American Psychologist
(1969) **24**: 409-429

REFORMS AS EXPERIMENTS[1]

DONALD T. CAMPBELL[2]
Northwestern University

The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available. Our readiness for this stage is indicated by the inclusion of specific provisions for program evaluation in the first wave of the "Great Society" legislation, and by the current congressional proposals for establishing "social indicators" and socially relevant "data banks". So long have we had good intentions in this regard that many may feel we are already at this stage, that we already are continuing or discontinuing programs on the basis of assessed effectiveness. It is a theme of this article that this is not at all so, that most ameliorative programs end up with *no* interpretable evaluation (Etzioni, 1968; Hyman & Wright, 1967; Schwartz, 1961). We must look hard at the sources of this condition, and design ways of overcoming the difficulties. This article is a preliminary effort in this regard.

Many of the difficulties lie in the intransigencies of the research setting and in the presence of recurrent seductive pitfalls of interpretation. The bulk of this article will be devoted to these problems. But the few available solutions turn out to depend upon correct administrative decisions in the initiation and execution of the program. These decisions are made in a political arena, and involve political jeopardies that are often sufficient to explain the lack of hard-headed evaluation of effects. Removing reform administrators from the political spotlight seems both highly unlikely, and undesirable even if it were possible. What is instead essential is that the social scientist research advisor understand the political realities of the situation, and that he aid by helping create a public demand for hard-headed evaluation, by contributing to those political interventions that reduce the liability of honest evaluation, and by educating future administrators to the problems and possibilities.

---

[1] The preparation of this paper has been supported by National Science Foundation Grant GS1309X. Versions of this paper have been presented as the Northwestern University Alumni Fund Lecture, January 24, 1968; to the Social Psychology Section of the British Psychological Society at Oxford, September 20, 1968; to the International Conference on Social Psychology at Prague, October 7, 1968 (under a different title); and to several other groups.

[2] Requests for reprints should be sent to Donald T. Campbell, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

For this reason, there is also an attempt in this article to consider the political setting of program evaluation, and to offer suggestions as to political postures that might further a truly experimental approach to social reform. Although such considerations will be distributed as a minor theme throughout this article, it seems convenient to begin with some general points of this political nature.

Political Vulnerability from Knowing Outcomes

It is one of the most characteristic aspects of the present situation that *specific reforms are advocated as though they were certain to be successful.* For this reason, knowing outcomes has immediate political implications. Given the inherent difficulty of making significant improvements by the means usually provided and given the discrepancy between promise and possibility, most administrators wisely prefer to limit the evaluations to those the outcomes of which they can control, particularly insofar as published outcomes or press releases are concerned. Ambiguity, lack of truly comparable comparison bases, and lack of concrete evidence all work to increase the administrator's control over what gets said, or at least to reduce the bite of criticism in the case of actual failure. There is safety under the cloak of ignorance. Over and above this tie-in of advocacy and administration, there is another source of vulnerability in that the facts relevant to experimental program evaluation are also available to argue the general efficiency and honesty of administrators. The public availability of such facts reduces the privacy and security of at least some administrators.

Even where there are ideological commitments to a hard-headed evaluation or organizational efficiency, or to a scientific organization of society, these two jeopardies lead to the failure to evaluate organizational experiments realistically. If the political and administrative system has committed itself in advance to the correctness of efficacy of its reforms, it cannot tolerate learning of failure. To be truly scientific we must be able to experiment. We must be able to advocate without that excess of commitment that blinds us to reality testing.

This predicament, abetted by public apathy and by deliberate corruption, may prove in the long run to permanently preclude a truly experimental approach to social amelioration. But our needs and our hopes for a better society demand we make the effort. There are a few signs of hope. In the United States we have been able to achieve cost-of-living and unemployment indices that, however imperfect, have embarrassed the administrations that published them. We are able to conduct censuses that reduce the number of representatives a state has in Congress. These are grounds for optimism, although the corrupt tardiness of state governments in following their own constitutions in revising legislative districts illustrates the problem.

One simple shift in political posture which would reduce the problem is the shift from the advocacy of a specific reform to the advocacy of the seriousness of the problem, and hence to the advocacy of persistence in alternative reform efforts should the first one fail. The political stance would become: "This is a serious problem. We propose to initiate Policy A on an experimental basis. If after five years there has been no significant improvement, we will shift to Policy B." By making explicit that a given problem solution was only one of several that the administrator or party could in good conscience advocate, and by having ready a plausible alternative, the administrator could afford honest evaluation of outcomes. Negative results, a failure of the first program, would not jeopardize his job, for his job would be to keep after the problem until something was found that worked.

Coupled with this should be a general moratorium on ad hominum evaluation research, that is, on research designed to evaluate specific administrators rather than alternative policies. If we worry about the invasion-of-privacy problem in the data banks and social indicators of the future (e.g. Sawyer & Schechter, 1968), the touchiest point is the privacy of administrators. If we threaten this, the measurement system will surely be sabotaged in the innumerable ways possible. While this may sound unduly pessimistic, the recurrent anecdotes of administrators attempting to squelch unwanted research findings convince me of its accuracy. But we should be able to evaluate those alternative policies that a given administrator has the option of implementing.

Field Experiments and Quasi-Experimental Designs

In efforts to extend the logic of laboratory experimentation into the "field", and into settings not fully experimental, an inventory of threats to experimental validity has been assembled, in terms of which some 15 or 20 experimental and quasi-experimental designs have been evaluated (Campbell, 1957, 1963; Campbell & Stanley, 1963). In the present article only three or four designs will be examined, and therefore not all of the validity threats will be relevant, but it will provide useful background to look briefly at them all. Following are nine threats to internal validity.[3]

1. *History*: events, other than the experimental treatment, occurring between pretest and posttest and thus providing alternate explanations of effects.
2. *Maturation*: processes within the respondents or observed social units producing changes as a function of the passage of time per se, such as growth, fatigue, secular trends, etc.
3. *Instability*: unreliability of measures, fluctuations in sampling persons or components, autonomous instability of repeated or "equivalent" measures. (This is the only threat to which statistical tests of significance are relevant.)
4. *Testing*: the effect of taking a test upon the scores of a second testing. The effect of publication of a social indicator upon subsequent readings of that indicator.
5. *Instrumentation*: in which changes in the calibration of a measuring instrument or changes in the observers or scores used may produce changes in the obtained measurements.
6. *Regression artifacts*: pseudo-shifts occurring when persons or treatment units have been selected upon the basis of their extreme scores.
7. *Selection*: biases resulting from differential recruitment of comparison groups, producing different mean levels on the measure of effects.
8. *Experimental mortality*: the differential loss of respondents from comparison groups.

