Assessing the Impact of Planned Social Change*

Originally published as Paper #8, Occasional Paper Series, December, 1976

* Reprinted with permission of The Public Affairs Center, Dartmouth College[1].

Donald T. Campbell

It is a special characteristic of all modern societies that we consciously decide on and plan projects designed to improve our social systems. It is our universal predicament that our projects do not always have their intended effects. Very probably we all share in the experience that often we cannot tell whether the project had any impact at all, so complex is the flux of historical changes that would have been going anyway, and so many are the other projects that might be expected to modify the same indicators.

It seems inevitable that in most countries this common set of problems, combined with the obvious relevance of social science research procedures, will have generated a methodology and methodological specialists focused on the problem of assessing the impact of planned social change. It is an assumption of this paper that, in spite of differences in forms of government and approaches to social planning and problem-solving, much of this methodology can be usefully shared–that social project evaluation methodology is one of the fields of science that has enough universality to make scientific sharing mutually beneficial. As a part of this sharing, this paper reports on program impact assessment methodology as it is developing in the United States today.

The most common name in the U.S. for this developing specialty is evaluation research, which now almost always implies program evaluation (even though the term "evaluation has a well-established usage in assessing the adequacy of persons in the execution of specific social roles). Already there are a number of anthologies and textbooks in this area. (Suchman, 1967; Caro, 1971; Weiss, 1972a, 1972b; Rivlin, 1971; Rossi & Williams, 1972; Glaser, 1973; Fairweather, 1967; Wholey, et al., 1970;Caporaso & Roos, 1973; Riecken, Boruch, Campbell, Caplan, Gennan, Pratt, Rees, & Williams, 1974.) There is a journal, Evaluation,

[1] Preparation of this paper has been supported in part by grants from the Russell Sage Foundation and the National Science Foundation, Grant Number SOC-7103704-03. A version of this paper was presented to the Visegrad, Hungary, Conference on Social Psychology, May 5-10, 1974.

which is being given free distribution during a trial period and after three issues seems likely to survive (address: 501 S. Park Ave., Minneapolis, Minnesota 55415). There is also Evaluation Comment: The Journal of Educational Evaluation (address: 145 Moore Hall, University of California, Los Angeles, 90024). Two other journals, which frequently have materials of this sort, are Social Science Research, edited by two leaders in the field in the U.S., James Coleman and Peter Rossi, and Law & Society Review, Founded by Richard D. Schwartz. Many other journals covering social science research methods carry important contributions to this area.

The participants in this new area come from a variety of social science disciplines. Economists are well represented. Operations research and other forms of "Scientific management" contribute. Statisticians, sociologists, psychologists, political scientists, social service administration researchers, and educational researchers all participate. The similarity of what they all end up recommending and doing testifies to the rapid emergence of a new and separate discipline that may soon have its own identity divorced from this diverse parentage.

Since my own disciplinary background is social psychology, I feel some need to comment on the special contribution that this field can make even though I regard what I am now doing as "applied social science" rather than social psychology. First, of all of the contributing disciplines, psychology is the only one with a laboratory experimental orientation, and social psychologists in particular have had the most experience in extending laboratory experimental design to social situations. Since the model of experimental science emerges as a major alternative in reducing equivocality about what caused what in program evaluation (from Suchman's, 1967, founding book onward), this is a very important contribution of both general orientation and specific skills.

Second, psychologists are best prepared with appropriately critical and analytic measurement concepts. Economists have an admirably skeptical book on monetary records (Morgenstern, 1963), but most economists treat the figures available as though they were perfect. Sociologists have a literature on interviewer bias and under enumeration, but usually treat census figures as though they were unbiased. Psychology, through its long tradition of building and criticizing its own measures, has developed concepts and mathematical models of reliability and validity which are greatly needed in program evaluation, even though they are probably not yet adequate for the study of the cognitive growth of groups differing in ability. The concept of bias, as developed in the older psychophysics in the distinction between "constant error" (bias) and "variable error" (unreliability), and the more recent work in personality and attitude measurement on response sets, halo effects, social desirability factors, index correlations, methods factors, etc. (Cronbach, 1946, 1950; Edwards, 1957; Jackson & Messick, 1962; Campbell, Siegman & Rees, 1967; Campbell & Fiske, 1959) is also very important and is apt to be missing in the concept of validity if that is defined in terms of correlation coefficient with a criterion. All this, of course, is not our monopoly. Indeed, it is the qualitative sociologists who do studies of the conditions under which social statistics get laid down (e.g., Becker, et al., 1968, 1970; Douglas, 1967; Garfinkel, 1967; Kitsuse & Ciccourel, 1963; Beck,

1970) who best provide the needed skepticism of such measures as suicide rates and crime rates. But even here, it is psychologists who have had the depth of experience sufficient to distinguish degrees of validity lying between total worthlessness and utter perfection, and who have been willing to use, albeit critically, measures they knew were partially biased and errorful.

Third, many of the methodological problems of social implementation and impact measurement have to do with the social psychology of interaction between citizens and projects, or between citizens and modes of experimental implementation (randomization, control groups), or between citizens and the special measurement procedures introduced as a part of the evaluation. These are special problems of attitude formation and of the effects of attitudes on responses, and are clearly within the domain of our intended competence.

Having said something about U.S. evaluation research in its professional aspects, I would like to spend the rest of the time telling about the problems we have encountered so far and the solutions we have proposed. It is with regret that I note that we have progressed very far from my earlier review (Campbell, 1969b); however, I will attempt to provide new illustrations.

The focus of what follows is so much on troubles and problems that I feel the necessity of warning and apologizing. If we set out to be methodologists, we set out to be experts in problems and, hopefully, inventors of solutions. The need for such a specialty would not exist except for the problems. From this point of view, no apology is needed. But I would also like to be engaged in recruiting members to a new profession and in inspiring them to invest great effort in activities with only long-range payoff. For potential recruits, or for those already engaged in it, a full account of our difficulties, including the problem of getting our skills used in ways we can condone, is bound to be discouraging. We cannot yet promise a set of professional skills guaranteed to make an important difference. In the few success stories of beneficial programs unequivocally evaluated, society has gotten by, or could have gotten by, without our help. We still lack instances of important contributions to societal innovation which were abetted by our methodological skills. The need for our specialty, and the specific recommendations we make, must still be justified by promise rather than by past performance. They are a priori in that they represent extrapolations into a new context not yet cross-validated in that context. I myself believe that the importance of the problem of social system reality-testing is so great that our efforts and professional commitment are fully justified by promise. I believe that the problems of equivocality of evidence for program effectiveness are so akin to the general problems of scientific inference that our extrapolations into recommendations about program evaluation procedures can be, with proper mutual criticism, well-grounded. Nonetheless, motivated in part by the reflexive consideration that promising too much turns out to be a major obstacle to meaningful program evaluation, I aim, however ambivalently, to present an honestly pessimistic picture of the problem.

A second problem with the problem focus comes from the fact that inevitably many of the methodological difficulties are generated from the interaction of aspects of the political situation surrounding programs and their

evaluation. Thus the U.S. experience in evaluation research, combined with the focus on problems, may make my presentation seem inappropriately and tactlessly critical of the U.S. system of government and our current political climate. Ideally, this would be balanced out by the sharing experiences from many nations. In the absence of this, I can only ask you not to be misled by this by-product of the otherwise sensible approach of focusing on problems.

It is in the area of methodological problems generated by political considerations that the assumptions of universality for the methodological principles will fail as we compare experiences from widely differing social, economic, and political systems. You listeners will have to judge the extent, if any, to which the same politico-methodological problems would emerge in your program evaluation settings. Most international conferences of scientists can avoid the political issues which divide nations by concentrating on the scientific issues which unite them as scientists. On the topic of assessing the impact of planned social change we do not have that luxury. Even so, I have hopes for a technology that would be useful to any political system. I believe that much of the methodology of program evaluation will be independent of the content of the program, and politically neutral in this sense. This stance is augmented by emphasizing the social scientist's role in helping society keep track of the effects of changes that its political process has initiated and by playing down the role of the social scientist in the design of program innovations. Whether this independence of ideology is possible, and even if it is, how it is to be integrated with our social scientist's duty to participate in the development of more authentic human consciousness and more humane forms of social life are questions I have not adequately faced, to say nothing of resolved.

In what follows, I have grouped our problems under three general headings, but have made little effort to keep the discussion segregated along these lines. First comes issues that are internal to our scientific community and would be present even if scientist program evaluators were running society solely for the purpose of unambiguous program evaluation. These are: 1) Metascientific issues; and 2) Statistical issues. The remaining heading involves interaction with the societal context. Under 3) Political system problems, I deal with issues that specifically involve political processes and governmental institutions, some of which are perhaps common to all large, bureaucratic nations, other unique to the U.S. setting.

Metascientific Issues

Quantitative vs. Qualitative Methodology

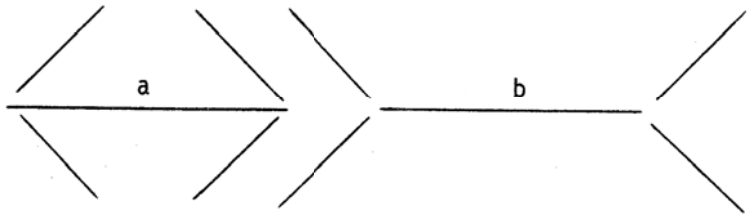
A controversy between "qualitative" versus "quantitative" modes of knowing, between *geisteswissenschaftlich* and *naturewissenchaftlich* approaches, between "humanitistic" and "scientistic" approaches is characteristic of most of the social sciences in the U.S.A. today. In fields such as sociology and social psychology, many of our ablest and most dedicated graduate students are increasingly opting for the qualitative, humanitistic mode. In political science, there has been a continuous division along these lines. Only economics and geography seem relatively immune.

Inevitably, this split has spilled over into evaluation research, taking the form of a controversy over the legitimacy of the quantitative-experimental paradigm for program evaluation (e.g., Weiss & Rein, 1969, 1970; Guttentag, 1971, 1973; Campbell, 1970, 1973). The issue has not, to be sure, been argued in quite these terms. The critics taking what I am calling the humanitistic position are often well-trained in quantitative-experimental methods. Their specific criticisms are often well-grounded in the experimentalist's own framework: experiments implementing a single treatment in a single setting are profoundly ambiguous as to what caused what; there is a precarious rigidity in the measurement system, limiting recorded outcomes to those dimensions anticipated in advance; process is often neglected in an experimental program focused on the overall effect of a complex treatment, and thus knowing such effects has only equivocal implications for program replication or improvement; broad-gauge programs are often hopelessly ambiguous as to goals and relevant indicators; change of treatment program during the course of an ameliorative experiment, while practically essential, make input-output experimental comparisons uninterpretable; social programs are often implemented in ways that are poor from an experimental design point of view; even under well-controlled situations, experimentation is a profoundly tedious and equivocal process; experimentation is too slow to be politically useful; etc. All these are true enough, often enough to motivate a vigorous search for alternatives. So far, the qualitative-knowing alternatives suggested (e.g., Weiss & Rein, 1969, 1970; Guttentag, 1971, 1973) have not been persuasive to me. Indeed, I believe that naturalistic observation of events is an intrinsically equivocal arena for causal inference, by qualitative or quantitative means, because of the ubiquitous confounding of selection and treatment. Any efforts to reduce that equivocality will have the effect of making conditions more "experimental." "Experiments" are, in fact, just that type of contrived observational setting optimal for causal inference. The problems of inference surrounding program evaluation are intrinsic to program settings in ongoing social processes. Experimental designs do not cause these problems and, in fact, alleviate them, though often only slightly so.

In such protests, there often seems implicitly a plea for the substitution of qualitative clairvoyance for the indirect and presumptive processes of science. But while I must reject this aspect of the humanistic protest, there are other aspects of it that have motivated these critics in which I can wholeheartedly join. These other criticisms may be entitled "neglect of relevant qualitative contextual evidence" or "over dependence upon a few quantified abstractions to the neglect of contradictory and supplementary qualitative evidence."

Too often qualitative social scientists, under the influence of missionaries from logical positivism, presume that in true science, quantitative knowing replaces qualitative, common-sense knowing. The situation is in fact quite different. Rather, science depends upon qualitative, common-sense knowing even though at best it goes beyond it. Science in the end contradicts some items of common sense, but it only does so by trusting the great bulk of the rest of common-sense knowledge. Such revision of common sense by common sense, which, paradoxically, can only be done by trusting more common sense. Let us

consider as an example the Muller-Lyer illustration (Figure 1).



