

Can We Be Scientific in Applied Social Science?

Can we be scientific in applied social science? My ability to take a middle-of-the-road, sensible position in a militant polemic way makes you know that my answer will be both *yes* and *no*. Certainly it is much harder to be scientific where financially enormous policy decisions hang upon our fragile social science tools. Let me give you one preliminary *yes* and two preliminary *no*'s.

A feeble *yes*: We can be somewhat more scientific than we are now or have been (in educational program evaluation, for example). Changes feasible within the current financial, administrative, and political climate could make us able to be more scientific. An equally feeble *no*: If we present our resulting improved truth claims as though they were definitive achievements comparable to those in the physical sciences, and thus deserving to override ordinary wisdom when they disagree, we can be socially destructive. We can be engaged in a political misuse of the authority of science that has not been fully earned in our own field. Another *no*: Using quantitative social science measures for administrative control and budgetary decision making (as in the accountability movement) can be destructive of the institutions and processes over which control is intended, and destructive as well of whatever prior validity the social science measures employed once had.

I want to come to these conclusions, or get close to them, by briefly reviewing recent developments in the philosophy of science, sociology of science, and sociology of knowledge, including the argument within our own program evaluation community as to whether we should employ the methods of physics or the methods of the humanities. In light of a fragmentary, modern, postpositivist theory of science, I will then discuss the special problems of policy-relevant social science research, including problems resulting from the politicization of our own mistaken view as to what an applied social science should look like, which we offered in the heyday of the Great Society Program of the 1964–68 period, under the regime of one of our two presidents named Lyndon Johnson (that is, “Lyndon Johnson the Good”). I am thinking of the Office of Economic Opportunity, Program Planning and Budgeting Systems, and program evaluation.

From *Evaluation Studies Review Annual*, edited by Ross F. Connor, David G. Attman, and Christine Jackson, vol. 9 (1984): 26–48. © 1984 by Sage Publications, Inc. Reprinted by permission of the publisher.

830

Postpositivist Theory of Science

Twenty years ago logical positivism dominated the philosophy of science and, through concepts like *operational definition*, dominated our thoughts about research methods. Today the tide has completely turned among the theorists of science in philosophy, sociology, and elsewhere. Logical positivism is almost universally rejected. This rejection, in which I have participated, has left our theory of science in disarray. Under some interpretations it has undermined our determination to be scientific and our faith that validity and truth are rational and reasonable goals. What we should have learned instead was that logical positivism was a gross misreading of the method of the already successful sciences. Logical positivism was wrong in rejecting causal processes imputed to unobserved variables. Logical positivism failed to recognize that even at its best, experimental research is equivocal and ambiguous in its relation both to the real physical process involved and to scientific theory; and that attention to this equivocality calls for use of multiple methods, none of them definitional, triangulating on causal processes that are imperfectly exemplified in our experimental treatments and measurement processes. Properly interpreted, the dethronement of logical positivism should have led to an *increase* in methodological concern rather than its abandonment. Positivism's worst gift to the social sciences was definitional operationism, and this still persists in applied social science, as in the accountability movement in which goal statements and achievement claims are rigidly defined in terms of singular, quantitative indicators. (In practice, the use of such indicators for practical decision making reduces or emasculates the validity of the measures involved [Campbell 1979a, 84-86].)

Campbell and Stanley's *Experimental and Quasiexperimental Designs for Research* (written in 1961 and 1962, first published in 1963), was lucky to be already postpositivist. (At least in a whiggish rewriting of history, I can claim that. In Cook and Campbell [1979, 10-14] the assessment is more mixed.) First of all, we cited N. R. Hanson (1958), who was the first in the Hanson-Kuhn-Feyerabend tradition to emphasize the theory-ladenness of the factual observations of science. It cited Popper (1959) with approval (although it didn't cite my favorite slogan of his: "We don't know, we can only guess"). We emphasized the equivocality of both the treatment implementations and the observations. We gave a section head and two paragraphs to evolutionary epistemology (in Campbell [1959a] version, if not the 1974b, and thus did not include my now standard, "Cousins to the amoeba, how could we know for certain?"). Most importantly postpositivist was the concept of "plausible rival hypotheses," putting so much scientific weight on that squishy concept of *plausibility*.

I would like to point out five postpositivist points with which I agree and with which I think you also should agree. I am borrowing from Hanson (1958), Polanyi (1958), Popper (1959), Toulmin (1961), Kuhn (1970), Feyerabend (1975 and before) and other wild characters including Quine (1951, 1969).

1. *Judgmental, discretionary components are unavoidable in science.* They appear in the choice of experimental design, the choice of a specific apparatus, the wording of the particular questions in our questionnaires, in the interpretation of results, and in the choice between competing theories. These subjective discretionary links cannot be avoided. Logical positivism wanted to remove all discretion. This effort to achieve foundationalist explicitness took two forms: completely explicit observational foundations (meter readings, sense data, and so on) and logical deductive manipulation of these sense data. Logical positivism failed at both levels.

Campbell and Stanley ([1963] 1966, 35) joined in this rejection of logical positivism when they said that "true" experiments at their very best only *probe* theories; they do not *prove* them. But the rejection was most important in our emphasis upon the role of *plausibility*. We took the position that there could be lots of threats to validity that were logically uncontrolled but that one should not worry about unless they were plausible. The general spirit was that any interpretation of a body of data or research procedure should be regarded as innocent until judged guilty for plausible reasons, as determined through the scientific method of mutual criticism.