[3] This list has been expanded from the major previous presentations by the addition of *Instability* (but see Campbell, 1968; Campbell & Ross, 1968). This has been done in reaction to the sociological discussion of the use of tests of significance in nonexperimental or quasi-experimental research (e.g. Selvin, 1957; and as reviewed by Galtung, 1967, pp. 353-389). On the one hand, I join with the critics in criticizing the exaggerated status of "statistically significant differences" in establishing convictions of validity. Statistical tests are relevant to at best 1 out of 15 or so threats to validity. On the other hand, I join with those who defend their use in situations where randomization has not been employed. Even in those situations, it is relevant to say or to deny, "This is a trivial difference. It is of the order that would have occurred frequently *had* these measures been assigned to these classes solely by chance." Tests of significance, making use of random reassignments of the actual scores, are particularly useful in communicating this point.

9. *Selection-maturation interaction*: selection biases resulting in differential rates of "maturation" or autonomous change.

If a change or difference occurs, these are rival explanations that could be used to explain away an effect and thus to deny that in this specific experiment any genuine effect of the experimental treatment had been demonstrated. These are faults that true experiments avoid, primarily through the use of randomization and control groups. In the approach here advocated, this checklist is used to evaluate specific quasi-experimental designs. This is evaluation, not rejection, for it often turns out that for a specific design in a specific setting the threat is implausible, or that there are supplementary data that can help rule it out even where randomization is impossible. The general ethic, here advocated for public administrators as well as social scientists, is to use the very best method possible, aiming at "true experiments" with random control groups. But where randomized treatments are not possible, a self-critical use of quasi-experimental designs is advocated. We must do the best we can with what is available to us.

Our posture vis-à-vis perfectionist critics from laboratory experimentation is more militant than this: the only threats to validity that we will allow to invalidate an experiment are those that admit of the status of empirical laws more dependable and more plausible than the law involving the treatment. The mere possibility of some alternative explanation is not enough – it is only the *plausible* rival hypotheses that are invalidating. Vis-à-vis correlational studies and common-sense descriptive studies, on the other hand, our stance is one of greater conservatism. For example, because of the specific methodological trap of regression artifacts, the sociological tradition of "ex post facto" designs (Chapin, 1947; Greenwood, 1945) is totally rejected (Campbell & Stanley, 1963, pp. 240-241; 1966, pp. 70-71).

Threats to external validity, which follow, cover the validity problems involved in interpreting experimental results, the threats to valid generalization of the results to other settings, or to other measures of the effect:[4]

1. *Interaction effects of testing*: the effect of a pretest in increasing or decreasing the respondent's sensitivity or responsiveness to the experimental variable, thus making the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.
2. *Interaction of selection and experimental treatment*: unrepresentative responsiveness of the treated population.
3. *Reactive effects of experimental arrangements*: "artificiality"; conditions making the experimental setting atypical of conditions of regular application of the treatment: "Hawthorne effects".
4. *Multiple-treatment interference*: where multiple treatments are jointly applied, effects atypical of the separate application of the treatments.
5. *Irrelevant responsiveness of measures*: all measures are complex, and all include irrelevant components that may produce apparent effects.
6. *Irrelevant replicability of treatments*: treatments are complex, and replications of them may fail to include those components actually responsible for the effects.

[4] This list has been lengthened from previous presentations to make more salient Threats 5 and 6 which are particularly relevant to social experimentation. Discussion in previous presentations (Campbell, 1957, pp. 309-310; Campbell & Stanley, 1963, pp. 203-204) had covered these points, but they had not been included in the checklist.

These threats apply equally to true experiments and quasi-experiments. They are particularly relevant to applied experimentation. In the cumulative history of our methodology, this class of threats was first noted as a critique of true experiments involving pretests (Schanck & Goodman, 1939; Solomon, 1949). Such experiments provided a sound basis for generalizing to other *pretested* populations, but the reactions of unpretested populations to the treatment might well be quite different. As a result, there has been an advocacy of true experimental designs obviating the pretest (Campbell, 1957; Schanck & Goodman, 1939; Solomon, 1949) and a search for nonreactive measures (Webb, Campbell, Schwartz, & Sechrest, 1966).

These threats to validity will serve as a background against which we will discuss several research designs particularly appropriate for evaluating specific programs of social amelioration. These are the "interrupted time-series design", the "control series design", "regression discontinuity design", and various "true experiments". The order is from a weak but generally available design to stronger ones that require more administrative foresight and determination.

Interrupted Time-Series Design

By and large, when a political unit initiates a reform it is put into effect across the board, with the total unit being affected. In this setting the only comparison base is the record of previous years. The usual mode of utilization is a casual version of a very weak quasi-experimental design, the one-group pretest-posttest design.

A convenient illustration comes from the 1955 Connecticut crackdown on speeding, which Sociologist H. Laurence Ross and I have been analyzing as a methodological illustration (Campbell & Ross, 1968; Glass, 1968; Ross & Campbell, 1968). After a record high of traffic fatalities in 1955, Governor Abraham Ribicoff instituted an unprecedentedly severe crackdown on speeding. At the end of a year of such enforcement there had been but 284 traffic deaths as compared with 324 the year before. In announcing this the Governor stated "With the saving of 40 lives in 1956, a reduction of 12.3% from the 1955 motor vehicle death toll we can say that the program is definitely worth-while." These results are graphed in Figure 1 with a deliberate effort to make them look impressive.

In what follows, while we in the end decide that the crackdown had some beneficial effects, we criticise Ribicoff's interpretation of his results from the point of view of the social scientist's proper standards of evidence. Were the now Senator Ribicoff not the man of stature that he is, this would be most unpolitic, because we could be alienating one of the strongest proponents of social experimentation in our nation. Given his character, however, we may feel sure that he shares our interests both in a progressive program of experimental social amelioration, and in making the most hard-headed evaluation possible of the experiments. Indeed, it was his integrity in using every available means at his disposal as Governor to make sure that the unpopular speeding crackdown was indeed enforced that make these data worth examining at all. But the potentials of that one illustration and our political temptation to substitute for it a less touchy one, point to the political problems that must be faced in experimenting with social reform.