If you ask the normal resident of a "carpentered" culture (Segall, et al., 1966) which line is longer, a or b, he will reply b. If you supply him with a ruler, or allow him to use the edge of another piece of paper as a makeshift ruler, he will eventually convince himself that he is wrong, and that line a is longer. In so deciding he will have rejected as inaccurate one product of visual perception by trusting a larger set of other visual perceptions. He will also have made many presumptions, inexplicit for the most part, including the assumption that the lengths of lines have remained relatively constant during the measurement process, that the ruler was rigid rather than elastic, that the heat and moisture of his hand have not changed the ruler's length in such a coincidental way as to product the different measurements, expanding it when approaching line a and contracting it when approaching line b, etc.

Let us take as another example a scientific paper containing theory and experimental results demonstrating the particulate nature of light, in dramatic contract to common-sense understanding. Or a scientific paper demonstrating that what ordinary perception deems "solids" are in fact open lattices. Were such a paper to limit itself to mathematical symbols and purely scientific terms, omitting ordinary language, it would fail to communicate to another scientist in such a way as to enable him to replicate the experiment and verify the observations. Instead, the few scientific terms have been imbedded in a discourse of elliptical prescientific ordinary language which the reader is presumed to (and presumes to) understand. And in the laboratory work of the original and replicating laboratory, a common-sense, prescientific language and perception of objects, solids, and light was employed and trusted in coming to the conclusions that thus revise the ordinary understanding. To challenge and correct the common-sense understanding in one detail, common-sense understanding in general had to be trusted.

Related to this is the epistemological emphasis on qualitative pattern identification as prior to an identification of quantifiable atomic particles, in reverse of the logical atomist's intuition, still to widespread (Campbell, 1966). Such an epistemology is fallibilist, rather than clairvoyant, emphasizing the presumptive error-proneness of such pattern identification, rather than perception as a dependable ground of certainty. But it also recognizes this fallible, intuitive, presumptive, ordinary perception to be the only route. This is not to make perceptions uncriticizable (Campbell,

1969a), but they are, as we have seen, only criticizable by trusting many other perceptions of the same epistemic level.

If we apply such an epistemology to evaluation research, it immediately legitimizes the "narrative history" portion of most reports and suggests that this activity be given formal recognition in the planning and execution of the study, rather than only receiving attention as an afterthought. Evaluation studies are uninterpretable without this, and most would be better interpreted with more. That this content is subjective and guilty of perspectival biases should lead us to better select those who are invited to record the events, and to prepare formal procedures whereby all interested participants can offer additions and corrections to the official story. The use of professionally trained historians, anthropologists, and qualitative sociologists should be considered. The narrative history is indispensable sociologists should be considered. The narrative history is an indispensable part of the final report, and the best qualitative methods should be used in preparing it.

We should also recognize that participants and observers have been evaluating program innovations for centuries without benefit of quantification or scientific method. This is the common-sense knowing which our scientific evidence should build upon and go beyond, not replace. But it is usually neglected in quantitative evaluations, unless a few supporting anecdotes haphazardly collected are included. Under the epistemology I advocate, one should attempt to systematically tap all the qualitative common-sense program critiques and evaluations that have been generated among the program staff, program clients and their families, and community observers. While quantitative procedures such as questionnaires and rating scales will often be introduced at this stage for reasons of convenience in collecting and summarizing, non-quantitative methods of collection and compiling should also be considered, such as hierarchically organized discussion groups. Where such evaluations are contrary to the quantitative results, the quantitative results should be regarded as suspect until the reasons for the discrepancy are well understood. Neither is infallible, of course. But for many of us, what needs to be emphasized is that the quantitative results may be as mistaken as the qualitative. After all, in physical science laboratories, the meters often work improperly, and it is usually qualitative knowing, plus assumptions about what the meter ought to be showing, this discovers the malfunction. (This is a far cry from the myth that meter readings operationally define theoretical parameters.)

It is with regret that I report that in U.S. program evaluations, this sensible joint use of modes of knowing is not yet practiced. Instead, there seems to be an all or none flip-flop. Where, as in Model Cities evaluation, anthropologists have been used as observers, this has often been in place of, rather than in addition to, quantitative indicators, pretests, posttests, and control-group comparisons. A current example of the use of anthropologists in the "Experimental Schools" program started in the U.S. Office of Education and now in the national Institute of Education. In this program, school-system initiative is encouraged, and winning programs receive substantial increments to their budgets (say 25%) for use in implementing the innovations. To evaluate some of these programs, very expensive contracts have been let for

anthropological process evaluations of single programs. In one case, this was to involve a team of five anthropologists for five years, studying the school system for a unique city with a population of 100,000 persons. The anthropologists have no prior experience with any other U.S. school system. They have been allowed no base-line period of study before the program was introduced; they arrived instead after the program had started. They were not scheduled to study any other comparable school system not undergoing this change. To believe that under these disadvantaged observational conditions, these qualitative observers could infer what aspects of the processes they observe were due to the new program innovation requires more faith than I have, although I should withhold judgment until I see the products. Furthermore, the emphasis of the study is on the primary observations of the anthropologists themselves, rather than on their role in using participants as informants. As a result there is apt to be a neglect of the observations of other qualitative observers better placed than the anthropologists. These include the parents who have had other children in the school prior to the change; the teachers who have observed this one system before, during, and after the change; the teachers who have transferred in with prior experience in otherwise comparable systems; and the students themselves. Such observations one would perhaps want to mass produce in the form of questionnaires. If so, one would wish that appropriate questions had also been asked prior to the experimental program, and on both occasions in some comparable school system undergoing no such reform, thus reestablishing experimental design and quantitative summaries of qualitative judgments. (For a more extended discussion of the qualitative-quantitative issues, see Campbell, 1975.)

While the issue of quantitative vs. qualitative orientations has important practical implications, it is still, as I see it, primarily an issue among us social scientists and relatively independent of the larger political process. Whether one or the other is used has pretty much been up to the advice of the segment of the social science community from which advice was sought, motivated in part by frustration with a previously used model. The issue, in other words, is up to us to decide.

The remaining issues in the metascientific group are much more involved with extrascientific issues of human nature, social systems, and political process. I have classified them here only because I judge that a first step in their resolution would be developing a consensus among evaluation methodologists, and such a consensus would involve agreement on metascientific issues rather than on details of method.

Separation of Implementation and Evaluation

A well-established policy in those U.S. government agencies most committed to program evaluation is to have program implementation organizationally separated from program evaluation. This recommendation comes from the academic community of scientific management theory, proliferated in the governmental circles of the late 1960's as “Programming, Planning, and Budgeting System,” or PPBS, in which these functions, plus program monitoring or evaluation, were to be place in a separated

organizational unit independent of the operating agencies. (William & Evans, 1969, provide one relevant statement of this policy.) This recommendation is based on an organizational control theory of check and balances. It is supported not only by general observations on human reluctance to engage in self-criticism, but more particularly on observations of a long standing self-defeating U.S. practice in which progress reports and other program evaluations are of necessity designed with the primary purpose of justifying the following year's budget. For the typical administrator of an ameliorative program in the U.S.A., be it a new experimental program or one of long-standing, budgets must be continually justified, and are usually on a year-to-year basis with six months or more lead-time rare. For such an administrator, program evaluations can hardly be separated from this continual desperate battle. In this context, it makes excellent sense to turn program evaluations over to a separate unit having no budgetary constraints on an honest evaluation. And so far, the policy is unchallenged.

My own observations, however, lead me to the conclusion that this policy is not working either. The separation works against modes of implementation that would optimize interpretability of evaluation data. There are such, and low cost ones too, but these require advance planning and close implementer/ evaluator cooperation. The external evaluators also tend to lack the essential qualitative knowledge of what happened. The chronic conflict between evaluators and implementers, which will be bad enough under a unified local direction, tend to be exacerbated. Particularly when combined with U.S. research contracting procedures, the relevance of the measures to local program goals and dangers is weakened. Evaluation becomes a demoralizing influence and a source of distracting conflict. It might be hoped that through specialization, more technically proficient methodologists would be employed. If there is such a gain, it is more than lost through reduced experimental control.

These problems are, of course, not entirely due to the separation of implementation and evaluation. And the reasons that argue for the separation remain strong. Yet the problems are troublesome and related enough to justify reconsidering the principle, particularly when it is noted that the separation seems totally lacking in experimental science. This raises the metascientific issue of how objectivity in science is obtained in spite of the partisan bias of scientists, and of the relevance of this model for objectivity in program evaluation.

In ordinary science, the one who designs the experiment also reads the meter. Comparably biasing motivational problems exist. Almost inevitably, the scientist is a partisan advocate of one particular outcome. Ambiguities of the interpretation present themselves. Fame and Careers are at stake. Errors are made, and not all get corrected before publication, with the hypothesis-supporting errors much less likely to be caught, etc. The puzzle of how science gets its objectivity (if any) is a metascientific issue still unresolved. While scientists are probably more honest, cautious, and self-critical than most groups, this is more apt to be a by-product of the social forces that produce scientific objectivity than the source. Probably the tradition and possibility of independent replication is a major factor. As the philosophers and sociologists of science better clarify this issue, evaluation research methodologists should be alert to the possibility of models

applicable to their area. Jumping ahead speculatively, I come to the following tentative stance.

Ameliorative program implementation and evaluation in the U.S.A. today need more zeal, dedication, and morale. These would be increased by adopting the scientist's model of experimenter-evaluator. If the conditions for cross-validating replication could be established, and if budgetary jeopardy from negative evaluations could be removed (for example, by allowing program implementers to shift to alternative programs in pursuit of the same goal), then the policy separation of implementation and evaluation should be abandoned.

The issue does not have to be one or the other. External evaluations can be combined with in-house evaluations. Certainly, even under present budgeting systems, program implementers should be funded to do their own evaluations and to argue their validity in competition with external evaluations. The organizational arrangement separating evaluation from implementation is borrowed from the model of external auditors, and it should be remembered that in accounting, auditors check on the internal records, rather than creating new data. Perhaps some such evaluation methodologist's audit of internal evaluation records would be enough of an external evaluation.

Maximizing Replication and Criticism

Continuing the same metascience theme as in the previous section: a number of other recommendations about policy research emerge, some of which run counter to current U.S. orthodoxy and practice.

At present, the preference is for single, coordinated, national evaluations, even where the program innovation is implemented in many separate, discrete sites. If one were to imitate science's approach to objectivity, it would instead seem optimal to split up the big experiments and evaluations into two or more contracts with the same mission so that some degree of simultaneous replication would be achieved. Our major evaluations of compensatory education programs (e.g., Head Start, Follow Through) offer instances which were of such magnitude that costs would not have been appreciably increased by this process. We could often, if we so planned, build in some of the competitive replication that keeps science objective.

One positive feature of the U.S. evaluation research scene in this regard is the widespread advocacy and occasional practice of reanalysis by others of program evaluation data. The Russell Sage Foundation has funded a series of these, including one on the "Sesame Street" preschool television programs (Cook, et al., 1975). The original governmental evaluation of the Head Start compensatory preschool program (Circirelli, 1969) has been reanalyzed by Smith and Bissell (1970) and Barnow (1973), and others are in progress. Similarly, for several other classic bodies of evaluation data, although this is still a rare activity and many sets of data are not made available.

One needed change in research customs or ethics is toward the encouragement of "minority reports" from the research staff. The ethic that the data should be available for critical reanalysis should be explicitly extended to include the staff members who did the data collection and analysis and who very frequently have the detailed insight to see

how the data might be assembled to support quite different conclusions than the official report presents. At present, any such activity would be seen as reprehensible organizational disloyalty. Because of this, an especially competent source of criticism, and through this a source of objectivity, is lost. An official invitation by sponsor and administrator to every member of the professional evaluation team to prepare minority reports would be of considerable help in reducing both guilt and censure in this regard.