I've often wondered why there were no hostile logical-positivist reviews of Campbell and Stanley, accusing it of undermining scientific standards. We failed to get one as far as I know.¹ It is with mixed pride that I note we are now regularly being used as an exemplar of logical positivism, and of the mistaken effort to import into the social sciences the inappropriate methods of the physical sciences. (While I am grateful for every citation, I think this is a misreading, as will be argued below.)

2. *The paradigm theme.* We are inevitably encapsulated in some paradigm of presuppositions, inexplicit or explicit. Historically, we can look back and see how provincially we were imbedded. We cannot do without presuppositions. We cannot pull each presupposition out individually and prove them one at a time. In every expansion of scientific knowledge we have to expand the number of things we assume are true and that have to go unproven. In the evolutionary-epistemology version of this, with the recipe of variation, selection, and retention, there is emphasis upon the presuppositions about the nature of the world that are built into our retinas, the nerve wiring of our brains, our language, and our own research tradition. From evolutionary epistemology comes the crucial question of balance between

1. Suppes's "Facts and Fantasies of Education" (1973, 14ff.) comes closest. Listing us under "second-order fantasies," he chides us, sympathetically, for offering no reasons for our "wholly enthusiastic support of experiments," no "abstract principles for which . . . principles of experimentation are derived," no "collection of empirical evidence bearing on the theory of experimentation," no "defense of the reasons for randomizing in experiments," and as needing "derivation from first principles in at least one example." While Cook and Campbell (1979) may have gone part way to meeting these and Suppes's other objections, probably Charles Reichardt's (1983) paper best fills the conceptual gap he and others have noted.

variation and retention. These are incompatible, and knowledge becomes impossible if either totally dominates.

In accepting paradigm-embeddedness again we are rejecting the *foundationalism* that was so central to logical positivism. There are no untouchable axioms: All are criticizable and revisable. Nor are there any foundational observations or facts. There are indeed at any historic period of time in any successful science a vast array of trusted facts, but none is immune from revision. For the atomistic (sense data, observations, or axioms) foundationalism of the positivists, we must substitute a holistic, squishy, quasi-foundationalism that I call the 99 to 1 trust-doubt ratio. This is like the holistic network imagery of Quine (1951), but I'll give it to you in my version.

For the cumulative evolutionary process of knowing, our only available tactic is to trust most of our current beliefs while we use that distributed fulcrum to revise a few of them. The ratio of the trusted to the doubted has got to be in the order of 99 percent trust to 1 percent doubt. In biological evolution, 99 percent of the genes are trusted while mutation and recombination vary 1 percent of them. However wrong-headed the initial beginnings, nature is stuck with this great mass of pre-suppositions on how to design an animal. Similarly, in a science such as physics, the great revolutions have been achieved by trusting 99 percent of the cumulated facts and using that basis to revise 1 percent of its beliefs and their theoretical integration. This produces a kind of gradualism at the level of facts (wherein lies my only real disagreement with Kuhn).

Don Moyer (1979) has studied the belief changes following the 1919 eclipse observations, where English physicists and astronomers moved from 5 percent adoption of Einsteinian general relativity to 99 percent adoption in a five-year period. He documents the ways this revolution was based upon profound trust of previous physics, which provided the factual leverage for overthrowing the dominant Newtonian theory. It was palace revolution in conceptual organization and theory, in which most of the facts (all being theory-laden "facts") were retained.

Before going on to the other three points, I would like to use these first two points (paradigm dependence, and discretionary human judgments) to discuss the qualitative versus quantitative agenda which is so important right now in educational research and program evaluation. Should we be using the methods of the humanities or the methods of the physical sciences? I would like to argue that if we had not misread the record of the physical sciences we would recognize that these methods are very similar.

Let us start out with that old tradition, at one time called philology and now called hermeneutics, which asks such questions as, What did Homer mean by this particular phrasing? or, What did Saint Paul mean in this particular verse? In philology and hermeneutics, one had generations of scholars quarreling about these issues, but remaining within the same social communication net, a quarreling collective committed to getting the truth. Now, part of this hermeneutic tradition is this presuppositional and contextual dependence that I have called the 1 to 99

doubt-trust ratio, a composite fallibilist foundationalism generating and criticizing plausible rival hypotheses as to alternative interpretations, including the hypothesis that some copyist had made a clerical error that was subsequently transmitted by other loyal copiers, and so on. This self-critical community of interpreters, by looking at a wider range of manuscripts from this same time, and thus extending the grounds of judgment, often eventually arrived at consensual decisions as to the best interpretations of a particular manuscript.

Or, look at the method of the historians as taught and exemplified by Collingwood (Levine and Levine 1970), who was a historical relativist with a historical paradigm theme. His method was explicitly the method of a detective in a detective story. The method is epitomized by trying to rule out plausible rival hypotheses.

When we get down to our own practical work, a plausible-rival-hypothesis approach is absolutely essential, and must for the most part be implemented by common-sense, humanistic, qualitative approaches. In program evaluation the details of program implementation history, the site-specific wisdom, and the gossip about where the bodies are buried are all essential to interpreting the *quantitative* data (Campbell 1974e, 1975c, 1979a; Cook and Reichardt 1979).

3. *Historicism*. At any given time, even in the best of science (even in physics), we are in a historical context and our experiments and our theoretical arguments are historically imbedded. They have a historical provincialism; they are reactions to what has gone before; they are dated and uninterpretable outside of that context. The contrasts with the past are, in some kind of a problem-solving way, almost necessarily exaggerated. So we have a dialectic of contrast, in which exaggerated, oversimplified corrections for what has gone before are an essential part of the process, and the past that has gone before is essential for understanding the new terms and new experiments that are introduced. In an effort to speak in the extreme forms of postpositivist jargon, I have called this the "dialectic historical indexicality of scientific terms" (Campbell 1982a). Gergen (1982) presents the historicist argument for social psychology.