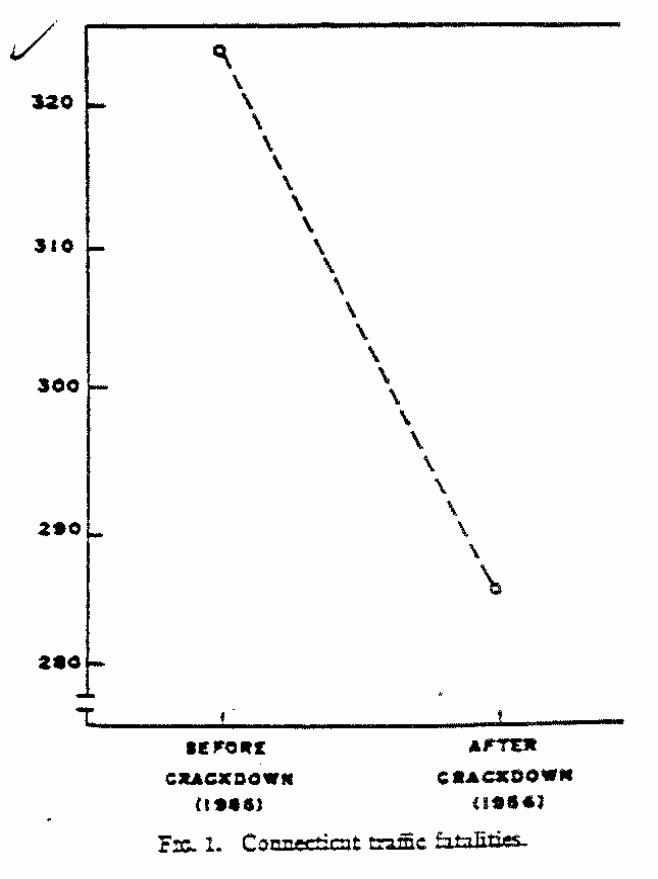

FIG. 1. Connecticut traffic fatalities.

Keeping Figure 1 and Ribicoff's statement in mind, let us look at the same data presented as a part of an extended time series in Figure 2, and go over the relevant threats to internal validity. First, *History* Both presentations fail to control for the effects of other potential change agents. For instance, 1956 might have been a particularly dry year, with fewer accidents due to rain or snow. Or there might have been a dramatic increase in use of seat belts, or other safety features. The advocated strategy in quasi-experimentation is not to throw up one's hands and refuse to use the evidence because of this lack of control, but rather to generate by informed criticism appropriate to this specific setting as many *plausible* rival hypotheses as possible, and then to do the supplementary research, as into weather records and safety-belt sales, for example, which would reflect on these rival hypotheses.

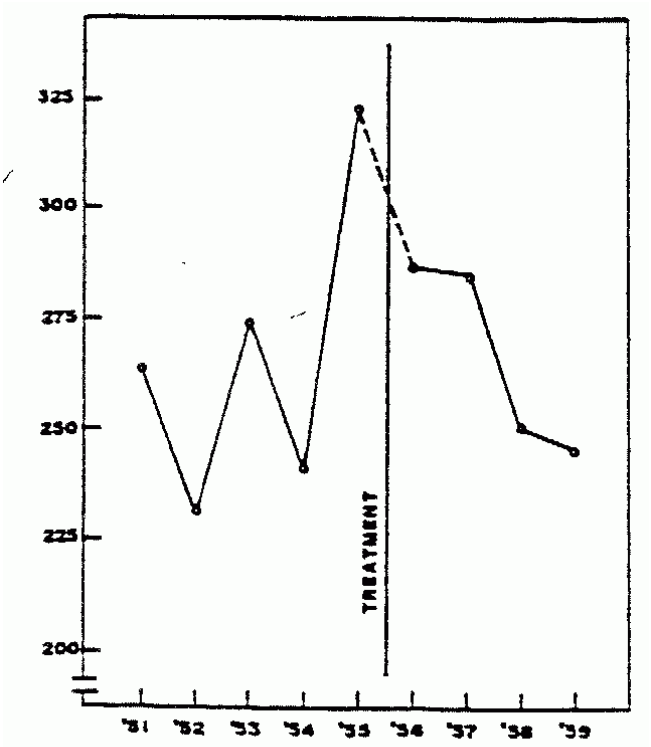

FIG. 2. Connecticut traffic fatalities. (Same data as in Figure 1 presented as part of an extended time series.)

Maturation. This is a term coming from criticisms of training studies of children. Applied here to the simple pretest-posttest data of Figure 1, it could be the plausible rival hypothesis that death rates were steadily going down year after year (as indeed they are, relative to miles driven or population of automobiles). Here the extended time series has a strong methodological advantage, and rules out this threat to validity. The general trend is inconsistently up prior to the crackdown, and steadily down thereafter.

Instability. Seemingly implicit in the public pronouncement was the assumption that all of the change from 1955 to 1956 was due to the crackdown. There was no recognition of the fact that all time series are unstable even when no treatments are being applied. The degree of this normal instability is the crucial issue, and one of the main advantages of the extended time series is that it samples this instability. The great pre-treatment instability now makes the treatment effect look relatively trivial. The 1955-56 shift is less than the gains of both 1954-55 and 1952-53. It is the largest drop in the series, but it exceeds the drops of 1951-52, 1953-54, and 1957-58 by trivial amounts. Thus the unexplained instabilities of the series are such as to make the 1955-56 drop understandable as more of the same. On the other hand, it is noteworthy that after the crackdown there are no year-to-year gains, and in this respect the character of the time series seems definitely to have changed.

The threat of instability is the only threat to which tests of significance are relevant. Box and Tiao (1965) have an elegant Bayesian model for the interrupted time series. Applied by Glass (1968) to our monthly data, with seasonal trends removed, it shows a statistically significant downward shift in the series after the crackdown. But as we shall see, an alternative explanation of at least part of this significant effect exists.