In this regard, we need to keep in mind two important models of social experimentation. On the one hand there is the big-science model, exemplified in the Negative Income Tax experiments to be discussed below (See also Kershaw's paper in this volume). On the other hand, there is the low-budget "administrative experiment" (Campbell, 1967; Thompson, 1974), in which an administrative unit such as a city or state (or factory, or school) introduces a new policy in such a way as to achieve experimental or quasi-experimental tests of its efficacy. Wholey's paper (in this volume) describes such studies and the Urban Institute in general has pioneered in this regard. Hatry, Winnie, and Fisk's Practical Program Evaluation for State and Local government Officials (1973) exemplifies this emphasis. For administrative experimentation to produce objectivity, cross-validating diffusion is needed, in which those cities or states, etc., adopting a promising innovation confirm its efficacy by means of their own evaluation effort.

Decentralization of decision-making has the advantage of creating more social units that can replicate and cross-validate social ameliorative inventions or that can explore a wide variety of alternative solutions simultaneously. Even without planning, the existence in the U.S.A. of state governments creates quasi-experimental comparisons that would be unavailable in a more integrated system. Zeisel (1971) has argued this well, and it is illustrated in the study of Baldus (1973) cited more extensively below. If factories, schools, and units of similar size, are allowed independent choice of programs and if borrowed programs are evaluated as well as novel ones, the contagious borrowing of the most promising programs would provide something of the validation of science.

Evaluation Research as Normal Rather than Extraordinary Science

The metascientific points so far have shown little explicit reference to the hot metascientific issues in the U.S. today. Of these, the main focus of discussion is still Thomas Kuhn's Structure of Scientific Revolution (1970). While I would emphasize the continuity and the relative objectivity of science more than he (as you have already seen), I recognize much of value in what he says, and some of it is relevant here. To summarize: There are normal periods of scientific growth during which there is general consensus on the rules for deciding which theory is more valid. There are extraordinary or revolutionary periods in science in which the choices facing scientists have to be made on the basis of decision rules which are not party of the old paradigm. Initially, the choice of the new dominant theory after such a revolution is unjustified in terms of decision rules of the prior period of normal science.

For evaluation research, the Kuhnian metaphor of revolution can be returned to the political scene. Evaluation research is

clearly something done by, or at least tolerated by, a government in power. It presumes a stable social system generating social indicators that remain relatively constant in meaning so that they can be used to measure the program's impact. The programs which are implemented must be small enough not to seriously disturb the encompassing social system. The technology I am discussing is not available to measure the social impact of a revolution. Even within a stable political continuity, it may be limited to the relatively minor innovations, as Zeisel has argued in the case of experimentation with the U.S. legal system. (Needless to say, I do not intend this to constitute a valid argument against making changes of a magnitude that precludes evaluation.)

Statistical Issues

In this section I will get into down-to-earth issues where we quantitative evaluation methodologists feel most at home. Here are issues that clearly call for a professional skill. Here are issues that both need solving and give promise of being solvable. These statistical issues are ones that assume a solution to the metascientific issues in favor of a quantitative experimental approach. In this section, I will start with a useful common-sense method–the interrupted time-series. Next will come some popular but unacceptable regression approaches to quasi-experimental design. Following that, problems with randomization experiments will be discussed, and following that, a novel comprise design.

The Interrupted Time-Series Design

By this term I cover the formalization of the widespread common practice of plotting a time-series on some social statistic and attempting to interpret it. This practice, the problem encountered, and the solution have been developed independently in many nations. I will start from some non-U.S. examples.

Figure 2 shows data on sex crimes in Denmark and possibly the effect of removing restrictions on sale of pornography (Kutchinsky, 1973). Kutchinsky is cautious about drawing causal conclusions, emphasizing the changes in the tolerance of citizens in lodging complaints and of policemen may have produced a drop in number of reported offenses without a drop in actual offenses. By studying attitudes of citizens and police over time, and by other subtle analyses, he concludes that for child molestation these other explanations do not hold, and one must conclude that a genuine drop in the frequency of this crime occurred. However, the graphic portrayal of these trends in Figure 3 is less convincing than Figure 2 because of a marked downward trend prior to the increased availability of pornography. In both cases interpretation of effects is made more difficult by the problem of when to define onset. In 1965 hard-core pornographic magazines became readily available. In 1969 the sale of pornographic pictures to those 16 or older was legalized, etc. Kutchinsky's presentation is a model of good quasi-experimental analysis in its careful searching out of other relevant data to evaluate plausible rival hypotheses.

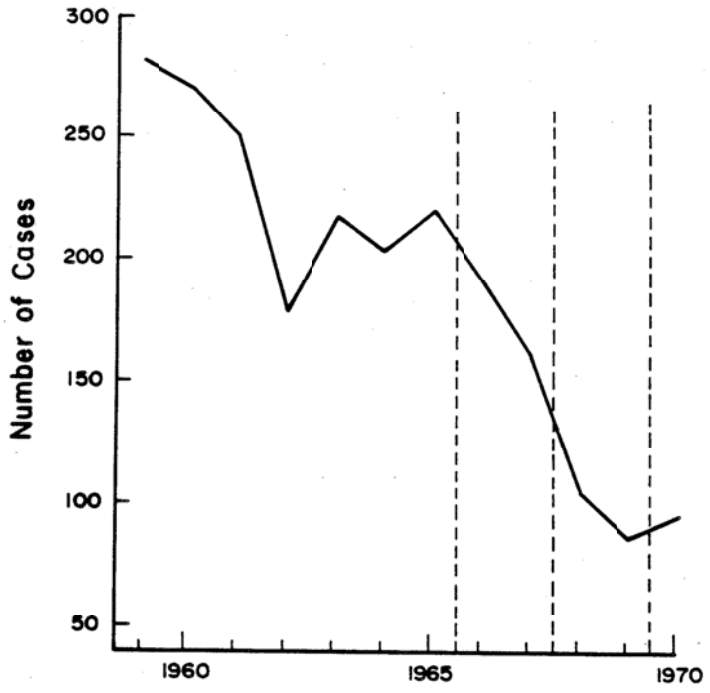


Figure 3. Child molestation (offenses against girls) in the city of Copenhagen. (From Kutchinsky's 1973 data.)

to the police in Denmark during the period 1948-1970. (Kutchinsky, 1973, p.164.)

Figure 4 shows the impact in Romania of the October 1966 population policy change which greatly restricted the use of abortion, reduced the availability of contraceptives, and provided several new incentives for large families. (David & Wright, 1971, David, 1970.) The combined effect is clear and convincing, with the change in the abortion law probably the main factor, particularly for the July-September 1967 peak. Presumably the subsequent decline represents a shift to other means of birth control. While clear visually, the data offer problems for the application of tests of significance. The strong seasonal trend rules out the application of the best statistical models (Glass, Willson, & Gottman, 1972), and there are not enough data points plotted here upon which to base a good seasonal adjustment. The point of onset for computing purposes is also ambiguous. Is it October 1, 1966, or six months later as per the prior rule permitting abortions in the first three months? Or nine months later? A shift to annual data obviates these two problems, but usually there are too few years or too many other changes to permit the use of tests of significance. Figure 5 shows annual data and also provides an opportunity to look for the effect of legalizing abortion in 1957. This occurred at a time when the rate of use of all means of birth control, including abortion, was increasing, and there is no graphic evidence that the 1957 law accelerated that trend. In other data not presented here, there is illustrated a chronic methodological problem with this design: Social systems react to abrupt changes by using all of the discretionary decision points to minimize that change. Thus the abrupt onset of the October 1966 decrees also produced an immediate increase in stillbirths, many of which were no doubt substitutes for the newly outlawed abortions. Such compensation was, however, too minimal to prevent an immediate increase in births.

Figure 4. Monthly birth rates per 1,000 population, Romania: 1966-1970. (David & Wright, 1971, p.207.)

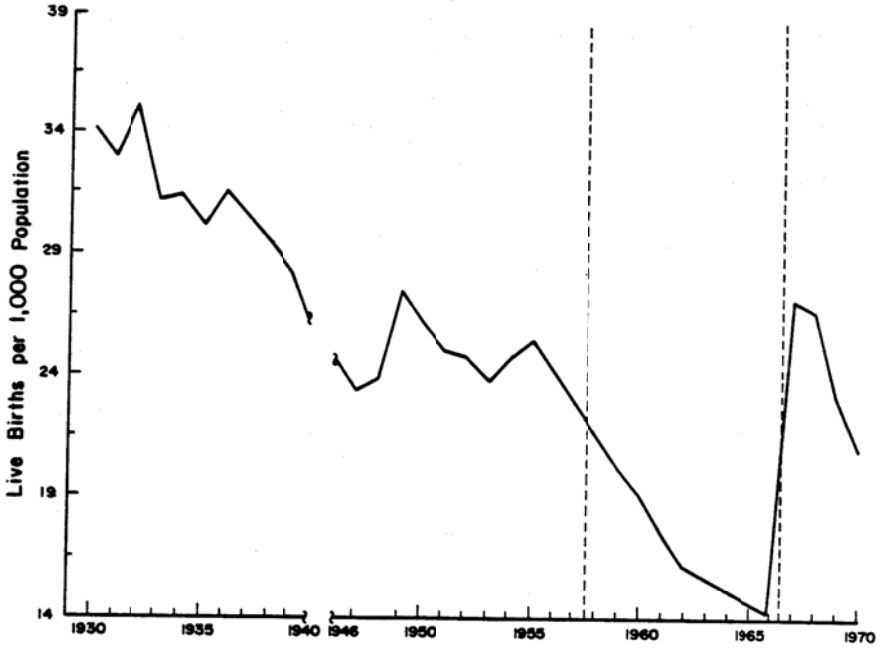


Figure 5. Live Births in Romania (1930-1970, excluding 1941-1945): total.

Figure 6 shows the effect of the British Breathalyser crackdown of 1967, illustrated here more dramatically than it has yet appeared in any British publication. The British Ministry of Transport dutifully reported strong results during the year following. Their mode of report was in terms of the percentage of decline in a given month compared with the same month one year earlier. This is better than total neglect of seasonal effects, but it is an inefficient method because unusual "effects" are often due to much to the eccentricity of the prior period as to that of the current one. It is also precludes presentation of the over-all picture. The newspapers duly noted the success, but interest soon faded, and today most British social scientists are unaware of the program's effectiveness. In figure 6 the data have been adjusted to attempt to correct for seasonal trend, uneven number of days per month, uneven number of weekends per month, and, for the month of October 1969, the fact that the crackdown did not begin until October 9. All of these adjustments have problems and alternative solutions. In this particular case, the effects are so strong that any approach would have shown them, but in many instances this will not be so. The data on commuting hours serves as a control for weekend nights.

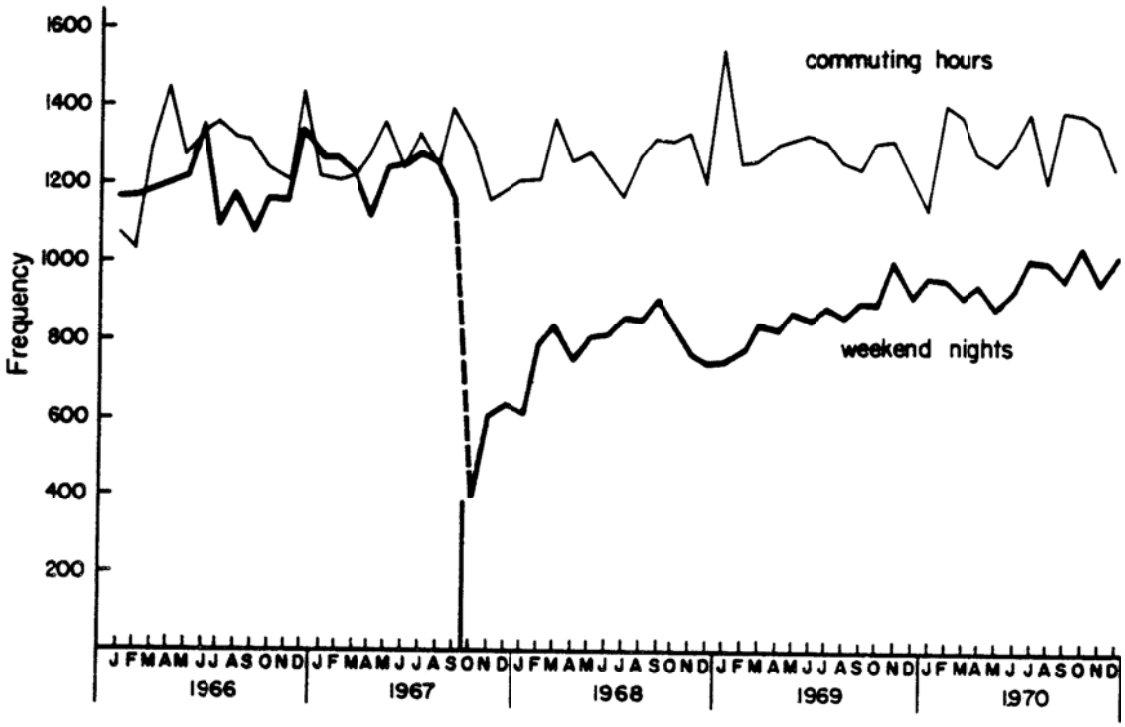


Figure 6. British traffic casualties (fatalities plus serious injuries) before and after the British Breathalyser crackdown of October 1967, seasonally adjusted. (Ross, 1973).