4. *Relativism*. This treasure of post-positivism encompasses epistemological, historical, cultural, and paradigm relativism. In the evolutionary epistemology tradition (Campbell 1974b) my slogan is "blind variation and selective retention." This is an emphasis on exploring in the dark, with the fumbling of a blind person being a better model for epistemology than clairvoyant vision. All of this commits me to a profound epistemological relativism.

Now, while I am a thoroughgoing epistemological relativist, I reject an ontological relativity, or, since Quine (1969) has used that term in a different sense, an ontological nihilism. Evolutionary epistemology has in it an unproven assumption of a real world external to the organism, with which the organism is in dialectical interaction. I have been spending a lot of time recently reading and meeting with (Campbell 1981g) exciting young sociologists of science such as Barry Barnes

(1976), David Bloor (1976), and Michael Mulkay (1979). Karen Knorr-Cetina (1981), Bruno Latour and Steven Woolgar (1979), and Harry Collins (1981). Also relevant is the book Robert Merton and Thomas Kuhn have resurrected, Ludwig Fleck's 1935 *Genesis and Development of Scientific Fact* (1979). Harry Collins calls this the "relativist" program in the sociology of science. Latour and Woolgar and Knorr-Cetina call it the "social constructivist" program. David Bloor and Barry Barnes (Bloor 1976; Barnes and Bloor 1982), call it the "strong programme," meaning that in doing sociological, historical studies of science (asking the question, What were the causes for their changing their scientific beliefs?) it is illegitimate to use our current confidence in the truth of the belief as an explanation for why, back then, they came to believe it.

This agnosticism I find methodologically correct. After all, those past scientists were not clairvoyant, and many of the changes we now regard as in the mainstream of scientific development we do not now regard as "true." But these new sociologists of science carry this agnosticism too far. They refuse to speculate on an ontologically-realist way about what kinds of social processes, what kinds of systems of interaction among scientists and between scientists and society, could produce *improved* beliefs. They refuse to undertake what I call an epistemologically relevant internalist sociology of science (Campbell 1979d, 1981g). I am continuing to work on such a sociology of science (Campbell 1986a).

5. *Sociologism and psychologism.* Science is a social process, scientists are thoroughly human beings: greedily ambitious, competitive, unscrupulous, self-interested, clique-partisan, biased by tradition and cultural memberships, given to mutual backscratching, and the like. James Watson's *The Double Helix* (1968; but see Olby 1974, for Crick's perspective) is one of the most used texts in the sociology of science relevant to this.

Out of this, I want us to keep the goal of *truth*, and to attempt to understand and foster a social system of science (differing greatly from our recent dominant theory of applied social science for policy purposes) in which it becomes sociologically plausible that the processes would lead to beliefs of increasing validity. The scientific method itself is a social system product. Science is itself a social system, it is "tribal" in that sense (Campbell 1979d), but strangely, its norms preach against that very tribalism: against deference to authority, against deference to majority rule. A key part of this sociology of successful science is a mutual criticism that keeps those who are criticizing each other still remaining in the same group, rather than splitting off into their own insulated cults. Competitive replication, threat of replication, a reward system that encourages competitive innovation but punishes dishonesty in the resulting competition (Merton 1973) are all parts of it (Campbell 1986a).

From this sociological point of view, combined with an evolutionary-epistemology point of view, it follows that large numbers of independent decision

makers are essential for objectivity in science. It follows too that we must maintain scientists' collective interests in the trust given the system of science by the larger public (Merton and Geiryn 1982). We must maintain the individual scientist's interest in reputation, recognition, and fame, without allowing these interests to undermine the self-interest in science's collective validity. We scientists cannot avoid being dependent on the trust of fellow scientists. We must avoid creating a motivational system that generates truth claims or belief assertions that we distrust. We need a scientific method (as a social invention and social process) that will counteract the ill effects that a cynical and nihilistic interpretation of point 4 (relativism) and point 5 (sociologism and psychologism) can produce.

This epistemologically relevant internalist sociology of science will not deny the scientist's paradigmatic provincialism, self-seeking competitiveness, and human fallibility, but will rather propose a social system designed to curb side effects that produce invalid beliefs. Inevitably our model of science will show science as a fragile and vulnerable social institution, one that is capable of flourishing only now and then, only here and there, on the face of the earth. A validity-producing social system of science is nothing we should take for granted.

Application to Applied Social Science

If we move such a post-positivist theory of science into the problem of the validity of applied social science, we find that we need all of the social system features of pure science (e.g. physics, laboratory psychology, and biology). From this perspective, when we move into the arena of policymaking, there are some regular features of applied social science for policy purposes that come to our attention.

First is clearly the *greater equivocality of causal inference for research done in policy settings*. There are many, many more plausible rival hypotheses. There is much less control. Looking back at the "artificiality" of physical science laboratories (their soundproof walls, atmospheric controls, insulation against electromagnetic and magnetic fields, achievement of vacuums, and all of the other accoutrements of "experimental isolation"), we can see that all of this laboratory apparatus is designed to control or to rule out *plausible rival hypotheses*, or at least to render them "implausible," thus achieving an *artificial* situation in which causal inference can be done more competently.

When biologists left the insulated laboratory where apparatus and walls are the essence of the scientific method, to move out into the agricultural experimental station where the winds blew and the rains rained, they invented another type of artificiality to render implausible large classes of plausible rival hypotheses. This was the *randomized* experiment. We should note that slightly before that, educational researchers such as E. L. Thorndike and his students, moving from the insulated psychology laboratory out into the classrooms, independently invented randomized assignment to experimental treatments and latin-square designs, again

833

as artificialities that operated somewhat like experimental isolation in generating controls, in reducing the plausibility of rival hypotheses such as selection, selection-treatment interaction, practice effects, and the like. While we educational psychologists did not do it with Fisher's mathematical elegance, we were *first* with these great tools of artificiality. McCall's (1923) *How to Experiment in Education* summarizes this early achievement.