Regression. In true experiments the treatment is applied independently of the prior state of the units. In natural experiments exposure to treatment is often a cosymptom of the treated group's condition. The treatment is apt to be an *effect* rather than, or in

addition to being, a cause. Psychotherapy is such a cosymptom treatment, as is any other in which the treated group is self-selected or assigned on the basis of need. These all present special problems of interpretation, of which the present illustration provides one type.

The selection-regression plausible rival hypothesis works this way: Given that the fatality rate has some degree of unreliability, then a subsample selected for its extremity in 1955 would on the average, merely as a reflection of that unreliability, be less extreme in 1956. Has there been selection for extremity in applying this treatment? Probably yes. Of all Connecticut fatality years, the most likely time for a crackdown would be after an exceptionally high year. If the time series showed instability, the subsequent year would on the average be les, *purely as a function of that instability*. Regression artifacts are probably the most recurrent form of self-deception in the experimental social reform literature. It is hard to make them intuitively obvious. Let us try again. Take any time series with variability, including one generated of pure error. Move along it as in a time dimension. Pick a point that is the "highest so far". Look then at the next point. On the average this next point will be lower, or nearer the general trend.

In our present setting the most striking shift in the whole series is the upward shift just prior to the crackdown. It is highly probable that this caused the crackdown, rather than, or in addition to, the crackdown causing the 1956 drop. At least part of the 1956 drop is an artifact of the 1955 extremity. While in principle the degree of expected regression can be computed from the auto-correlation of the series we lack here an extended-enough body of data to do this with any confidence.

Advice to administrators who want to do genuine reality-testing must include attention to this problem, and it will be a very hard problem to surmount. The most general advice would be to work on chronic problems of a persistent urgency or extremity, rather than reacting to momentary extremes. The administrator should look at the pre-treatment time series to judge whether or not instability plus momentary extremity will explain away his program gains. If it will, he should schedule the treatment for a year or two later, so that his decision is more independent of the one year's extremity. (The selection biases remaining under such a procedure need further examination.)

In giving advice to the *experimental* administrator, one is also inevitably giving advice to those *trapped* administrators whose political predicament requires a favorable outcome whether valid or not. To such trapped administrators the advice is pick the very worst year, and the very worst social unit. If there is inherent instability, there is nowhere to go but up, for the average case at least.

Two other threats to internal validity need discussion in regard to this design. By *testing* we typically have in mind the condition under which a test of attitude, ability, or personality is itself a change agent, persuading, informing, practicing, or otherwise setting processes of change in action. No artificially introduced testing procedures are involved here. However, for the simple before-and-after design of Figure 1, if the pretest were the first data collection of its kind ever publicized, this publicity in itself might produce a reduction in traffic deaths which would have taken place even without a speeding crackdown. Many traffic safety programs assume this. The long time-series evidence reassures us on this only to the extent that we can assume that the figures had been published each year with equivalent emphasis.[5]

[5] No doubt the public and press shared the Governor's special alarm over the 1955 death toll. This differential reaction could be seen as a negative feedback servosystem in which the dampening effect was proportional to the degree of upward deviation from the prior trend. Insofar as such alarm reduces

Instrumentation changes are not a likely flaw in this instance, but would be if recording practices and institutional responsibility had shifted simultaneously with the crackdown. Probably in a case like this it is better to use raw frequencies rather than indices whose correction parameters are subject to periodic revision. Thus per capita rates are subject to periodic jumps as new census figures become available correcting old extrapolations. Analogously, a change in the miles per gallon assumed in estimating traffic mileage for mileage-based mortality rates might explain a shift. Such biases can of course work to disguise a true effect. Almost certainly, Ribicoff's crackdown reduced traffic speed (Campbell & Ross, 1968). Such a decrease in speed increases the miles per gallon actually obtained, producing a concomitant drop in the estimate of miles driven, which would appear as an inflation of the estimate of mileage-based traffic fatalities if the same fixed approximation to actual miles per gallon were used, as it undoubtedly would be.

The "new broom" that introduces abrupt changes of policy is apt to reform the record keeping too, and thus confound reform treatments with instrumentation change. The ideal experimental administrator will, if possible, avoid doing this. He will prefer to keep comparable a partially imperfect measuring system rather than lose comparability altogether. The politics of the situation do not always make this possible, however. Consider, as an experimental reform, Orlando Wilson's reorganization of the police system in Chicago. Figure 3 shows his impact on petty larceny in Chicago – a striking *increase*! Wilson, of course, called this shot in advance, one aspect of his reform being a reform in the bookkeeping. (Note in the pre-Wilson records the suspicious absence of the expected upward secular trend.) In this situation Wilson had no choice. Had he left the record keeping as it was, for the purposes of better experimental design, his resentful patrolmen would have clobbered him with a crime wave by deliberately starting to record the many complaints that had not been getting into the books.[6]

---

traffic fatalities, it adds a negative component to the autocorrelation, increasing the regression effect. This component should probably be regarded as a rival cause or treatment rather than as artifact. (The regression effect is less as the positive autocorrelation is higher, and will be present to some degree insofar as this correlation is less than positive unity. Negative correlation in a time series would represent regression beyond the mean, in a way not quite analogous to negative correlation across persons. For an autocorrelation of Lag 1, high negative correlation would be represented by a series that oscillated maximally from one extreme to the other.)

[6] Wilson's inconsistency in utilization of records and the political problem of relevant records are ably documented in Kamisar (1964). Etzioni (1968) reports that in New York City in 1965 a crime wave was proclaimed that turned out to be due to an unpublicized improvement in record keeping.