Figure 7 shows data from Baldus (1973) on the substantial effects of a law that Baldus believes to be evil just because it is effective. This law requires that, when a recipient of old age assistance (charity to the poor from the government) dies and leaves money or property, the government must be repaid. In our capitalist ideology, shared even by the poor, many old people will starve themselves just to be able to leave their homes to their children. Baldus has examined the effects of such laws in some 40 cases where states have initiated them and in some 40 other cases where states have discontinued them. In each case, he has sought out nearby, comparable states that did not change their laws to use as comparisons. One such instance is shown in Figure 7.

Figure 8 is a weak example of a time-series study because it has so few time periods. It has a compensatory strength because the several comparison groups are themselves constituted on a quantitative dimension. I include it primarily because it seems to indicate that the U.S. Medicaid legislation of 1964 has had a most dramatic effect on the access to medical attention of the poorest group of U.S. citizens.

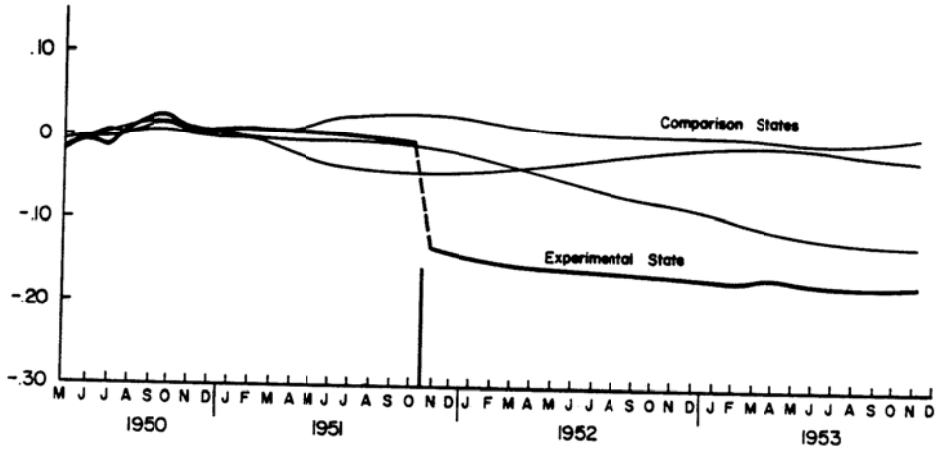


Figure 7. Effect of introducing a law in State A requiring repayment of welfare costs from the deceased recipient's estate on the old age assistance case loads. Modified from Baldus (1973, p. 204). Monthly data, with all values expressed as a percentage of the case load 18 months prior to the change in the law.

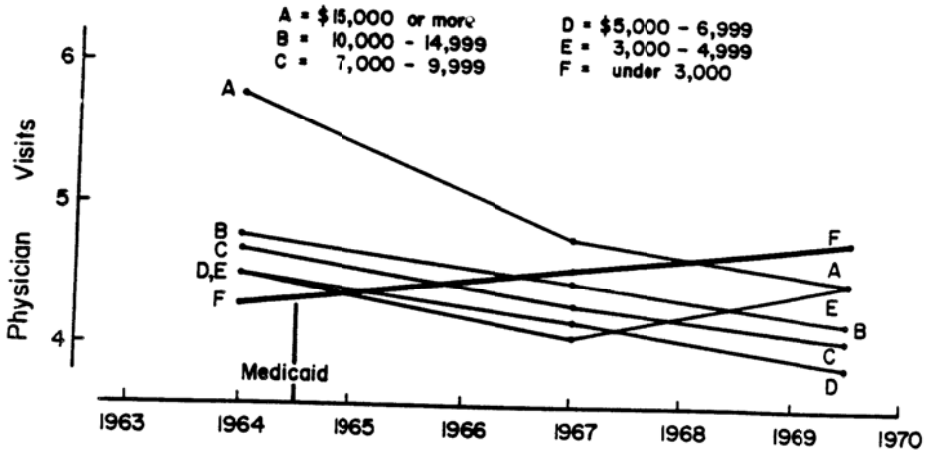


Figure 8. Effect of Medicaid on number of visits to physician per year by persons in low-income families. (From Lohr, 1972; Wilder, 1972, p.5, Table B.)

The interrupted time-series design is of the very greatest importance for program evaluation. It is available where the new program affects everyone and where, therefore, a proper control group can usually not be constituted. If comparison group data are available, it ranks as the strongest of all quasi-experimental designs (Campbell & Stanley, 1966). It can often be reconstructed from archival data. Graphically presented, it is really understood by administrators and legislators. Therefore, it is well worth the maximum of technical development. Following is a brief list of its methodological problems as we have encountered them.

1. Tests of significance are still a problem. Ordinary least squares estimation is usually inapplicable because of autoregressive error; therefore moving-average models seem most appropriate. Glass, Willson, and Gottman (1972) have assembled the best approach, which build on the work of Box and Tiao (1965) and box and Jenkins (1970). These models require that systematic cycles in the data be absent, but all methods of removing them tend to under-adjust. They also require large number of time-points, and will sometimes fail to confirm an effect which is compelling visually as graphed. They will also occasionally find a highly significant impact where visual inspection shows none.

2. Removing seasonal trends remains a problem. Seasonal trends are themselves unstable and require a moving-average model. The month-to-month change coincident with the program change should not be counted as purely seasonal; thus the series has to be split at this point for estimating the seasonal pattern. Therefore, the parts of the series just before and just after the program initiation become series ends, and corrections for these are much poorer than for mid-series points. (Kepka, 1971; McCain, in preparation.)

3. There is a tendency for new administrations that initiate new programs to make other changes in the record-keeping system. This often makes changes in indicators uninterpretable (Campbell, 1969b, pp.414-415). Where possible, this should be avoided.

4. Where programs are initiated in response to an acute problem (e.g., sudden change for the worse in a social indicator), ameliorative effects of the program are confounded with “regression artifacts” due to the fact that in an unstable series, points following an extreme deviation tend to be closer to the general trend (Campbell, 1969b, pp.413-414).

5. Gradually introduced changes are usually impossible to detect by this design. If an administrator wants to optimize evaluability using this design, program initiation should be postponed until preparations are such that it can be introduced abruptly. The British Breathalyser crackdown exemplifies this optimal practice (see Figure 6, above).

6. Because long series of observations are required, we tend to be limited to indicators that are already being recorded for other purposes. While these are often relevant (e.g., births and deaths) and while even the most deliberately designed indicators are never completely relevant this is a serious limitation. Particularly lacking are reports on the participants’ experiences and perceptions. On the other hand, it seems both impossible and undesirable to attempt to anticipate all future needs and to initiate bookkeeping procedures for them. Some intermediate compromise is desirable, even at the expense of adding to the forms to be filled

out and the records to be kept. For institutional settings, it would be valuable to receive from all participants "Annual Reports for Program Evaluation" (Gordon & Campbell, 1971). In educational settings teachers, students, and parents could file such a report. Note that at present the school system records how the pupil is doing but never records the pupil's report on how the school is doing. Teachers are annually rated for efficiency but never get a chance to systematically rate the policies they are asked to implement. Some first steps in this direction are being explored. (Weber, Cook, & Campbell, 1971; Anderson, 1973). In the U.S. social welfare system, both social worker and welfare recipient would be offered the opportunity to file reports (Gordon & Campbell, 1971). All ratings would be restricted to the evaluation of programs and policies, not persons, for reasons to be discussed below.

Regression Adjustments as Substitutes for Randomization

The commonest evaluation design in U.S. practice consists in administering a novel program to a single intact institution or administrative unit, with measures before and after. While this leaves much to be desired in the way of controls, it is still informative enough to be worth doing. Almost as frequently, this design is augmented by the addition of comparison group which is also measured before and after. This is typically another intact social unit which does not receive the new program and is judged comparable in other respects. It usually turns out that these two groups differ even before the treatment, and a natural tendency is to try to adjust away the difference. In statistical practice in the U.S. today, the means by which this is done are, in my opinion, almost always wrong. What has happened is that a set of statistical tools developed for and appropriate to prediction are applied to causal inference purposes for which they are inappropriate. Regression analysis, multivariate statistics, covariance analysis are some of the names of the statistical tools I have in mind. Whether from educational statistics or economics, the choice of methods seems to be the same. The economists have a phrase for the problem "error in variables" or, more specifically, "error in independent variables." But while theoretically aware of the problem, they are so used to regarding their indicators as essentially lacking in error that they neglect the problem in practice. What they forget is that irrelevant systematic components of variance create the same problem as does random error, leading to the same bias of underadjustment. Note that the presence of error and unique variance has a systematic effect, i.e. operates as a source of bias rather than as a source of instability in estimates. This fact, too, the economists and other neglect. Thus efforts to correct for pretreatment differences by "regression adjustments" on the means or by "partialing out" pretest differences or by covariance adjustments all lead to underadjustment unless the pretest (or other covariate) is a perfect measure of what pretest and posttest have in common. The older technique of using only cases matched on pretest scores is well known to produce "regression artifacts" (Thorndike, 1942; Campbell & Stanley, 1966). Covariance turns out to produce the same bias, the same degree of underadjustment only with greater precision (Lord, 1960, 1969; Porter, 1967; Campbell & Erlebacher, 1970), and also for multiple regression and partial

correlation (e.g., Cook & Campbell, 1975). Essentially the same problem emerges in ex post facto studies where, although there is no pretest, other covariates are available for adjustment. A common version of the problem occurs where some persons have received a treatment and there is a larger population of untreated individuals from which "controls" are sought and a comparison group assembled.

In U.S. experience it has become important to distinguish two types of setting in which this type of quasi-experimental design and these types of adjustments are used, since the social implications of the underadjustment are opposite. On the one hand, there are those special opportunity programs like university education which are given to those who need them least, or as more usually stated, who deserve them most or who are more likely to be able to profit from them. Let us call these "distributive" programs in contrast with the "compensatory" programs, or those special opportunities given to those who need them most.

For the regressive programs, the treatment group will usually be superior to the control group or the population from which the quasi-experimental controls are chosen. In this setting the inevitable underadjustment due to unique variance and error in the pretest and/or other covariates (the "regression artifacts") works to make the treatment seem effective if it is actually worthless and to exaggerate its effectiveness in any case. For most of us, this seems a benign error, confirming our belief in treatments we know in our hearts are good. (It may come as a surprise, but the U.S. Sesame Street preschool educational television program is "distributive," in that children from better-educated parents watch it more.) (Cook, et al., 1975.)

For compensatory programs usually, although not always, the control group start out superior to the treatment group, or are selected from a larger population whose average is superior. In this setting, the biases of underadjustment, the regression artifacts, are in the direction of underestimating program effectiveness and of making our program seem harmful when they are merely worthless. This quasi-experimental research setting has occurred for our major evaluations of compensatory education programs going under the names of Head Start, Follow Through, Performance Contracting, Job Corps (for unemployed young men), and many others. In the major Head Start evaluation (Cicirelli, 1969; Campbell & Erlebacher, 1970), this almost certainly accounts for the significantly harmful effects shown in the short three month, ten-hours-a-week program. I am persuaded that the overwhelming prevalence of this quasi-experimental setting and adjustment procedures is one of the major sources of the pessimistic record for such programs of compensatory education efforts. The very few studies in compensatory education which have avoided this problem by random assignment of children experimental and control conditions have shown much more optimistic results.

