R. Popper*. LaSalle, IL : Open Court Publishing Co.

Campbell, D.T. and Boruch, R.F. (1970). Measurement and experimentation in social settings. Duplicated research proposal to the National Science Foundation (December).

Campbell, D.T. and Boruch, R.F. (In preparation). *On the possibility of unbiased matching.*

Campbell, D.T. and Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In: J. Hellmuth (Ed.), *Compensatory Education: A National Debate*. Vol. III of *The Disadvantaged Child*. New York: Brunner/Mazel. Also, *Reply to the replies* (same volume).

Campbell, D.T. and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In: N.L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand-McNally. (Reprinted as *Experimental and uasi-experimental design for research*. Chicago: Rand-McNally, 1966.)

Caro, F.G. (1971). Evaluation research: An overview. In: F.G. Caro (Ed.), *Readings in Evaluation Research*. New York: Russell Sage Foundation.

Cicirelli, V.G. (1970). The relevance of the regression artifact problem in the Westinghouse—Ohio evaluation of Head Start: A reply to Campbell and Erlebacher. In: J. Hellmuth (Ed.), *Compensatory Education: A National Debate*. Vol. III of *The Disadvantaged Child*. New York: Brunner/Mazel.

Cook, S.W. et al. (1971). Proposed ethical principles submitted to the APA membership for criticism and modification. *APA Monitor, 2*(7), 9–28.

Cook, T.D. and Diamond, S.S. (1971). True field experiments in pure and applied social psychology. Duplicated manuscript, Northwestern University.

Dubos, Rene. (1970). Address to the Centennial Celebration of Loyola University. Chicago (February).

Dunn, E.S. (1971). *Economic and Social Development: A Process of Social Learning*. Baltimore: The John Hopkins Press.

Etzioni, A. (1968). "Shortcuts" to social change? *The Public Interest, 12*, 40–51.

Evans, J.W. and Schiller, J. (1970). How preoccupation with possible regression artifacts can lead to faulty strategy for the evaluation of social action programs. A reply to Campbell and Erlebacher. In: J. Hellmuth (Ed.), *Compensatory Education: A National Debate*. Vol. III of *The Disadvantaged Child*. New York: Brunner/Mazel.

Feldman, A.S. and Campbell, D.T. (In preparation, 1971). *Experimental Socialism and Marxist Theories*.

Fischer, J.L. (1971). The uses of Internal Revenue Service income information for measuring the impact of manpower programs. In: M.E. Borus (Ed.), A Conference on the Evaluation of the Impact of Manpower Programs. Columbus, OH (June 15–17).

Glass, G.V, Tiao, G.C., and Maguire, T.O. (1971). Analysis of data on the 1900 revision of the German divorce laws as a quasi-experiment. *Law and Society Review, 3*(1), 539–562.

Gordon, A.C. (1971). The university-community interface: Progress and prospects. In: J.T. Walton and E.E. Carns (Eds.), *Sociology* (In preparation). Duplicated manuscript (August).

Gordon, A.C. and Campbell, D.T. (November 1970–June 1971). Recommended accounting procedures for the evaluation of improvements in the delivery of state social services. Report to the Illinois Institute for Social Policy. Evanston, IL: Northwestern University, Center for Urban Affairs.

Gordon, A.C. and Myers, J.R. (1970). Methodological recommendations for extensions of the Heussenstamm bumper sticker study. Duplicated report. Evanston, IL: Northwestern University, Center for Urban Affairs.

Guttman, L. (1946). An approach for quantifying paired comparisons and rank order. *Annals of Mathematical Statistics, 17*, 1–218.

Hammond, K.R. (1971). *Science*.

Harris, F.R. (Ed.) (1970). *Social Science and National Policy*. Chicago: Aldine Publishing Co.

Harris, Louis and Associates. (1970). Survival literacy study. Duplicated report for the National Reading Council, 1776 Massachusetts Avenue, Washington, DC 20036 (September).

Haworth, L. (1960). The experimenting society: Dewey and Jordan. *Ethics, 71*(1), 27–40 (October).

Heller, R.N. (1971). The uses of social security administration data for measuring the impact of manpower programs. In: M.E. Borus (Ed.), A Conference on the Evaluation of the Impact of Manpower Programs. Columbus, OH (June 15–17).

Heussenstamm, F. (1971). Bumper stickers and the cops. *Transaction, 8*(4), 32–33 (February).

Hyman, H.H. and Wright, C.R. (1967). Evaluating social action programs. In: R.F. Lazarsfeld, W.H. Sewell, and H.L. Wilensky (Eds.), *The Uses of Sociology*. New York: Basic Books .

Ikeda, K., Yinger, J.M., and Laycock, F. (1970). Reforms as experiments and experiments as reforms. Ohio Valley Sociological Society Meetings, Akron, OH. Duplicated paper (May).

Janousek, J. (1970). Comments on Campbell's "Reforms as Experiments." *American Psychologist, 25*, 191–193.

Kepka, E.J. (1971). Model representation and the threat of instability in the interrupted time series quasi-experiment. Ph.D. dissertation, Northwestern University.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakatos, I. and Musgrave, A. (1970). *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

Levenson, B. and McDill, M.S. (1966). Vocational graduates in auto mechanics: A follow-up study of negro and white youth. *Phylon, 27*(4), 347–357.

Levine, M. (1970). Scientific method and the adversary model: Some preliminary thoughts. Duplicated report, State University of New York at Buffalo.