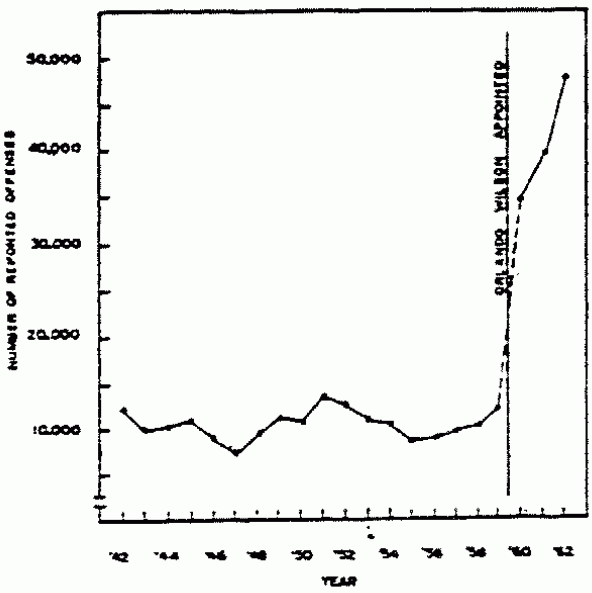

FIG. 3. Number of reported larcenies under $50 in Chicago, Illinois, from 1942 to 1962 (data from *Uniform Crime Reports for the United States*, 1942–62).

Those who advocate the use of archival measures as social indicators (Bauer, 1966; Gross, 1966, 1967; Kaysen, 1967; Webb et al., 1966) must face up not only to their high degree of chaotic error and systematic bias, but also to the politically motivated changes in record keeping that will follow upon their public use as social indicators (Etzioni & Lehman, 1967). Not all measures are equally susceptible. In Figure 4, Orland Wilson's effect on homicides seems negligible one way or the other.

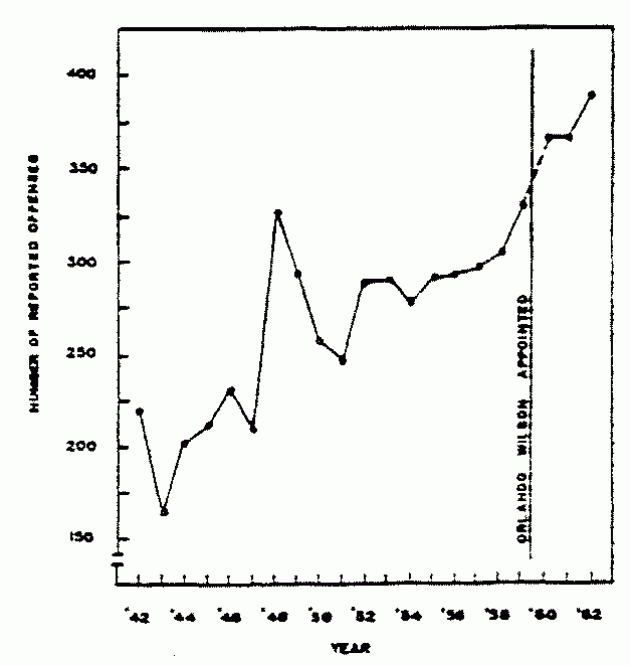

FIG. 4. Number of reported murders and nonnegligent manslaughters in Chicago, Illinois, from 1942 to 1962 (data from *Uniform Crime Reports for the United States*, 1942–62).

Of the threats to external validity, the one most relevant to social experimentation is *Irrelevant Responsiveness of Measures*. This seems best discussed in terms of the problem of generalizing from indicator to indicator or in terms of the imperfect validity of all measures that is only to be overcome by the use of multiple measures of independent imperfection (Campbell & Fiske, 1959; Webb et al., 1966).

For treatments on any given problem within any given governmental or business subunit, there will usually be something of a governmental monopoly on reform. Even though different divisions may optimally be trying different reforms, within each division there will usually be only one reform on a given problem going on at a time. But for measures of effect this need not and should not be the case. The administrative machinery should itself make multiple measures of potential benefits and of unwanted side effects. In addition, the loyal opposition should be allowed to add still other indicators, with the political process and adversary argument challenging both validity and relative importance, with social science methodologists testifying for both parties, and with the basic records kept public and under bipartisan audit (as are voting records under optimal conditions). This competitive scrutiny is indeed the main source of objectivity in sciences (Polanyi, 1966, 1967; Popper, 1963) and epitomizes an ideal of democratic practice in both judicial and legislative procedures.

The next few figures return again to the Connecticut crackdown on speeding and look to some other measures of effect. They are relevant to the confirming that there was indeed a crackdown and to the issue of side effects. They also provide the methodological comfort of assuring us that in some cases the interrupted time-series design can provide clear-cut evidence of effect. Figure 5 shows the jump in suspensions of licenses for speeding – evidence that severe punishment was abruptly instituted. Again a note to experimental administrators: with this weak design, *it is only abrupt and decisive changes that we have any chance of evaluating*. A gradually introduced reform will be indistinguishable from the background of secular change, from the net effect of the innumerable change agents continually impinging.

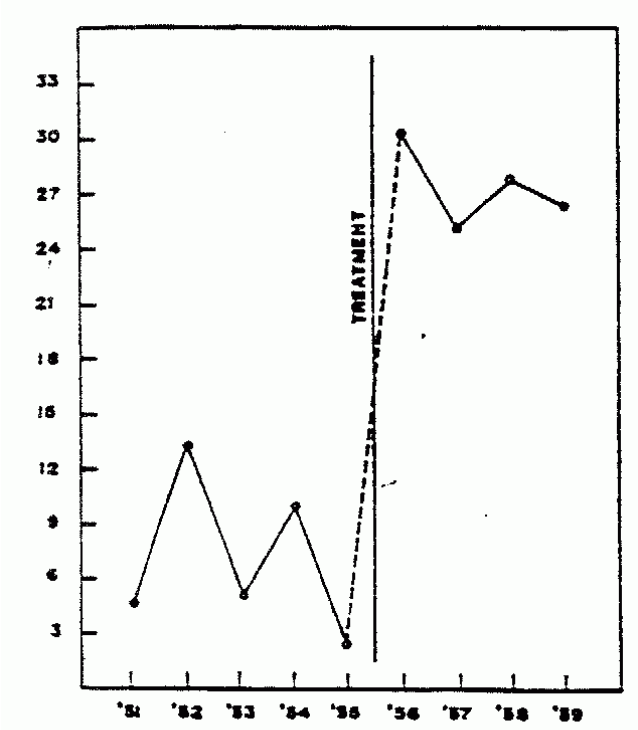

FIG. 5. Suspensions of licenses for speeding, as a percentage of all suspensions.