In the compensatory education situation, there are several other problems which also work to make the program look harmful in quasi-experimental studies. These include tests that are too difficult, differential growth rates combined with age-based, grade-equivalent, absolute, or raw scores, and the fact that test reliability is higher for the post-test than for the pretest, and higher for the control group than for the experimental (Campbell,

1973). These require major revisions of our test score practice. When various scoring models are applied to a single population on a single occasion, all scoring procedures correlate so highly that one might as well use the simplest. But when two groups that differ initially are measured at two different times in a period of rapid growth, our standard test score practices have the effect of making the gap appear to increase, if, as is usual, test reliability is increasing. The use of a correction for guessing becomes important. The common model that assumes "true score" and "error" are independent needs to be abandoned, substituting one that sees error and true score negatively correlated across persons (the larger the error component, the smaller the true score component).

Problems with Randomized Experiments

The focal example of a good social experiment in the U.S. today is the New Jersey Negative Income Tax Experiment. (Watts & Rees, 1973; The Journal of Human Resources, Vol. 9, No. 2, Spring 1974; Kershaw's paper in this volume; Kershaw, 1972; Kershaw & Fair, 1973). This is an experiment dealing with a guaranteed annual income as an alternative to present U.S. welfare systems. It gets its name from the notion that when incomes fall below a given level, the tax should become negative, that is, the government should pay the citizen rather than the citizen paying a tax to the government. It also proposes substituting income-tax like procedures for citizen reports of income in place of the present social worker supervision. In this experiment some 600 families with a working male head of household received income support payments bringing their income up to some level between $3,000 and $4,000 per year for a family of four, under one of eight plans which differ as to support level and incentive for increasing own earnings. Another 600 families received no income support but cooperated with the quarterly interviews. The experiment lasted for three years, and preliminary final results are now available. This study when completed will have cost some $8,000,000 of which $3,000,000 represented to participants and necessary administrative costs, and $5,000,000 research costs, the costs of program evaluation. Before this study was half completed, three other negative income tax experiments were started, some much bigger (rural North Carolina and Iowa, Gary, and Seattle and Denver). The total U.S. investment for these experiments now totals $65,000,000. It is to me amazing, and inspiring, that our nation achieved, for a while at least, this great willingness to deliberately "experiment" with policy alternatives using the best of scientific methods.

This requires a brief historical note. The key period is 1964-68. L.B. Johnson was President and proclaimed a "Great Society" program and a "War on Poverty." In Washington, D.C. administrative circles, the spirit of scientific management (the PPBS I've already criticized and will again) had already created a sophisticated interest in hard-headed program evaluation. Congress was already writing into its legislation for new programs the requirement that 1% (or some other proportion) of the program budget be devoted to evaluation of effectiveness. In a new agency, the Office of Economic Opportunity, a particularly creative and dedicated group of young economist was recruited, and these scientist-evaluators were given an especially strong role in

guiding over-all agency policy. This OEO initiated the first two of the Negative Income Tax experiments. (Two others were initiated from the Department of Health, Education and Welfare.) Under the first Nixon administration, 1968-72, OEO programs were continued, although on a reduced scale. Under the second Nixon administration, OEO itself was dismantled, although several programs were transferred to different agencies. I believe that all four of the Negative Income Tax Experiments are still being carried out much as planned. Initiation of new programs has not entirely ceased, but it is greatly reduced. My general advice to my fellow U.S. social scientists about the attitude they should take toward this historical period is as follow: Let us use this experience so as to be ready when this political will returns again. In spite of the good example provided by the New Jersey Negative Income Tax Experiment, over all we were not ready last time. Competent evaluation researchers were not available when Model Cities Programs, Job Corps Programs, etc. went to their local universities for help. Perhaps 90% of the funds designated for program evaluation were wasted; at any rate, 90% of the programs came out with no interpretable evidence of their effectiveness. The available evaluation experts grossly overestimated the usefulness of statistical adjustments as substitutes for good experimental design, including especially randomized assignment to treatment.

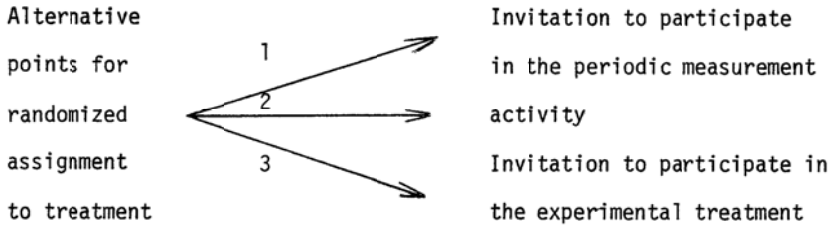
In this spirit I would like to use the experience of the New Jersey Negative Income Tax Experiment to elucidate the methodological problems remaining to be solved in the best of social experiments. In this spirit, my comments are apt to sound predominately critical. My over-all attitude, however, is one of highest approval. Indeed, in lectures in the U.S. I often try to shock audiences with the comment that the New Jersey experiment is the greatest example of applied social science since the Russian Revolution.

The major finding of the NJNITE is that income guarantees do <u>not</u> reduce the effective work effort of employed poor people. This finding, if believed, removes the principal argument against such a program−-for on a purely cost basis, it would be cheaper than the present welfare system unless it tempted many persons now employed to cease working. The major methodological criticisms of the study are focused on the credibility of the belief that this "laboratory" finding would continue to hold up if the support program became standard, permanent U.S. policy. These are questions of "external validity" (Campbell & Stanley, 1966) or of "construct validity," as Cook (Cook & Campbell, 1975) applies the term developed initially for measurement theory. Two specific criticisms are prominent: One, there was a "Hawthorne Effect" or a "Guinea-pig Effect." The experimental families knew they were the exceptional participant in an artificial arrangement and that the spotlight of public attention was upon them. Therefore, they behaved in a "good" industrious, respectable way, producing the results obtained. Such motivation would be lacking once the program was universal. Two features in the NJNITE implementation can be supposed to accentuate this. There was publicity about the experiment at its start, including television interviews with selected experimental subjects; and the random selection was by families rather than neighborhoods, so each experimental family was surrounded by equally poor neighbors, who were not getting this beneficence. The second common criticism, particularly among economists,

might be called the time-limit effect. Participants were offered the support for exactly three years. It was made clear that the experiment would terminate at that time. This being the case, prudent participants would hang on to their jobs unless they could get better ones, so that they would be ready for a return to their normal financial predicament.

It should be recognized that these two problems are in no way specific to randomized experiments, and would also have characterized the most casual of pilot programs. They can only be avoided by the evaluation of the adoption of NIT as a national policy. Such an evaluation will have to be quasi-experimental, as by time-series, perhaps using several Canadian cities as comparisons. Such evaluations would be stronger on external, construct validity, but weaker on internal validity. It is characteristic of our national attitudes, however, that this quasi-experimental evaluation is not apt to be done well, if at all–-once we've chosen a policy we lose interest in evaluating it. Had the NJNITE shown a reduction in work effort, national adoption of the policy would have been very unlikely. For this reason alone, it was well worth doing and doing well.

The details of the experiment draw attention to a number of problems of method that need detailed attention from creative statisticians and social psychologists. These will only be mentioned here, but are being treated more extensively elsewhere (Riecken, et al., 1974). The issue of the unit of randomization has already been raised in passing. Often there is a choice of randomizing larger social units than persons or families–residential blocks, census tracts, classrooms, schools, etc. are often usable. For reasons of statistical efficiency, the smaller, more numerous units are to be preferred, maximizing the degrees of freedom and the efficacy of randomization. But the use of larger units often increases construct validity. Problems of losses due to refusals and later attrition interact with the choice of the stage of respondent recruitment at which to randomize. NJNITE used census statistics on poverty areas and sample survey approaches to locate eligible participants. Rethinking their problem shows the usefulness of distinguishing two types of assent required, for measurement and for treatment; thus two separate stages for refusals occur. There emerge three crucial alternative points at which randomization could be done:



In NJNITE, alternative 1 was employed. Subsequently control group respondents were asked to participate in the measurement, and experimental subjects were asked to participate in both measurement and treatment. As a result, there is the possibility that the experimental group contains persons who would not have put up with the bother of measurement had they by change been invited into the control group. Staging the invitations separately, and randomizing from among those who had agreed to the survey (i.e., to the control group condition) would have ensured comparability. In NJNITE there were some refusals to experimental treatment because of unwillingness to accept charity. This products dissimilarity, but the bias due to such differential refusal can be estimated if those refusing treatment continue in the measurement activity. Alternative 2 is the point we would now recommend.

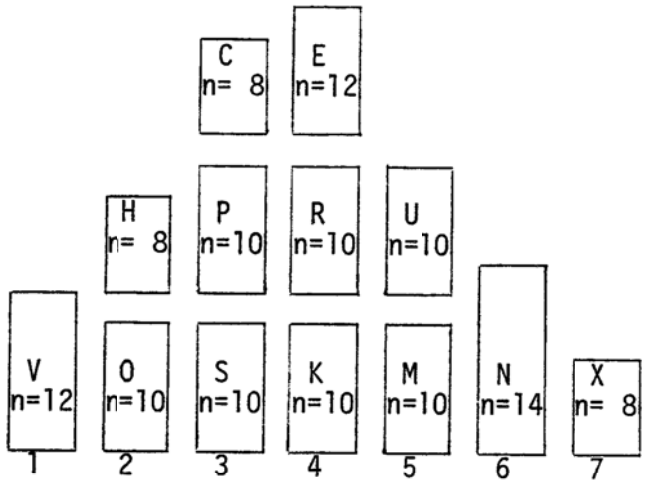
One could consider deferring the randomizing still further, to alternative 3. Under this procedure, all potential participants would be given a description of each of the experimental conditions and his chances for each. He would then be asked to agree to participate no matter what outcome he drew by chance. From those who agreed to all this, randomized assignment would be made. This alternative is one that is bound to see increased use. The opportunity for differential refusal is minimized (though some will still refuse when they learn their lot). This seems to maximize the "informed consent" required by the U.S. National Institutes of Health for all of the medical and behavioral research funded by them. The U.S. Social Science Research Council's Committee on Experimentation as a Method for Planning and Evaluating Social Programs (Riecken, et al., 1974) failed to recommend alternative 3 however. In net, they judge informed consent to be adequately achieved when the participant is fully informed of the treatment he is to receive. Informing the control group participants of the benefits that others were getting and that they almost got would have caused discontent, and have made the control treatment an unusual experience rather than the representative of the absence of treatment. The tendency of control subjects to drop out more frequently than experimental subjects would have been accentuated, and thus this approach to greater comparability would be in the end self-defeating. These are reasonable arguments on both sides. More discussion and research are needed on the problem.

Attrition and in particular differential attrition become major problems on which the work of inventive statisticians is still needed. In the NJNITE, attrition rates over the three-year period range from 25.3% in the control group to only 6.5% in the most remunerative

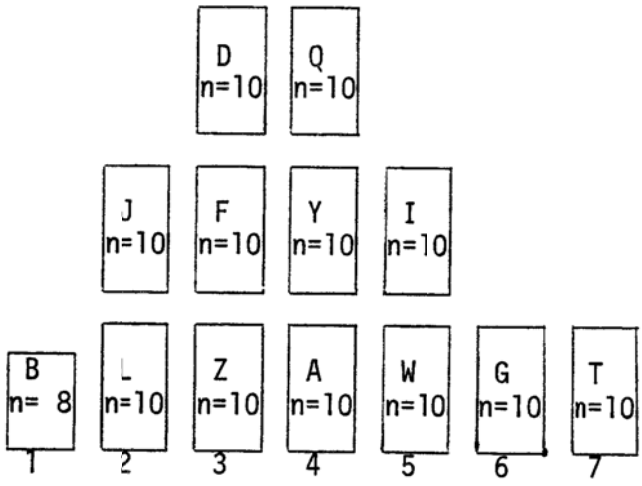
experimental group. These differences are large enough to create pseudo effects in post-test values. The availability of pretest scores provides some information on the probable direction of bias, but covariance on these values underadjusts and is thus not an adequate correction. Methods for bracketing maximum and minimum biases under various specified assumption need to be developed. Where there is an encompassing periodic measurement framework that still retains persons who have ceased cooperating with the experiment, other alternatives are present that need developing.