Today, as so many of us react to the frustrations of social science research with the hope that humanistic methods will turn out to be more appropriate than physical science ones—an exploration that I too favor (Campbell 1974e, 1975c, 1986a), our troubles are often blamed on a prior, mistaken, subservient borrowing of physical science methods. Indeed, Campbell and Stanley (1966) are often accused of this fallacy. Close analysis will, I believe, show that this is unfair. Thorndike and McCall were *not* borrowing random assignment and the "rotation experiment" (latin-square) from the physical sciences, nor from R. A. Fisher and his agricultural experiment stations. Instead, they were reacting to the mutual criticisms of their own educational-psychology research community, and inventing research designs that would help rule out the recurrent very plausible rival hypotheses generated by their fellow critics.

So too, Campbell and Stanley's list of threats to validity is an accumulation of our field's criticisms of each other's research. The list of quasi-experimental designs is a cumulative listing of our discipline's inventions of ways of ruling out some of the very plausible rival hypotheses. We can thank Campbell and Stanley for being conscientious collectors of the achievements of this tradition of collective self-criticism. (That's what they were: collators, bookkeepers, reviewers of the literature.) Their collection of designs is not at all drawn from physical science. Of course, from the quasi-experimental perspective, just as from that of physical science methodology, it is obvious that moving out into the real world increases the number of plausible rival hypotheses. Experiments move to quasi-experiments, and on into queasy experiments, all too easily.

A second difference between applied social science and laboratory research is that the still greater likelihood of *extraneous, nondescriptive interests and biases* entering through the inevitable discretionary judgmental components that exist in all science at the levels of data collection, instrument design and selection, data interpretation, and choice of theory. As we move into the policy arena there is much less social-system-of-science control over such discretionary judgment favoring descriptive validity, and there are much stronger nondescriptive motives to consciously or unconsciously use that discretionary judgment, to, so to speak, break the glass of the galvanometer and get in there and push the needle one way or the other so that it provides the meter-reading wanted for nondescriptive reasons (Campbell 1982a; 1979a, 84-86).

The next few points about moving the theory of science into the applied social science arena stem in considerable part from the seriously mistaken model of

applied social science that we social science methodologists offered to ourselves and to government in the 1960s, in the period of the Great Society, in the era of the Office of Economic Opportunity, and Program Planning Budgeting and Systems. Many of these I have gone over on previous occasions (Campbell, 1974e; 1979a).

My third point is the *mistaken belief that quantitative measures replace qualitative knowing*. Instead, qualitative knowing is absolutely essential as a pre-requisite foundation for quantification in any science. Without competence at the qualitative level, one's computer printout is misleading or meaningless. We failed in our thinking about program evaluation methods to emphasize the need for a qualitative context that could be depended upon. One example is frequent separation of data collection, data analysis and program implementation that was once characteristic of Washington's funding of programs, in which one firm would collect the pretest, another firm would collect the post-test, and a third firm would analyze the data. This easily led to a gullible credulity about the numbers on the computer tape, with the analyst in total innocence about what was actually going on in the program implementation and testing situations.

To rule out plausible rival hypotheses we need situation-specific wisdom. The lack of this knowledge (whether it be called ethnography, or program history, or gossip) makes us incompetent estimators of program impacts, turning out conclusions that are not only wrong, but are often wrong in socially destructive ways.

Fourth, the evaluation model we offered mistakenly bought into the logical positivist's *definitional operationalism*, specifying as program goals fallible measures open to bureaucratic manipulation (Campbell 1969e, 414-17; 1979a, 84-86).

Fifth, a *one decision/one research* ideal was a central feature of our original program evaluation model. (This is diametrically opposed to the social system of the successful physical and biological sciences.) Each program evaluation was to be done to support a specific administrative decision. One researcher-evaluator was to have a monopoly on the resulting truth claims. This one study was to be the basis for the decision. With this often went a disregard of prior wisdom and prior science in making the decisions about the future of the program (Lindblom and Cohen 1979). The program evaluation was conceptually tied to refunding, to be the sole or an important base for expanding or contracting the program.

Such a policy violates common sense as well as the sociology of knowledge. Had we sat down and thought, What will it do to all of those discretionary points in data collection if next year's funding is going to ride on them? Where are the discretionary points and how can they be distorted?, we would have recognized that this program evaluation model belied our common experience, the sociology of bureaucracy (Blau 1955, 1956; Ginsberg 1984) and of our knowledge as psychologists as to the multiple motives the individuals implementing programs have, including the motive of being able to feed one's children next year. ("Where will another job come from if this program is discontinued?" or, "If we report to our client our unpleasant results, where will next year's contract come from?" and so

on.) These considerations add into the recurring conflict we all have observed between the evaluation staff and the program delivery staff. Program evaluation became destructive of program delivery morale.