We would want intermediate evidence that traffic speed was modified. A sampling each year of a few hundred five-minute highway movies (random as to location and time) could have provided this at a moderate cost, but they were not collected. Of the public records available, perhaps the data of Figure 6, showing a reduction in speeding violations, indicate a reduction in traffic speed. But the effects on the legal system were complex, and in part undesirable. Driving with a suspended license markedly increased (Figure 7), at least in the biased sample of those arrested. Presumably because of the harshness of the punishment if guilty, judges may have become more lenient (Figure 8) although this effect is of marginal significance.

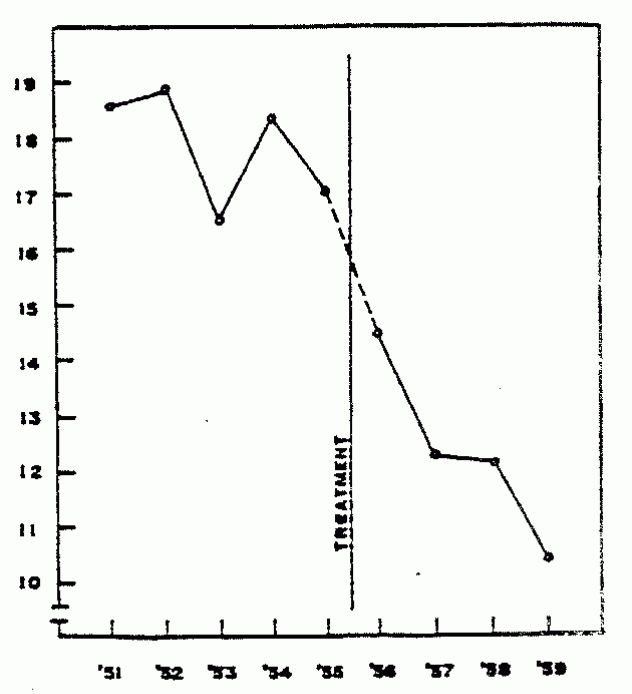

FIG. 6. Speeding violations, as a percentage of all traffic violations.

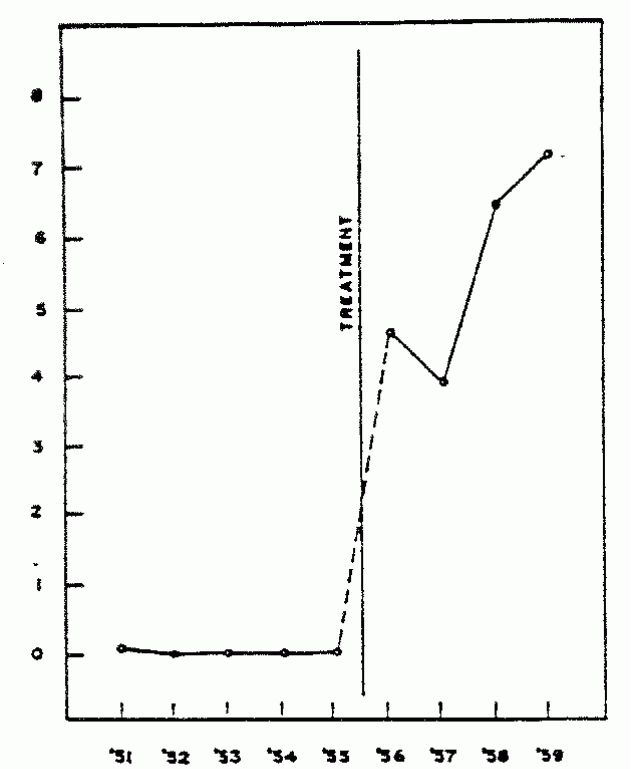

FIG. 7. Arrested while driving with a suspended license, as a percentage of suspensions.

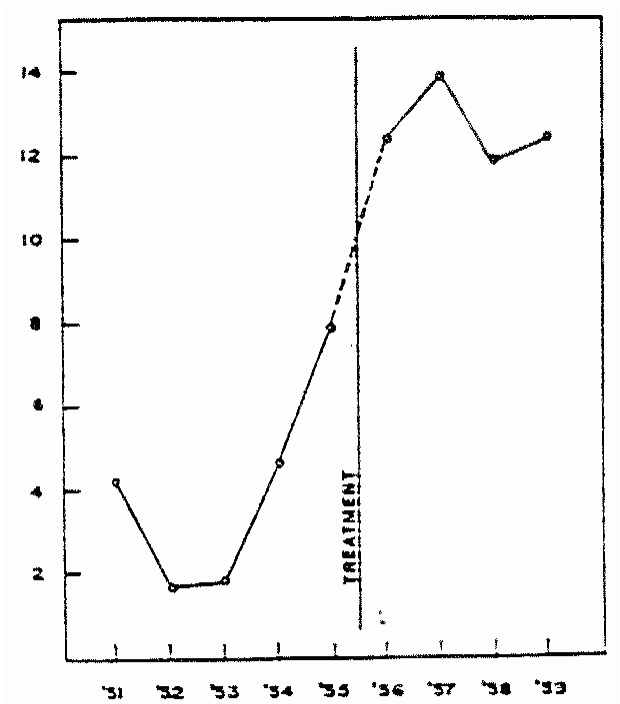

FIG. 8. Percentage of speeding violations judged not guilty.

The relevance of indicators for the social problems we wish to cure must be kept continually in focus. The social indicators approach will tend to make the indicators themselves the goal of social action, rather than the social problems they but imperfectly indicate. There are apt to be tendencies to legislate changes in the indicators per se rather than changes in the social problems.

To illustrate the problem of the irrelevant responsiveness of measures, Figure 9 shows a result of the 1900 change in divorce law in Germany. In a recent reanalysis of the data with the Box and Tiao (1965) statistic, Glass (Glass, Tiao, & Maguire, 1969) has found the change highly significant, in contrast to earlier statistical analyses (Rheinstein, 1959; Wolf, Lüke, & Hax, 1959). But Rheinstein's emphasis would still be relevant: This indicator change indicates no likely improvement in marital harmony, or even in marital stability. Rather than reducing them, the legal change has made the divorce rate a less valid indicator of marital discord and separation than it had been earlier (see also Etzioni & Lehman, 1967).