Such measurement frameworks appropriate to NJNITE would include the Social Security Administration's records on earnings subject to social security withholding and claims on unemployment insurance, the Internal Revenue Service records on withholding taxes, hospitalization insurance records, etc. These records are occasionally usable in evaluation research (e.g., Levenson & McDill, 1966; Bauman, David & Miller, 1979; Fischer, 1972, Heller, 1972) but the facilities for doing so are not adequately developed, and the concept of such usage may seem to run counter to the current U.S. emphasis on preserving the confidentiality of administrative records (e.g., Reubhausen & Brim, 1965; Sawyer & Schechter, 1968; Goslin, 1970; Miller, 1971; Westin, 1967; Wheeler, 1969). Because research access to administrative files would make possible so much valuable low cost follow-up on program innovations, this issue is worthy of discussion somewhere in this paper. For convenience, if not organizational consistency, I will insert the discussion here.

There is a way to statistically relate research data and administrative records without revealing confidential data on individual (Schwartz & Orleans, 1967; Campbell, Boruch, Schwartz & Steinberg, 1975; Boruch & Campbell, 1974). Let us call this "mutually insulated interfile exchange." It requires that the administrative file have the capacity for internal statistical analysis of its own records. Without going into detail, I would nonetheless like to communicate the nub of the idea. Figure 9 shows a hypothetical experiment with one experimental group and one control group. In this case there are enough cases to allow a further breakdown by socio-economic level. From these data some 26 lists are prepared ranging from 8 to 14 persons in length. These lists are assigned designations at random (in this case A to Z) so that list designation communicates no information to the administrative file. The list provides the name of the person, his Social Security number, and perhaps his birth date and birth place. These lists are then turned over to the administrative file which deletes one person at random from each list, retrieves the desired data from the files on each of the others for whom it is available, and computes mean, variance, and number of cases with data available for each list for each variable. These values for each list designation are then returned to the evaluation researchers who reassemble them into statistically meaningful composites and then compute experimental and control group means and variances, correlations, interactions with socio-economic level, etc. Thus neither the research file nor the administrative file has learned individual data from the other file, yet the statistical estimates of program effectiveness can be made. In the U.S. it would be a great achievement were this facility for program evaluation to become feasible for regular use.



SES Level: Experimental Group



SES Level: Control Group

Figure 9. Hypothetical data from two treatment groups in a social experiment, grouped by SES level and given coded list designators A through Z.

To return to attrition problems in randomized experiments, not only do we need new statistical tools for the attrition problem, we also need social-psychological inventions. In long-term experiments such as that of Ikeda, Yinger, and Laycock (1970 in which a university starts working with underprivileged twelve-year-old during the summers, trying to motivate and guide their high school activities (ages 14 to 18) to be university preparatory, two of the reasons why attrition is so differential are that more recent home addresses are available for the experimentals (who have been in continuous contact) and that the experimentals answer more follow-up inquires because of gratitude to the project. This suggests that controls in long-term studies might be given some useful service on a continuing basis–less than the experimentals but enough to motivate keeping the project informed of address changes and cooperating with follow-up inquires (Ikeda, et al., 1970). If one recognizes that comparability between experimental and control groups is more important than completeness per se, it becomes conceivable that comparability might be improved by deliberately degrading experimental group data to the level of the control group. In an exploration of the possibility, Ikeda, Richardson, and I (in preparation) are conducting an extra follow-up of this same study using five-year-old addresses, a remote unrelated inquiring agency, and questions that do not refer specifically to the Ideka, Yinger and Laycock program. (I offer this unpromising example to communicate my feeling that we need wide-ranging explorations of possible solutions to this problem.)

It is apparent that through refusal and attrition, true experiments tend to become quasi-experiments. Worse than that, starting with randomization makes the many potential sources of bias more troubling in that it focuses awareness on them. I am convinced, however, that while the biases are more obvious, they are in fact considerably less than those accompanying more casual forms of selecting comparison groups. In addition, our ability to estimate the biases is immeasurably greater. Thus, we should, in my judgment, greatly increase our use of random assignment, including in regular admissions procedures in ongoing programs, having a surplus of applicants. To do this requires that we develop practical procedures and rationales that overcome the resistance to randomization met with in those settings. Just to communicate to you that there are problems to be solved in these areas, I will briefly sketch several of them.

Administrators raise many objections to randomization (Conner, 1974). While at one time lotteries were used to “let God decide,” now a program administrator feels he is “playing God” himself when he uses a randomization procedure, but not when he is using his own incompetent and partisan judgment based on inadequate and irrelevant information (Campbell, 1971). Participants also resist randomization, though less so when they themselves choose the capsule from the lottery bowl than when the administrator does the randomization in private (Wortman, et al., 1974). Collecting a full list of eligible applicants and then randomizing often causes burdensome delays, and it may be better to offer a 50-50 lottery to each applicant as he applies, closing off all applications when the program openings have been filled, at which time the controls would be approximately the same in number. For settings like specially equipped old people’s homes, the control group ceases

to be representative of non-experimental conditions if those losing the lottery are allowed to get on the waiting list–waiting for an opening forestalls normal problem-solving. For such settings, a three-outcome lottery is suggested: (1) admitted; (2) waiting list; (3) rejected. Group 3 would be the appropriate control. For agencies having a few new openings each week or so, special "trickle processing" procedures are needed rather than large-batch randomization. Where the program is in genuinely short supply, one might think that the fact that most people were going without it would reconcile control group subjects to their lot; however, experimental procedures including randomization and measurement may create an acute focal deprivation, making control status itself an unusual treatment. This may result in compensatory striving or low moral (Cook & Campbell, 1975).

Regression-Discontinuity Design

The arguments against randomizing admissions to an ameliorative program (one with more eligible applications than there is space for) include the fact that there are degrees of eligibility, degrees of need or worthiness, and that the special program should go to the most eligible, needy, or worthy. If eligibility can be quantified (e.g., through ranks, ratings, scores, or composite scores) and if admission for some or all of the applicants can be made on the basis of a strict application of this score, then a powerful quasi-experimental design, Regression-discontinuity, is made possible. General explanation and discussion of administrative details are to be found in Campbell (1969b) and Riecken, et al. (1974). Sween (1971) has provided appropriate tests of significance. Goldberger (1971), working from an econometric background, has made an essentially equivalent recommendation.

The application of quantified eligibility procedures usually involves at least as great a departure from ordinary admission procedures as does randomization. Developing specific routines appropriate to the setting is necessary. But once instituted, their economic costs would be low and would be more than compensated for by increased equity of the procedures. Resistance, however, occurs. Administrators like the freedom to make exceptions eve to the rules they themselves have designed. "Validity" or "reliability" for the quantified eligibility criterion is not required; indeed, as it approaches zero reliability, it becomes the equivalent of randomization.

Political/Methodological Problems

Resistance to Evaluation

In the U.S., one of the pervasive reasons why interpretable program evaluations are so rare is the widespread resistance of institutions and administrators to having their programs evaluated. The methodology of evaluation research should include the reasons for this resistance and ways of overcoming it.

A major source of this resistance in the U.S. is the identification of the administrator and the administrative unit with the program. An evaluation of a program under our political climate becomes an evaluation of the agency and its directors. In addition the machinery for evaluating programs can be used deliberately to evaluate administrators. Combined with this, there are a number of

factors that lead administrators to correctly anticipate a disappointing outcome. As Rossi (1969) has pointed out, the special programs that are the focus of evaluation interests have usually been assigned the chronically unsolvable problems, those on which the usually successful standard institutions have failed. This in itself provides a pessimistic prognosis. Furthermore, the funding is usually inadequate, both through the inevitable competition of many worthy causes for limited funds and because of a tendency on the part of our legislatures and executives to generate token or cosmetic efforts designed more to convince the public that action is being taken than to solve the problem. Even for genuinely valuable programs, the great effort required to overcome institutional inertia in establishing any new program leads to grossly exaggerated claims. This produces the "overadvocacy trap" (Campbell, 1969b, 1971), so that even good and effective programs fall short of what has been promised, which intensifies fear and evaluation.

The seriousness of these and related problems can hardly be exaggerated. While I have spent more time in this presentation on more optimistic cases, the preceding paragraph is more typical of evaluation research in the U.S. today. As methodologists, we in the U.S. are called upon to participate in political process in efforts to remedy the situation. But before we do so, we should sit back in our armchairs in our ivory towers and invent political/organizational alternatives which would avoid the problem. This task we have hardly begun, and it is one in which we may not succeed. Two minor suggestions will illustrate. I recommend that we evaluation research methodologists should refuse to use our skills in ad hominem research. While the expensive machinery of social experimentation can be used to evaluate persons, it should not be. Such results are of very limited generalizability. Our skills should be reserved for the evaluation of policies and programs that can be applied in more than one setting and that any well-intentioned administrator with proper funding could adopt. We should meticulously edit our opinion surveys to that only attitudes toward program alternatives are collected and such topics as supervisory efficiency excluded. This prohibition on ad hominem research should also be extended to program clients. We should be evaluating not students or welfare recipients but alternative policies for dealing with their problems. It is clear that I feel such a prohibition is morally justified. But I should also confess that in our U.S. settings it is also recommended our of cowardice. Program administrators and clients have it in their power to sabotage our evaluation efforts, and they will attempt to do so if their own careers and interests are at stake. While such a policy on our part will not entirely placate administrators' fears, I do believe that if we conscientiously lived up to it, it would initiate a change toward a less self-defeating political climate.

A second recommendation is for advocates to justify new programs on the basis of the seriousness of the problem rather than the certainty of any one answer and combine this with the emphasis on the need to go on to other attempts at solution should the first one fail (Campbell, 1969b). Shaver and Staines (1971) have challenged this suggestion, arguing that for an administrator to take this attitude of scientific tentativeness constitutes a default of leadership. Conviction, zeal, enthusiasm, faith are required for any effective effort to change

traditional institutional practice. To acknowledge only a tentative faith in the new program is to guarantee a half-hearted implementation of it. But the problem remains; the overadvocacy trap continues to sabotage program evaluation. Clearly, social-psychological and organization-theoretical skills are needed.

Corrupting Effect of Quantitative Indicators

Evaluation research in the U.S.A. is becoming a recognized tool for social decision-making. Certain social indicators, collected through such social science methods as sample surveys, have already achieved this status; for example, the unemployment and cost-of-living indices of the Bureau of Labor Statistics. As regular parts of the political decision process, it seems useful to consider them as akin to voting in political elections (Gordon & Campbell, 1971; Campbell, 1971). From this enlarged perspective, which is supported by qualitative sociological studies of how public statistics get created, I come to the following pessimistic laws (at least for the U.S. scene): The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. Let me illustrate these two laws with some evidence which I take seriously, although it is predominantly anecdotal.

Take, for example, a comparison between voting statistics and census data in the city of Chicago: Surrounding the voting process, there are elaborate precautionary devices designed to ensure its honesty; surrounding the census-taking process, there are few, and these could be easily evaded. Yet, in our region, the voting statistics are regarded with suspicion while the census statistics are widely trusted (despite underenumeration of young adult, black males). I believe this order of relative trust to be justified. The best explanation for it is that votes have continually been used–have had real implications as far as jobs, money, and power are concerned–and have therefore been under great pressure from efforts to corrupt. On the other hand, until recently our census data were unused for political decision-making. (Even the constitutional requirement that electoral districts be changed to match population distribution after every census was neglected for decades.)

Another example: In the spirit of scientific management, accountability, the PPBS movement, etc., police departments in some jurisdictions have been evaluated by "clearance rates," i.e., the proportion of crimes solved, and considerable administrative and public pressure is generated when the rate is low. Skolnick (1966) provide illustrations of how this pressure has produced both corruption of the indicator itself and a corruption of the criminal justice administered. Failure to record all citizens' complaints, or to postpone recording them unless solved, are simple evasions which are hard to check, since there is no independent record of the complaints. A more complicated corruption emerges in combination with "plea-bargaining." Plea-bargaining is a process whereby the prosecutor and court bargain with the prisoner and agree on a crime and a punishment to which the prisoner is willing to plead guilty, thus saving the cost and delays of a trial. While this is only a semilegal custom, it is probably not undesirable in most instances. However, combined with the clearance rate,

Skolnick finds the following miscarriage of justice. A burglar who is caught in the act can end up getting a lighter sentence the more prior unsolved burglaries he is willing to confess to. In the bargaining, he is doing the police a great favor by improving the clearance rate, and in return, they provide reduced punishment. Skolnick believes that in many cases the burglar is confessing to crimes he did not in fact commit. Crime rates are in general very corruptible indicators. For many crimes, changes in rates are a reflection of changes in the activity of the police rather than changes in the number of criminal acts (Gardiner, 1969; Zeisel, 1971). It seems to be well documented that a well-publicized, deliberate effort at social change–Nixon's crackdown on crime–had as its main effect the corruption of crime-rate indicators (Seidman & Couzens, 1972; Morrissey, 1972; Twigg, 1972), achieved through underrecording and by downgrading the crimes to less serious classifications.