6
A sixth mistake in the model that we in the 1964-68 period recommended to government was the emphasis on *external evaluation* of programs rather than evaluation by the delivery team itself. This again is the complete opposite of the social customs of the physical sciences, in which passionate believers in new theories design the research and carry it out. The objectivity of physical science does not come from turning over the running of experiments to people who could not care less about the outcome, nor from having a separate staff to read the meters. It comes from a social process that can be called competitive cross-validation (Campbell 1986a), and from the fact that there are many independent decision makers capable of rerunning an experiment, at least in a theoretically essential form. The resulting dependability of reports (such as it is, and I judge it usually to be high in the physical sciences) comes from a social process rather than from dependence upon the honesty and competence of any single experimenter. Somehow in the social system of science a systematic norm of distrust (Merton's [1973] "organized skepticism") combined with ambitiousness leads people to monitor each other for improved validity. Organized distrust produces trustworthy reports. In contrast, in program evaluation, the monopoly of a single evaluation for each program, with but one decision maker to use it, and the dogma of external evaluation, all combined to make impossible this crucial aspect of the social system of the successful sciences.

7
Another type of mistake involved *immediate evaluation*, evaluation long before programs were debugged, long before those who were implementing a program believed there was anything worth imitating.

8
A totally unnecessary feature was recommending a *single national once-and-for-all evaluation* that would settle the issue forever.

9
Point nine: There was *gross overvaluing of, and financial investment in, external validity*, in the sense of representative samples at the nationwide level. In contrast, the physical sciences are so provincial that they have established major discoveries like the hydrolysis of water (in which electrical anodes and cathodes generate bubbles of oxygen and hydrogen) by a single water sample taken from the Soho neighborhood of London in 1903 (see Campbell [1969d] for a more extended and complex discussion), never cross-validating the discovery on a "representative sample" of all of the water of the world.

The so-called Northwestern School—whose center of strength is still at Northwestern with Bob Boruch, Tom Cook, and their colleagues, and within which I still include Lee Sechrest, Paul Wortman, myself, and most of our Northwestern Ph.D.s—has been criticized for overemphasis of internal validity at the expense of external validity. This accusation must be, in a historical sense at least, wrong. Who, after all, introduced the great emphasis on, itemized all of the threats to, and assembled the controls for external validity (Campbell 1957a; Campbell

and Stanley 1963c; Cook and Campbell 1979)? Of course, we are interested in external validity, but we see no point in having a representative national sample of a repeated regression artifact, or of some other internally invalid research design.

Tenth, is *the neglect of the fact that scientific truths are a collective product of a community of scientists at any given time*. Such a community is self-critical, gets into the guts and looks under the cover and tries to decide what was going on in specific experiments. There was a neglect of this insulating layer of human judgments that are well informed and mutually disciplined. We somehow assumed in our OEO-PPB&S model that a single computer output could speak directly to the administrator. Now, however, as postpositivist fallibilist critical realists, we want our realism to include the real and fallible processes of data collection and conclusion drawing. We can see vision as the product of imperfect lenses, imperfect nervous systems, and oversimplified presumption systems, which lead to generally valid perceptions but also to optical illusions (Campbell 1987a).

This physicalization, this materialization of the process of knowing, is a very important part of the historical development of epistemology. Extended to science we should have seen from the very beginning that social data collection and social experimentation were social system intrusions into the ongoing processes, and that putting policy-decision pressure on them would distort every crushable, squishy, little discretionary link. We were guilty of a doctrine of "immaculate perception" (as it has been called in epistemology), guilty of assuming a noninteractive acausal observational process in which all of our questionnaires and arrangements could describe without disturbing what they were describing, and in which the people being described as well as the describers would be unmotivated to bias the meter readings.

Better Strategies for Applied Social Science

Our postpositivist theory of science with its social system of science emphasis is far from complete. Nor have we yet applied it adequately as an alternative ideology for applied social science, ready as advice to Washington whenever the spirit of the experimenting society, that existed under the regime of the good President Johnson, returns. To be so ready, we must start arguing now about the pros and cons of alternative models. To help initiate this I offer the following.

1. *I'll call the first alternative the contagious cross-validation model* of program evaluation. A generous and concerned government provides funds for developing local programs addressed to chronic sores of society. This local program funding includes funds for whatever evaluation the program designers want, including funds for academic consultants. Lots of local programs result. When any one of them, after a year or so of debugging, feels they have something hot, a program worth others borrowing, we will worry about program evaluation in a serious sense. Our slogan would be, "Evaluate only proud programs!" (Think of the con-

trast with our present ideology, in which Washington planners in Congress and the executive branch design a new program, command immediate nationwide implementation, with no debugging, plus an immediate nationwide evaluation).

When the high morale program and program results were disseminated, there would no doubt emerge a group of willing adopters. (Note that before we had our program evaluation ideology, such borrowing was usually on the basis of persuasive program *plans*, and took place prior to even the first full tryout, as Addie and Murray Levine [1970] have documented so well in one classic instance.) At this stage, our federal funding would support adoptions that include locally designed cross-validating evaluations, including funds for appropriate comparison groups not receiving the treatment. (We might at this or the next stage have large-scale "external" evaluations, as long as these did not preclude interpretable comparisons at each site not depending upon full national implementation.)

After five years we might have 100 locally interpretable experiments. We would also have a community of applied social scientists familiar with them all, that had cross-examined each others' data, suggested and done reanalyses, performed bias-sensitive meta-analyses, and so on. Many of these scholars would be tenured university or public school faculty, whose job security would not depend upon the outcome. From the consensus of this mutually monitoring research community we would advise government and potential adopters.

I leave it an open question whether or not the full-scale dissemination of a clearly successful program would be done without local cross-validation by adopters. Fully facing the problems of external validity, and the social historicity (Gergen 1982) as to what will work when, would require this. I do believe we could make it feasible for many programs, and provide classroom teachers, for example, with realistic means of evaluating the competence of their own practice, albeit usually without synchronous parallel comparison groups except for exploratory innovations.