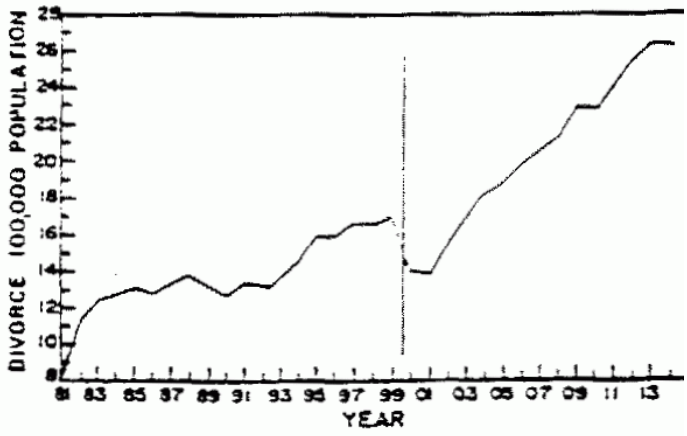

FIG. 9. Divorce rate for German Empire, 1881–1914

Control Series Design

The interrupted time-series design as discussed so far is available for those settings in which no control group is possible, in which the total governmental unit has received the experimental treatment, the social reform measure. In the general program of quasi-experimental design, we argue the great advantage of untreated comparison groups even where these cannot be assigned at random. The most common of such designs is the non-equivalent control-group pretest-posttest design, in which for each of two natural groups, one of which receives the treatment, a pretest and posttest measure is taken. If the traditional mistaken practice is avoided of matching on pretest scores (with resultant regression artifacts), this design provides a useful control over those aspects of history, maturation, and test-retest effects shared by both groups. But it does not control for the plausible rival hypothesis of *selection-maturation interaction* – that is, the hypothesis that the selection differences in the natural aggregations involve not only differences in mean level, but differences in maturation rate.

This point can be illustrated in terms of the traditional quasi-experimental design problem of the effects of Latin on English vocabulary (Campbell, 1963). In the hypothetical data of Figure 10B, two alternative interpretations remain open. Latin may have had effect, for those taking Latin gained more than those not. But on the other hand, those students taking Latin may have a greater annual rate of vocabulary growth that would manifest itself whether or not they took Latin. Extending this common design into two time series provides relevant evidence, as comparison of the two alternative outcomes of Figure 10C and 10D shows. Thus approaching quasi-experimental design from either improving the non-equivalent control-group design or from improving the interrupted time-series design, we arrive at the control series design. Figure 11 shows this for the Connecticut speeding crackdown, adding evidence from the fatality rates of neighboring states. Here the data are presented as population-based fatality rates so as to make the two series of comparable magnitude.

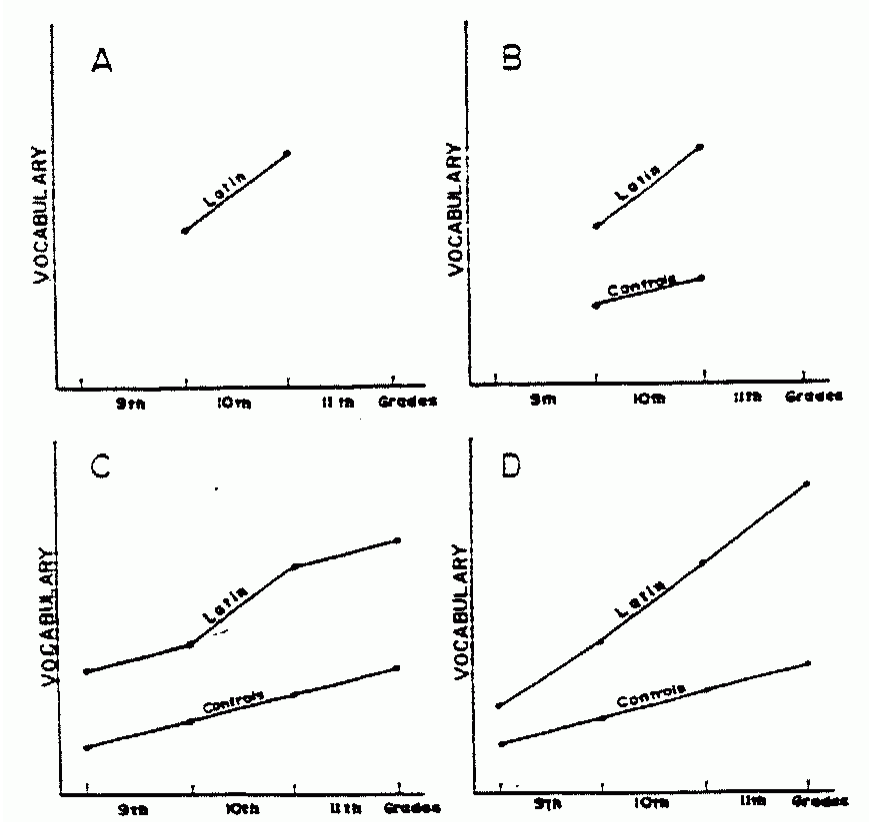

FIG. 10. Forms of quasi-experimental analysis for the effect of specific course work, including control series design.

The control series design of Figure 11 shows that downward trends were available in the other states for 1955-56 as due to history and maturation, that is, due to shared secular trends, weather, automotive safety features, etc. But the data also show a general trend for Connecticut to rise relatively closer to the other states prior to 1955, and to steadily drop more rapidly than other states from 1956 on. Glass (1968) has used our monthly data for Connecticut and the control states to generate a monthly difference score, and this too shows a significant shift in trend in the Box and Tiao (1965) statistic. Impressed particularly by the 1957, 1958, and 1959 trend, we are willing to conclude that the crackdown had some effect, over and above the undeniable pseudo-effects of regression (Campbell & Ross, 1968).