For other types of administrative records, similar use-related distortions are reported (Kitsuse & Cicourel, 1963; Garfinkel, 1967). Blau (1963) provides a variety of examples of how productivity standards set for workers in government offices distort their efforts in ways deleterious to program effectiveness. In an employment office, evaluating staff members by the number of cases handled led to quick, ineffective interviews and placements. Rating the staff by the number of persons placed led to concentration of efforts on the easiest cases, neglecting those most needing the service, in a tactic know as "creaming" (Miller, et al., 1970). Ridgeway's pessimistic essay on the dysfunctional effects of performance measures (1956) provides still other examples.

From the experimental program in compensatory education comes a very clear-cut illustration of the principle. In the Texarkana "performance contracting" experiment (Stake, 1971), supplementary teaching for undereducated children was provided by "contractors" who came to the schools with special teaching machines and individualized instruction. The corruption pressure was high because the contractors were to be paid on the basis of the achievement test score gains of individual pupils. It turned out that the contractors were teaching the answers to specific test items that were to be used on the final play-off testing. Although they defended themselves with a logical-positivist, operational-definitionalist argument that their agreed-upon goal was defined as improving scores on that one test, this was generally regarded as scandalous. However, the acceptability of tutoring the students on similar items from other tests is still being debated. From my own point of view, achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways. (Similar biases of course surround the use of objective tests in courses or as entrance examinations.) In compensatory education in general there are rumors of other subversions of the measurement process, such as administering pretests in a way designed to make scores as low as possible so that larger gains will be shown on the post test, or limiting treatment to those scoring lowest on the pretest so that regression to the mean will provide apparent gains. Stake (1971) lists still

other problems. Achievement tests are, in fact, highly corruptible indicators.

That this serious methodological problem may be a universal one is demonstrated by the extensive U.S.S.R. literature (reviewed in Granick, 1954; and Berliner, 1957) on the harmful effects of setting quantitative industrial production goals. Prior to the use of such goals, several indices were useful in summarizing factory productivity–e.g., monetary value of total product, total weight of all products produced, or number of items produced. Each of these, however, created dysfunctional distortions of production when used as the official goal in terms of which factory production was evaluated. If monetary value, then factories would tool up for and produce only one product to avoid the production interruptions of retooling. If weight, then factories would produce only their heaviest item (e.g., the largest nails in a nail factory). If number of items, then only their easiest item to produce (e.g., the smallest nails). All these distortions led to overproduction of unneeded items and underproduction of much needed ones.

To return to the U.S. experience in a final example. During the first period of U.S. involvement in Viet Nam, the estimates of enemy casualties put out by both the South Vietnamese and our own military were both unverifiable and unbelievably large. In the spirit of McNamara and PPBS, an effort was then instituted to substitute a more conservative and verifiable form of reporting, even if it underestimated total enemy casualties. Thus the "body count" was introduced, an enumeration of only those bodies left by the enemy on the battlefield. This became used not only for overall reflection of the tides of war, but also for evaluating the effectiveness of specific battalions and other military units. There was thus created a new military goal, that of having bodies to count, a goal that came to function instead of or in addition to more traditional goals, such as gaining control over territory. Pressure to score well in this regard was passed down from higher officers to field commanders. The realities of guerrilla warfare participation by persons of a wide variety of sexes and ages added a permissive ambiguity to the situation. Thus poor Lt. Calley was merely engaged in getting bodies to count for the weekly effectiveness report when he participated in the tragedy at My Lai. His goals had been corrupted by the worship of a quantitative indicator, leading both to a reduction in the validity of that indicator for its original military purposes, and a corruption of the social processes it was designed to reflect.

I am convinced that this is one of the major problems to be solved if we are to achieve meaningful evaluations of our efforts at planned social change. It is a problem that will get worse, the more common quantitative evaluations of social programs become. We must develop ways of avoiding this problem if we are to move ahead. We should study the social processes through which corruption is being uncovered and try to design social systems that incorporate these features. In the Texarkana performance-contracting study, it was an "outside evaluator" who uncovered the problem. In a later U.S. performance-contracting study, the Seattle Teachers' Union provided the watchdog role. We must seek out and institutionalize such objectivity-preserving features. We should also study the institutional form of those indicator systems, such as the census or the cost-of-living index in the U.S., which seem relatively immune to distortion. Many

commentators, including myself (1969b), assume that the use of multiple indicators, all recognized as imperfect, will alleviate the problem, although Ridgeway (1956) doubts this.

There are further problems that can be anticipated in the future. A very challenging group centers on the use of public opinion surveys, questionnaires, or attitude measures in program evaluation. Trends in the U.S. are such that before long, it will be required that all participants in such surveys, before they answer, will know the uses to which the survey will be put, and will receive copies of the results. Participants will have the right to use the results for their own political purposes. (Where opinion surveys are used by the U.S. Government, our present freedom of information statutes should be sufficient to establish this right now.) Under these conditions, using opinion surveys to evaluate local government service programs can be expected to produce the following new problems when employed in politically sophisticated communities such as we find in some of our poorest urban neighborhoods: There will be political campaigns to get respondents to reply in the particular ways the local political organizations see as desirable, just as there are campaigns to influence the vote. There will be efforts comparable to ballot-box stuffing. Interviewer bias will become even more of a problem. Bandwagon effects–i.e., conformity influence from the published results of prior surveys–must be anticipated. New biases, like exaggerated compliant, will emerge.

In my judgment, opinion surveys will still be useful if appropriate safeguards can be developed. Most of these are problems that we could be doing research on now in anticipation of future needs. (Gordon & Campbell, 1971 provide a detailed discussion of these problems in a social welfare service program evaluation setting.)

Summary Comment

This has been a condensed overview of some of the problems encountered in the U.S. experience with assessing the impact of planned social change. The sections of the paper dealing with the problems related to political processes have seemed predominantly pessimistic. While there are very serious problems, somehow the overall picture is not as gloomy as this seems. Note that the sections on time-series and on randomized designs contain success stories worthy of emulation. And many of the quasi-experimental evaluations that I have scolded could have been implemented in better ways–had the social science methodological community insisted upon it–within the present political system. There are, however, new methodological problems which emerge when we move experimentation out of the laboratory into social program evaluation. In solving these problems, we may need to make new social-organizational inventions.

References

Anderson, J.K. *Evaluation as a process in social systems*. Unpublished doctoral dissertation, Northwestern University, Department of Industrial Engineering & Management Sciences, 1973.

Baldus, D. C. Welfare as a loan: An empirical study of the recovery of public assistance payments in the United States. *Stanford Law Review*, 1973. 25.123-250. (No. 2)

Barnow, B. S. *The effects of Head Start and socioeconomic status on*

cognitive development of disadvantaged children. Unpublished doctoral dissertations, University of Wisconsin, Department of Economics, 1973. (252 pp.)

Bauman, R. A., David, M. H., & Miller, R. F. Working with complex data files: II. The Wisconsin assets and incomes studies archive. In R. L. Bisco (Ed.), *Data bases, computers, and the social sciences*. New York: Wiley-Interscience, 1970, 112-136.

Beck, B. Cooking the welfare stew. In R. W. Habenstein (Ed.), *Pathways to data: Field methods for studying ongoing social organizations*. Chicago: Aldine, 1979.

Becker, H. M., Geer, B., & Hughes, E. C. *Making the grade*. New York: Wiley, 1968

Becker, H. W. *Sociological work: Method and substance*. Chicago: Aldine, 1970.

Berliner, J. S. *Factory and manager in the U.S.S.R*. Cambridge Mass.: Harvard University Press, 1957.

Blau, P. M. *The dynamics of bureaucracy*. (Rev. Ed.) Chicago: University of Chicago Press, 1963.

Boruch, R. F., & Campbell, D. T. *Preserving confidentiality in evaluative social research: Intrafile and interfile data analysis*. Paper presented at the meeting of the American Association for the Advancement of Science, San Francisco, February/March 1974.

Box, G. E. P., & Tiao, G. C. A change in level of non-stationary time series. *Biometrika*, 1965, 52.181-192.

Box, G. E. P., & Jenkins, G. M. *Time-series analysis: Forecasting and control*. San Francisco: Holden Day, 1970.

Campbell, D. T. Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik*. New York: Holt, Rinehart & Winston, 1966, 81-106.

Campbell, D. T. Administrative experimentation, institutional records, and nonreactive measures. In J. C. Stanley & S. M. Elam (Eds.), *Improving experimental design and statistical analysis*. Chicago: Rand McNally, 1967, 257-291. Reprinted: in W. M. Evan (Ed.), Organizational experiments: Laboratory and field research. New York: Harper & Row, 1971, 169-179.

Campbell, D. T. A phenomenology of the other one: Corrigible, hypothetical and critical. In T. Mischel (Ed.), *Human action: conceptual and empirical issues*. New York: Academic Press, 1969a, 41-69.

Campbell, D. T. Reforms as experiments. *American Psychologist*, 1969b, 24.409-429. (No. 4, April)

Campbell, D. T. Considering the case against experimental evaluations of social innovations. *Administrative Science Quarterly*, 1970, 15.110-113 (No. 1, March)

Campbell, D. T. *Methods for an experimenting society*. Paper presented to the Eastern Psychological Association, April 17, 1971, and to the American Psychological Association, Sunday, September 5, 12:00, Diplomat Room, Shoreham Hotel, Washington D.C. to appear when revised in the American Psychologist.

Campbell, D. T. Experimentation revisited: A conversation with Donald T. Campbell (by S. Salasin). *Evaluation*, 1973, 1.7-13. (NO. 1)

Campbell, D. T. Qualitative Knowing in Action Research. *Journal of Social Issues*, probably 1975.

Campbell, D. T., & Boruch, R. F. *Making the case for randomized assignment to treatments by considering the*

alternatives: Six ways in which quasi-experimental evaluations in compensatory education trend to underestimate effects. Lecture delivered at the conference on Central Issues in Social Program Evaluation, C. A. Bennett & A. Lumsdaine, coordinators, Battelle Human Affairs Research Center, Seattle, Washington, July 1973. Publication pending.

Campbell, D. T., Boruch, R. F., Schwartz, R. D., & Steinberg, J. *Confidentiality-preserving modes of useful access to files and to interfile exchange for statistical analysis*, National Research Council, Committee on Federal Agency Evaluation Research, Final Report, Appendix A. 1975.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), D*isadvantaged child*. Vol. 3 Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrate-multimethod matrix. *Psychological Bulletin*, 1959, 56.81-105. (Also, Bobbs-Merrill Reprint series in the social sciences, S-354.)

Campbell, D. T., Siegman, C. R., & Rees, M. B. Direction-of-wording effects in the relationships between scales. *Psychological Bulletin*, 1967, 68.293-303. (No. 5)

Campbell, D. T. & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNaly, 1966.

Caporso, J. A., & Roos, L. L., Jr. (Eds.) *Quasi-experimental approaches: Testing theory and evaluating policy*. Evanston, Ill.: Northwestern University Press, 1973

Caro, F. G. (Ed.) *Readings in evaluation research*. New York: Russell Sage Foundation, 1972.

Cicirelli, V. G., et al. *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. A report presented to the Office of Economic Opportunity pursuant to Contract B89-4536, June 1969. Westinghouse Learning Corporation, Ohio University. (Distributed by Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce, National Bureau of Standards, Institute for Applied Technology. PB 184 328).

Conner, R. F. *A methodological analysis of twelve true experimental program evaluations*. Unpublished doctoral dissertation, Northwestern University, August, 1974.