By moving the primary evaluation to the dissemination stage, we are evaluating the transferable, borrowable aspects of the program. In the initial zeal of program developers, exceptional success is frequently due to heroic eighty-hour weeks on the part of key staff, and these are not aspects of the transferable program. We need to know about effectiveness for the program's routinized form. While the problem of generalizing in applied science is substantially different than in theoretical science, one essential of the "knowledge" produced is still reusability on different occasions and times.

The contagious cross-validation model is much closer to the model of the physical sciences, as noted in the previous section under point 6. Let us remember that applied social science has more, rather than less, need for mutual criticism, argumentative reanalysis, and cross-validation than does physical science. This is just because we lack the possibility of experimental isolation, just because our data have to be generated through the cooperation of persons with strong stakes in the outcome, and just because science (either physical or social) is one in an arena in

which the rival interests in what the outcome is are so powerful that objective description can become a minor motive.

Let me give a concrete illustration that is banal and simple. I was an observer from a nearby but safe distance from the Chicago school system for many years. Here they were spending millions of dollars on testing programs that used national norms for an annual humiliation of half of the grammar schools in the city. That testing program was destructive in its net effect. The annual humiliation did nothing to improve the schools, told them nothing about what they could do to make education better, and put tremendous corruption pressure on test administrations. (Rumored practices were to classify as many children as possible as abnormal ineligibles, and to manipulate the time schedules to optimize performance). Thus the annual humiliation was destructive both of the validity of measurement and of the morale of the teachers.

While there were continual plans and expenditures for individual student data retrieval, neither the school system nor we designers of quasi-experiments ever provided a teacher with the ability to tell whether the text chosen for the current year was better than last year's.² We could have also provided individualized data retrieval disguising the scores so that no one knew what they were in terms of national norms, providing a comparison base for teachers based solely upon the previous performance of their own pupils. No national-norm humiliation need have been involved—merely an ability to tell whether one was doing better than last year. Such de-normed retrieval capability would also have provided adventuresome teachers the capacity to try out alternatives in teaching style. It's a great failure that we never got around to doing this. We program evaluation methodologists never provided the perspective nor the conceptual tools, nor lobbied the school system for this usage and against the other.

2. *Getting competitive replication into national policy pilot studies.* The contagious cross-validation model is appropriate only where the program under study can be implemented autonomously by a local unit (be it school, classroom, city, retail store, or factory). Where the program being piloted has to be eventually implemented nationally, different sources of competitive cross-validation must be sought. I am thinking of such heroic studies as the New Jersey Negative Income Tax Experiment (Watts and Rees 1977; Kershaw and Fair 1976; Pechman and Timpane 1975; Rossi and Lyall 1976) and the several subsequent still larger experiments with guaranteed annual incomes in rural North Carolina, Gary, Seattle, and Denver. Belonging here too are the Housing Alliance Experiments (see Lowry [1982] for the Supply experiments, Abt Associates and the Urban Institute for the

2. Chicago, for all its reputation for corruption, still allowed teachers a list of texts they could choose among, so they could have experimented with textbooks. Textbook evaluation is a good place for a science of program evaluation to cut its teeth. A text obviously differs depending on who is using it, but still it is a relatively specifiable and disseminable program package.

Demand experiments) and the big health insurance experiments. We need such enormous studies, but should run them in the future with deliberate efforts to build in some degree of independent replication and mutual monitoring. Here are several ways this might be done.

A. Rather than awarding a single contract, each should be *split into two or more independent experiments*, so that all of the hundreds of discretionary decisions as to how to present the experimental treatment and design the questionnaires and interviews would be made and implemented by at least two independent research teams. Such heteromethod replication (Campbell 1969d; Cook and Campbell 1979) is needed for interpretive validity. It would also provide a small group of informed scientists for competitive cross-examination.

B. There should be *adversarial stakeholder* participation in the design of each pilot experiment or program evaluation, and again in the interpretation of results (Krause and Howard 1976; Bryk 1983). We should be consulting with the legislative and administrative opponents of the program as well as the advocates, generating measures of feared undesirable outcomes as well as promised benefits.

C. There should be *competitive reanalysis* of data from the big studies. The Office of Economic Opportunity set a great precedent to which we have inadequately responded. The Institute for Research on Poverty, University of Wisconsin, has available for reanalysis the data tapes for the New Jersey Negative Income Tax Experiment, and proper scientific disagreements are emerging, for example, as to how they handled the attrition problem (Boeckmann 1981). They have the data from the first big Head Start evaluation, a data set with a fine record for productive second-guessing (Smith and Bissell 1970; Barnow 1973; Magidson 1977; Bentler and Woodward 1978). I hope they have the big Performance Contracting study (Gramlich and Koschel 1975) with the rebuttals from the performance contractors. Major classics in this area come from my Northwestern colleagues (Cook et al. 1975; Boruch 1978; Boruch et al. 1981; Trochim 1982).

The original Coleman report (1966) on educational desegregation has been thoroughly reanalyzed, so that now we could assemble a half-dozen volumes the size of Mosteller and Moynihan's (1972); and from a modern postpositivist theory of science, we can recognize that only now do we have a competent applied social science community ready to use the Coleman report in conjunction with all related research, prior and subsequent, to guide governmental policy. The original image of one research (one data collection, one analysis, by one analyst team) to guide one governmental decision, was based upon a fallacious theory even for pure science, and still more wrong for applied social science.

While these secondary analyses are of great value, and should become obligatory for all expensive data collections, we should remember that they cannot fully correct for the hundreds of idiosyncratic discretionary judgments involved in the initial data collection.