Levine, R.A. (1969). Evaluating the war on poverty. In: J.L. Sundquist (Ed.), *On Fighting Poverty*. New York: Basic Books.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*, 34–46.
Lewin, K. (1948). *Resolving Social Conflicts*. New York: Harper.
Lord, F.M. (1960). Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307–321.
Lorenz, K. (1941). Kants lehre vom apriorischen im Lichte gegenwartiger Biologie. *Blatter fur Deutsche Philosophi, 15*, 94–125. Translated in: L. von Bertalanffy and A. Rapoport (Eds.) (1962). *General Systems*, (III), 23–35.
Michael, D. N. (1968). *The Unprepared Society: Planning for a Precarious Future*. New York: Harper and Row.
Montagna, P.D. (1971). The public accounting profession: Organization, ideology and social power. *American Behaviorist Scientist, 14*(4), 475–491 (March/April).
Moore, O.K. (1957). Divination, a new perspective. *American Anthropologist, 59*, 72.
Office of Economic Opportunity. (1971). Further preliminary results: The New Jersey graduated work incentive experiment. Washington, DC: OEO.
Orr, L.L., Hollister, R.G., Lefcowitz, M.J. and Hester, K. (1971). *Income Maintenance*. Chicago: Markham Publishing.
Parker, E.B. (1963). The effects of television on public library circulation. *Public Opinion Quarterly, 27*, 578–589.
Polanyi, M. (1969). The message of the Hungarian revolution. *The American Scholar, 35*, 261–276. Reprinted in: M. Grene (Ed.) (1969) *Knowing and Being: Essays by Michael Polanyi*. London: Routledge and Kegan Paul.
Polanyi, M. (1966). A society of explorers. In: M. Polanyi, *The Tacit Dimension*. New York: Doubleday.
Polanyi, M. (1967). The growth of science in society. Minerva, 5, 533–545.
Popper, K.R. (1945). *The Open Society and Its Enemies*. London: Routledge, (Harper Torchbooks, 1963).
Popper, K.R. (1963). *Conjectures and Refutations*. London: Routledge and Kegan Paul, New York: Basic Books.
Porter, A.C. (1967). The effects of using fallible variables in the analysis of covariance. Ph.D. dissertation, University of Wisconsin. (University Microfilms, Ann Arbor, Michigan, 1968).
Ross, H.L., Campbell, D.T., Glass, G.V (1970). Determining the social effects of a legal reform: The British "breathalyser" crackdown of 1967. *American Behavioral Scientist, 13*(4), 493–509 (March-April).
Rossi, R.H. (1969). Practice, method, and theory in evaluating social-action programs. In: J.L. Sundquist (Ed.), *On Fighting Poverty*. New York: Basic Books.
Sanford, N. (1970). Whatever happened to action research? *Journal of Social Issues, 26*, 3–23.
Sawyer, J. and Schechter, H. (1968). Computers, privacy, and the National Data Center: The responsibility of social scientists. *American Psychologist, 23*, 810–818.
Schaff, A. (1963). *A Philosophy of Man*. New York: Monthly Review Press.
Schwartz, R.D. (1961). Field experimentation of sociolegal research. *Journal of Legal Education, 13*, 401–410.
Schwartz, R.D. and Orleans, S. (1967). On legal sanctions. *University of Chicago Law Review, 34*(2), 274–300.
Seligman, M.E.P. (1969). Control group and conditioning: A comment on operationism. *Psychological Review, 76*, 484–491.
Shaver, P. and Staines, G. (1971). Problems facing Campbell's "experimenting society." *Urban Affairs Quarterly, 7*(2), 173–186.
Skolimowski, H. (In press). *Polish Marxism*.
Stapleton, V. (1970). Counsel in American juvenile courts: An experimental study of lawyers in delinquency hearings. Ph.D. dissertation, Northwestern University.
Stapleton, V. and Teitelbaum, L. (In press). *In Defense of Youth: A Study of the Role of Counsel in American Juvenile Courts*. New York: Russell Sage Foundation.
Stouffer, S.A. (1949). The point system for redeployment and discharge. In: S. A. Stouffer et al., *The American Soldier*. Vol. 2, *Combat and its Aftermath*. Princeton: Princeton University Press.

Sundquist, J.L. (Ed.) (1969). *On Fighting Poverty. Perspectives on Poverty*, Vol II. New York: Basic Books.

Sween, Joyce. (1971). The experimental regression design: An inquiry into the feasibility of non-random treatment allocation. Ph.D. dissertation, Northwestern University.

Watts, H.W. (1969). Graduated work incentives: An experiment in negative taxation. *American Economic Review, 57*, 463–472.

Watts, H.W. (1971). Midexperiment report on basic labor-supply response. Discussion paper, Institute for Research on Poverty, The University of Wisconsin, Madison.

Weiss, R.S. and Rein, M. (1969). The evaluation of broad-aim programs: A cautionary case and a moral. *Annals of the American Academy of Political and Social Sciences, 385*, 133–142.

Weiss, R.S. and Rein, M. (1970). The evaluation of broad-aim programs: Experimental design, its difficulties, an alternative. *Administrative Science Quarterly, 15*, 97–109.

Wholey, J.S., Scanlon, J.W., Fukumoto, J., and Vogt, L.M. (1970). *Federal Evaluation Policy*. Washington, DC: The Urban Institute.

Williams, W. and Evans, J.W. (1969). The politics of evaluation: The case of Head Start. *The Annals, 385*, 118–132 (September).

Zeisel, H. (1970). Reducing the hazards of human experiments through modifications in research design. *Annals of the New York Academy of Sciences*, 169, 475–486 (January 21).

Zeisel, H. (1971). *Say It With Figures* (Rev. Ed.) New York: Harper and Row.

ABOUT THE AUTHOR

Donald T. Campbell is University Professor at Lehigh University and professor emeritus of psychology at Northwestern University. He is a former president of the American Psychological Association and a member of the National Academy of Sciences.