Cook, T. D., Appleton, H. , Conner, R., Shaffer, A., Tamkin, G., & Weber, S. J. *Sesame Street revisited: A case study in evaluation research*. New York: Russell Sage Foundation, 1975.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette & J. P. Campbell (Eds.), *Handbook of industrial and organizational research*. Chicago: Rand McNally, 1975.

Cronbach, L. J. Response sets and test validity. *Educational and Psychological Measurement*, 1946, 6.475-494.

Cronbach, L. J. Further evidence on response sets and test design. *Educational and Psychological Measurement*, 1950, 10.3-31.

David, H. P. *Family planning and abortion in the socialist countries on*

Central and Eastern Europe. New York: The Population Council, 1970. Romanian data based primarily on Anuarul Statistic al Republicii Socialiste Romania.

Douglas, J. D. *The social meanings of suicide*. Princeton, N.J.: Princeton University Press, 1967.

Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.

Fairweather, G. W. *Methods for experimental social innovation*. New York: Wiley, 1967

Fischer, J. L. The uses of Internal Revenue Service data. In M.E. Borus (Ed.), *Evaluating the impact of manpower programs*. Lexington, Mass.: D. C. Health, 1972, 177-180

Gardiner, J. A. *Traffic and the police: Variations in law-enforcement policy*. Cambridge, Mass.: Harvard University Press, 1969.

Garfinkel, H. "Good" organizational reasons for "bad" clinic records. In H. Garfinkel, *Studies in ethnomethodology*. Englewood Cliffs, N.J.: Prentice Hall, 1967, 186-207.

Glaser, D. *Routinizing evaluation: Getting feedback on effectiveness of crime and delinquency programs*. Washington, D.C.: Center for Studies of Crime and Delinquency, NIMH., 1973. (Available from U.S. Government Printing Office, Washington, D.C. 20402; publication number 1724-00319, $1.55.)

Glass, G. V., Willson, V. L., & Gottman, J. M. *Design and analysis of time-series experiments*. Boulder, Colo.: Laboratory of Educational Research, University of Colorado, 1972.

Goldberger, A. S. Selection bias in evaluating treatment effects: Some formal illustrations. *Discussion Papers*, 123-72. Madison: Institute for Research on Poverty, University of Wisconsin, 1972.

Gordon, A. C., Campbell, D. T., et al. *Recommended accounting procedures for the evaluation of improvements in the delivery of state social services*. Duplicated paper, Center for Urban Affairs, Northwestern University, 1971.

Goslin, D. A. *Guidelines for the collection, maintenance, and dissemination of pupil records*. New York: Russell Sage Foundation, 1970.

Granick, D. *Management of the industrial firm in the U.S.S.R*. New York: Columbia University Press, 1954.

Guttentag, M. Models and methods in evaluation research. *Journal for the Theory of Social Behavior*, 1971, 1.75-95. (No. 1, April)

Guttentag, M. Evaluation of social intervention programs. *Annals of the New York Academy of Sciences*, 1973, 218.3-13.

Guttentag, M. Subjectivity and its use in evaluation research. *Evaluation*, 1973, 1.60-65. (No. 2)

Hatry, H. P., Winnie, R. E., & Fisk, D. M. *Practical program evaluation for state and local government officials*. Washington, D.C.: The Urban Institute (2100 M. Street NW, Washington, DC 20037), 1973.

Heller, R. N. The uses of social security administration data. In M. E. Borus (Ed.), *Evaluating the impact of manpower programs*. Lexington, Mass.: D. C. Heath, 1972, 197-201.

Ikeda, K., Yinger, J. M., & Laycock, F. *Reforms as experiments and experiments as reforms*. Paper presented to the meeting of the Ohio Valley Sociological Society, Akron, May 1970.

Jackson, D. N., & Messick, S. Response styles on the MMPI: Comparison of

clinical and normal samples. *Journal of Abnormal and Social Psychology*, 1962, 65.285-299.

Kepka, E. J. *Model representation and the threat of instability in the interrupted time series quasi-experiment*. Doctoral dissertation, Northwestern University, Department of Psychology, 1971.

Kershaw, D. N. A negative income tax experiment. *Scientific American*, 1972, 227.19-25.

Kershaw, D. N. The New Jersey negative income tax experiment: a summary of the design, operations, and results of the first large-scale social science experiment. Dartmouth/OECD Seminar on "Social Research and Public Policies," Sept. 13-15, 1974 (this volume).

Kershaw, D. N., & Fair, J. Operations, surveys, and administration, Volume IV of the Final Report of the New Jersey Graduated Work Incentive Experiment. Madison, Wisconsin: Institute for Research on Poverty, December 1973. Duplicated, 456 pp. (Mathematica, Inc., Princeton, N.J. as Vol. IV of the Report of the New Jersey Negative Income Tax Experiment) December, 1973. (To be published by Academic Press)

Kitsuse, J. K., & Circourel, A. V. A note on the uses of official statistics. *Social Problems*, 1963, 11.131-139. (Fall)

Kuhn, T. S. *The structure of scientific revolutions*. (2nd ed.) Chicago: University of Chicago Press, 1979. (Vol. 2, No. 2, International Encyclopedia of Unified Science.)

Kutchinsky, B. The effect of easy availability of pornography on the incident of sex crimes: The Danish experience. *Journal of Social Issues*, 1973, 29.163-181. (No. 3)

Levenson, B., & McDill, M. S. Vocational graduates in auto mechnaics: A follow-up study of Negro and white youth. *Phylon*, 1966, 27.347-357. (No. 4)

Lohr, B. W. *An historical view of the research on the factors related to the utilization of health services*. Duplicated Research Report, Bureau for Health Services Research and Evaluation, Social and Economic Analysis Division, Rockvill, Md., January 1973, 34 pp. In press, U.S. Government Printing Office.

Lord, F. M. Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 1960, 55.307-321.

Lord, F. M. Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 1969, 72.336-337.

McCain, L. J. *Data analysis in the interrupted time series quasi-experiment*. Master's thesis, Northwestern University, Department of Computer Sciences, in preparation.

Miller, A. R. The assault on privacy. Ann Arbor, Mich.: University of Michigan Press, 1971.

Miller, S. M., Roby, P., & Steenwijk, A. A. V. Creaming the poor. Trans-action, 1970, 7.38-45. (No. 8, June)

Morgenstern, O. On the accuracy of economic observations. (2nd ed.) Princeton, N.J.: Princeton University Press, 1963.

Morrissey, W. R. Nixon anti-crime plan undermines crime statistics. Justice Magazine, 1972, 1.8-11, 14. (No. 5/6, June/July) (Justice Magazine, 922 National Press Building, Washington, D.C. 20004)

Porter, A. C. The effects of using fallible variables in the analysis of covariance. Unpublished doctoral dissertation,

University of Wisconsin, 1967. (University Microfilms, Ann Arbor, Michigan, 1968)

Ridgeway, V. Dysfunctional consequences of performance measures. Administrative Science Quarterly, 1956, 1.240-247. (No. 2, September)

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Pratt, J., Rees, A., & Williams, W. Social Experimentation: a method for planning and evaluating social intervention. New York: Academic Press, 1974.

Rivlin, A. M. Systemic thinking for social action. Washington, D.C.: The Brookings Institution, 1971.

Ross, H. L. Law, science and accidents: The British Road Safety Act of 1967. Journal of Legal Studies, 1973, 2.1-75. (No. 1; American Bar Foundation)

Rossi, P. H. Practice, method, and theory in evaluating social-action programs. In J. L. Sundquist (Ed.), On fighting poverty. New York: Basic Books, 1969, 217-234. (Chapter 10)

Rossi, P. H., & Williams, W. (Eds.) *Evaluating social programs: Theory, practice, and politics*. New York: Seminar Press, 1972.

Ruehausen, O. M., & Brim, O. G., Jr. Privacy and behavioral research. *Columbia Law Review*, 1965, 65.1184-1211.

Sawyer, J., & Schecter, H. Computers, privacy, and the National Data Center: The responsibility of social scientists. *American Psychologists*, 1968, 23.810-818.

Schwartz, R. D., & Orleans, S. *On legal sanctions*. University of Chicago Law Review, 1967, 34.274-300.

Seidman, D., & Couzens, M. Crime statistics and the great American anti-crime crusade: Police misreporting of crime and political pressures. Paper presented at the meeting of the *American Political Science Association*, Washington, D.C., September 1972. (To appear in Law & Society Review).

Segall, M. H., Campbell, D. T., & Herskovits, M. J. *The influence of culture on visual perception*. Indianapolis, Ind.: Bobbs-Merrill, 1966.

Shaver, P., & Staines, G. Problems facing Campbell's "Experimenting Society." *Urban Affairs Quarterly*, 1971, 7.173-186. (No. 2, December)

Skolnick, J. H. Justice without trial: *Law enforcement in democratic society*. New York: Wiley, 1966. (Chapter 8, 164-181)

Smith, M. S., & Bissell, J. S. Report analysis: The impact of Head Start. *Harvard Educational Review*, 1970, 40.51-104.

Stake, R. E. Testing hazards in performance contracting. *Phi Delta Kappan*, 1971. 52.583-588. (No. 10, June)

Suchman, E. *Evaluation research*. New York: Russell Sage Foundation, 1967.

Sween, J. A. *The experimental regression design: An inquiry into the feasibility of non-random treatment allocation*. Unpublished doctoral dissertation, Northwestern University, 1971.

Thompson, C. W. N. Administrative experiments: The experience of fifty-eight engineers and engineering managers. IEEE Transactions on *Engineering Management*, 1974, EM-21, pp. 42-50. (No. 2, May)

Thorndike, R. L. Regression fallacies in the matched groups experiment. *Psychometrika*, 1942, 7.85-102.

Twigg, R. Downgrading of crimes verified in Baltimore. Justice Magazine, 1972, 1, 15, 18. (No. 5/6, June/July) (Justice

Magazine, 922 National Press Building, Washington, D.C. 20004)

Watts, H. W., & Rees, A. (Eds.) Final report of the New Jersey graduated work incentive experiment. Volume I. An overview of the labor supply results and of Central labor-supply results (700 pp.), Volume II. Studies relating to the validity and generalizability of the results (250 pp.), Volume III. Response with respect to expenditure, health, and social behavior and technical notes (300 pp.) Madison, Wisconsin: Institute for Research on Poverty, University of Wisconsin. December, 1973. (Duplicated)

Weber, S. J., Cook, T. D., & Campbell, D. T. *The effects of school integration on the academic self-concept of public school children*. Paper presented at the meeting of the Midwestern Psychological Association, Detroit, 1971.

Weiss, C. H. (Ed.) *Evaluating action programs: Readings in social action and education*. Boston: Allyn & Bacon, 1972a.

Weiss, C. H. *Evaluation research*. Englewood Cliffs, N.J.: Prentice-Hall, 1972b.

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

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: Experimental design, its difficulties, and an alternative. *Administrative Science Quarterly*, 1970, 15.97-109.

Westin, A. F. *Privacy and freedom*. New York: Atheneum, 1967.

Wheeler, S. (Ed.) *Files and dossiers in American life*. New York: Russell Sage Foundation, 1969.

Wholey, J. S., Nay, J. N., Scanlon, J. W., Schmidt, R. E. If you don't care where you get to, then it doesn't matter which way you go. Dartmouth/OECD Seminar on Social Research and Public Policies, Sept. 13-15, 1974 (this volume).

Wholey, J. S., Scanlon, J. W., Duffy, H. G., Fukumoto, J., & Vogt, L. M. Federal evaluation policy. Washington, D.C.: The Urban Institute, 1970.

Wilder, C. S. Physician Visits, volume and interval since last visit, U.S. 1969. Rockville, Md.: National Center for Health Statistics, series 10, No. 75, July 1972 (DHEW Pub. No. [HSM] 72-1064).

Williams, W., & Evans, J. W. The politics of evaluation: The case of Head Start. *The Annals*, 1969, 385.118-132. (September)

Wortman, C. B., Hendricks, M., & Hillis, J. Some reactions to the randomization process. Duplicated research report, Northwestern University, Department of Psychology, 1974.

Zeisel, H. And then there were none: The diminunition of the Federal jury. *University of Chicago Law Review*, 1971, 38.710-724.

Zeisel, J. The future of law enforcement statistics: A summary view. In *Federal statistics: Report of the President's Commission*. Vol. II, 1971, 527-555. Washington, D.C.: U.S. Government Printing Office. (Stock no. 4000-0269)