D. *Legitimizing dissenting-opinion research reports* from members of the research team. The Freedom of Information Act of the late 1960s was one of the

great social inventions increasing the possibility of a valid, policy-relevant, applied social science. While Rights of Subjects legislation (another great innovation) has been used to greatly curtail its practical implementation (needlessly so—see Campbell, Boruch, Schwartz, and Steinberg 1977d; Boruch and Cecil 1979, 1982) the legitimating value is still there. It should make possible competitive reanalyses. Indeed, the Seattle Teachers' Union had used it in demanding the data tapes from The Office of Economic Opportunity's (OEO) Performance Contracting Study before the final report was ready, and OEO had agreed to this in an out-of-court settlement. (This never lead to a rival analysis, in part at least because OEO's official analysis when it came out supported the interests of the teachers' union). In my "The Experimenting Society" (see chapter 11), drawing upon the unpublished and minimally circulated Gordon and Campbell (1971), we argue that the voting booth rather than the rat lab should be the methodological model in policy research, and that the right to reanalyze data employed in governmental decision making is fundamentally related to the right to demand a recount in an election.

Another background for my argument is the great value that whistle-blowing has had for the validity of physical and biological research results when these have been done under conditions of extreme policy relevance. (I am thinking of research on the dangers of chemicals to manufacturing workers and food consumers, the dangers to and effects on humans and sheep of irradiation from nuclear experiments and power generators.) While such whistle-blowing occurs, it is still experienced as a guilt-producing team disloyalty, both by the whistle blower and coworkers, who may react with ostracism. It would improve the scientific and political validity of applied physics, chemistry, and biology if whistle-blowing were legitimated by reconceptualizing it as the right and duty to generate dissenting-opinion research reports, and if all laboratory staff were provided official access to all data for this purpose. Insofar as our research results are inherently more ambiguous, even more do we need this in applied social science.

I am making a radical suggestion, but one that we in the American Educational Research Association, the Evaluation Network, and the Evaluation Research Society, could right now be put into our guidelines on research ethics (Stufflebeam 1981; Rossi 1982). Moreover, we as individuals could start it now with our own research assistants. Imagine if you gave every research assistant (including the neurotic ones with negative Oedipal resolutions whom you never should have hired in the first place) the right of access to all of the data and the right to generate minority reports. I have no doubt that this would increase the validity of the official report (as well as provide some of the needed competitive reevaluation). We research directors would write up our reports differently knowing that our righteous and sore-headed assistants were potentially free to dig up the items on the interview that we neglected in our final report, to dig up and publicize the disappointing analyses we failed to find room for in our final report, or to reanalyze the data with a different perspective. Our profession should start designing a model contract specifying such rights that could be given each employee when hired.

3. *Writing up our evaluation research reports for our fellow evaluation researchers* in and out of the universities, is my third suggested reform of our original OEO-PPB&S model of applied social science. I state it thus because we so often in those early days chided ourselves for letting our academic standards and interests get in the way of writing program evaluations geared to fluent administrative decision making. (I need help in assembling good examples of this literature.) While I am not attempting to condone irrelevant "pure" research smuggled in under the applied budget, I am insisting on having available (along with the data available for reanalysis) a full academic analysis for cross-examination by our applied social science colleagues.

Let me stress this through an aside to those of my students (face-to-face and by the printed word) who feel that the Campbell and Stanley superego has ill prepared them for life in the real world of program evaluation. Let your employer or the administrator whose neck is on the block write up the "executive summary." Be sympathetic to the social role and predicament of program administrators and developers. Do not be a "sadistician" (as one of our psychoanalysts might diagnose it), forcing them to live up to your own most punitive standards of scientific rigor (note Devereaux's *From Anxiety to Method in the Behavioral Sciences* [1967]). You protect your own superego by signing your name to the 200-page appendix addressed to your fellow scientists. We too should be like the physical scientists who advise government from the consensus of a well-informed, mutually monitoring scientific community focused on the problem area. These appendices, proper government funding of conferences, and reanalyses in terms of the plausible rival hypotheses we generate, will provide an applied social science base that is more optimal, politically and scientifically.

The complete sociology of applied-science validity, which I wish I had, would take into account environmental impacts on commitments to validity which applied science careers involve. I will use this future agenda, and my earned status as an academic garrulous grandfather, to permit inserting here (rather than properly reorganizing this paper) some further advice on maintaining the Campbell and Stanley superego in program evaluation careers. It will help if one recognizes that our initial OES-PPB&S rhetoric got fused with a legislative and administrative rhetoric in a way that we should avoid being mousetrapped by.

Still today, governmental funds are needed to provide relief to the ill, aged, and underfed. Let us call this *problem-specific revenue sharing*. But it became politically necessary for such relief funds to be disguised as "new programs" that would cure the problems they were designed to alleviate. Including in the legislation the requirement that the "program" be "scientifically evaluated" became in many, many instances just a part of the escalated rhetoric, a routine part of assuring conscientious, responsible custodianship of governmental funds, on a par with requiring proper bookkeeping and auditing. In such cases the genuinely worthy goal was achieved when funds were spent locally on the problem. *Local fund-spending on the need was the real "program."* (Paying too much attention to pork-barrel motives

supporting the same goal can distract from attending sympathetically to the local relief aspects, and the rhetorical requirements for providing for these needs.) Most such so-called programs involved no alternate disseminable program package. At best, funding and staff are added in ungeneralizable ways to preexisting agencies.

For these programs I recommend avoiding laying one's scientific superego on the line. Save up those negotiating energies and costs in interpersonal goodwill that comparable untreated comparison groups, meaningful pretests, and interpretable before-after comparisons involve to apply to that rare occasion when a potentially valuable innovation is being tried out, or that still rarer occasion when unique circumstances permit an impact assessment of current practice. For the "only-rhetorically programs," do evaluations that are low-cost in both rapport and money. Collect the opinions of well-placed observers as to what would have happened without the "program," and as to what aspects of it failed and what succeeded. Put in the final report appendix useful "input" descriptions. Include discussion on suggestions as to how promising disseminable aspects of local practice, or practitioners' suggested innovations as yet untried, might be implemented in the future in ways that might probe their usefulness.

For such nonprograms, evade (if you can) producing any quantitative estimate of impact. If you cannot, at least in the long appendix surround them with full discussion of how the setting makes them equivocal. If a cost-benefit analysis is required, try to get this subcontracted to an economist or operations researcher whose training has not troubled his conscience with all of the plausible threats to the validity of the "benefit" estimates available. Avoiding quantified guessing in highly equivocal evaluation settings is a matter of political conscience also. Evaluation reports should enter into political decision-making processes as one component to multiparty argument and negotiation. Due to the general prestige of quantified science, not yet earned in our area, quantitative guesses and computer output carry more weight than they should in competition with the qualitative judgments of well-placed observers.

"Street wisdom," or theoretical understanding of the encompassing social system and the political realities (sympathetic to actors and roles, avoiding hostile cynicism) are important components of our "methodology for the experimenting society" (see chapter 11). It will help to remember what Rossi (1969) has taught us. The legislative and administrative setting is always one in which many needs are competing for funds. The most important needs may indeed have priority for funding. But importance means that these are stubborn, unsolvable, chronic problems on which normal societal problem solving has failed. The competition for funds almost guarantees that the tentative solutions for these urgent chronic problems will be underfunded.

It often seems that programs are designed and implemented just so as to preclude interpretable comparisons. This may indeed be so, and so because the designers and administrators have been aware (perhaps unconsciously) that the program could only be a drop in the bucket, and had no chance of living up

to the claims for panacea that were politically necessary for getting even that drop (the "overadvocacy trap," see chapters 10 and 11; Campbell 1971d; Shaver and Staines 1971). We program evaluators, expanding our methodological responsibilities beyond the narrow issues of experimental design (while not at all abandoning these concerns) to include a sociology of applied-scientific validity, must be sympathetic to this predicament. We must avoid reacting with the hostile disdain of wounded idealism and methodological righteousness. We must avoid this not only for the health of the social system in which we participate, but also for our own mental health. Our economic and career predicament may give us no alternative but to keep our job. The reaction of unsympathetic hostile disgust can trap us in self-contempt for prostituting our scientific skills and ideals. I believe we can avoid this by aspiring to a sympathetic understanding of our program director's and our own social-system predicament, and by working as best we can within that system to produce validly interpretable evaluations whenever feasible and when there is a potentially disseminable program alternative worthy of such efforts.

4. *Avoiding "ad hominem" and "ad institutionem" research.* A final radical shift I would like us to consider is the recommendation that we stop using our fragile tools of experimental design and measurement for purposes of managerial control and "accountability." (I thus reverse the implicit recommendations of my early [1956c] view that leadership effectiveness is a causal hypothesis to be demonstrated optimally by quasi-experimental methods.) Financial costs are one reason; these tools are too expensive to be used for personnel selection purposes (for selecting the better teacher, principal, superintendent), nor is quasi-experimental comparability likely to be available to make such data interpretable as an effect of skill, effort, and merit. Nor can we really solve our organizational problems by promoting effective persons out of their current locations of effectiveness. There are not enough dedicated geniuses. Overall, we must improve organizations by discovering optimal use of the energies and abilities of current staffs, rather than by hiring those of proven effectiveness away from their current jobs.

But my main reason for recommending that we exclude the research goals of evaluating institutions, social organizations, and persons, is my conviction that this use, beyond all others, corrupts the validity of the measures, and may also corrupt the very social processes the measures are designed to monitor (Campbell 1979a, 84-86; Blau 1955; Ginsberg 1984). We are thoroughly dependent upon the staffs we evaluate for the qualitative background required by discretionary judgments, as well as for generating much of the data. The social control, organizational management, and personnel evaluation purposes maximize the nondescriptive motives, the motive to influence the decision rather than (or in addition to) provide a valid description. It would be my thesis and hope that these distortion pressures are at minimum when what is being evaluated is a *program*, an alternative that present staffs could adopt without losing their jobs. *Let us evaluate alternative programs, not persons or social units.*

This principle of postpositivist applied social science obviously also supports again the abandonment of the single evaluation, single-decision model, and the decoupling of evaluation from refunding decisions, or a radical reversal of the present coupling. I return to my near-but-safe-distance observation of the old Chicago: It was my sincere judgment that there would have been a substantial saving of program and evaluation funds had the evaluation-funding lineage read, "In the event of no-effect or undesirable-effect outcomes, the same staffs should continue to work on the same problem with an alternative program, and with a 10 percent budget increase above inflation." (We would, of course, have needed econometric tuning of that percentage to avoid pressures toward faking failure.)

Conclusion

The problem is turned over to you unfinished and inadequately formalized. But I hope that I have convinced you that we need sociology of scientific validity, and an applied social science specialty within it, as a part of the methodology we bring to our tasks. I hope that you share my conviction that this can be done in a way that still makes valid applied social science possible (or, at very least, that we can produce beliefs of enough improved validity and subtlety to make continuing in our profession worthwhile). If you are convinced of both need and possibility, I call upon you vigorous youngsters to take up the task of creating an adequate social theory of validity-increasing applied social science. But if you are convinced of the impossibility, then it is your moral duty to publicly denounce the pseudo-science in which we inadvertently find ourselves engaged. Let us at very least create around the problem a mutually monitoring, disputatious community of scholars who listen carefully to each other's arguments and rebuttals.

839