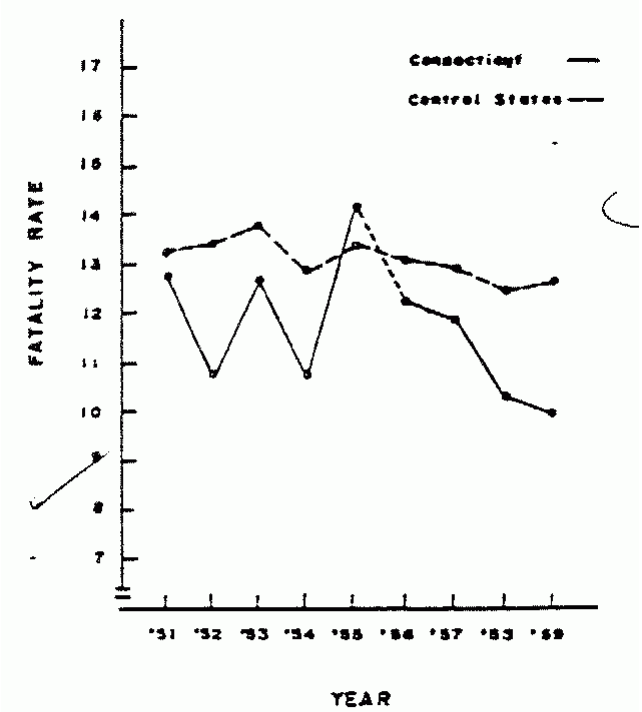

FIG. 11. Control series design comparing Connecticut fatalities with those of four comparable states.

The advantages of the control series design point to the advantages for social experimentation of a social system allowing subunit diversity. Our ability to estimate the effects of the speeding crackdown, Rose's (1952) and Stieber (1949) ability to estimate the effects on strikes of compulsory arbitration laws, and Simon's (1966) ability to estimate the price elasticity of liquor were made possible because the changes were not being put into effect in all states simultaneously, because they were matters of state legislation rather than national. I do not want to appear to justify on these grounds the wasteful and unjust diversity of laws and enforcement practices from state to state. But I would strongly advocate that social engineers make use of this diversity while it remains available, and plan cooperatively their changes in administrative policy and in record keeping so as to provide optimal experimental inference. More important is the recommendation that, for those aspects of social reform handled by the central government, a purposeful diversity of implementation be envisaged so that experimental and control groups be available for analysis. Properly planned, these can approach true experiments, better than the casual and ad hoc comparison groups now available. But without such fundamental planning, uniform central control can reduce the present possibilities of reality testing, that is, of true social experimentation. In the same spirit, decentralization of decision making, both within large government and within private monopolies, can provide a useful competition for both efficiency and innovation, reflected in a multiplicity of indicators.

Regression Discontinuity Design

We shift now to social ameliorations that are in short supply, and that therefore cannot be given to all individuals. Such scarcity is inevitable under many circumstances, and can make possible an evaluation of effects that would otherwise be impossible. Consider the heroic Salk poliomyelitis vaccine trials in which some

children were given the vaccine while others were given an inert saline placebo injection – and in which many more of these placebo controls would die than would have if they had been given the vaccine. Creation of these placebo controls would have been morally, psychologically, and socially impossible had there been enough vaccine for all. As it was, due to the scarcity, most children that year had to go without the vaccine anyway. The creation of experimental and control groups was the highly moral allocation of that scarcity so as to enable us to learn the true efficacy of the supposed good. The usual medical practice of introducing new cures on a so-called trial basis in general medical practice makes evaluation impossible by confounding prior status with treatment, that is, giving the drug to the most needy or most hopeless. It has the further social bias of giving the supposed benefit to those most assiduous in keeping their medical needs in the attention of the medical profession, that is, the upper and upper-middle classes. The political stance furthering social experimentation here is the recognition of randomization as the most democratic and moral means of allocating scarce resources (and scarce hazardous duties), plus the moral imperative to further utilize the randomization so that society may indeed learn true value of the supposed boon. This is the ideology that makes possible "true experiments" in a large class of social reforms.

But if randomization is not politically feasible or morally justifiable in a given setting, there is a powerful quasi-experimental design available that allows the scarce good to be given to the most needy or the most deserving. This is the regression discontinuity design. All it requires is strict and orderly attention to the priority dimension. The design originated through an advocacy of a tie-breaking experiment to measure the effects of receiving a fellowship (Thistlethwaite & Campbell, 1960), and it seems easiest to explain it in that light. Consider as in Figure 12, pre-award ability-and-merit dimension, which would have some relation to later success in life (finishing college, earnings 10 years later, etc.). Those higher on the premeasure are most deserving and receive the award. They do better in later life, but does the award have an effect? It is normally impossible to say because they would have done better in later life anyway. Full randomization of the award was impossible given the stated intention to reward merit and ability. But it might be possible to take a narrow band of ability at the cutting point, to regard all of these persons as tied, and to assign half of them to awards, half to no awards, by means of a tie-breaking randomization.

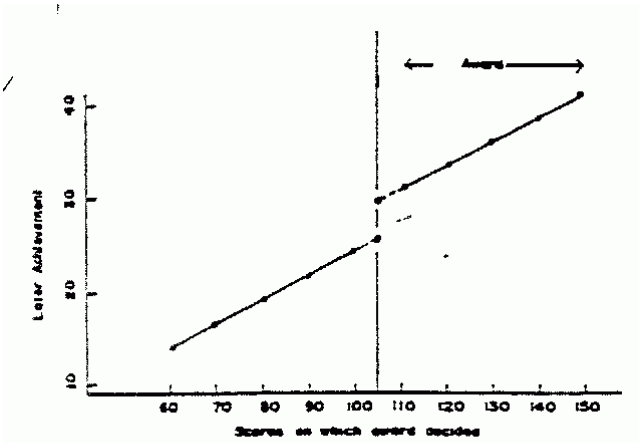

FIG. 12. Tie-breaking experiment and regression discontinuity analysis.

The tie-breaking rationale is still worth doing, but in considering that design it became obvious that, if the regression of premeasure on later effects were reasonably orderly, one should be able to extrapolate to the results of the tie-breaking experiment by plotting the regression of posttest on pretest separately for those in the award and non-award regions. If there is no significant difference for these at the decision-point intercept, then the tie-breaking experiment should show no difference. In cases where the tie breakers would show an effect, there should be an abrupt discontinuity in the regression line. Such a discontinuity cannot be explained away by the normal regression of the posttest on pretest, for this normal regression, as extensively sampled within the nonaward area and within the award area, provides no such expectation.

Figure 12 presents, in terms of column means, an instance in which higher pretest scores would have led to higher posttest scores even without the treatment, and in which there is in addition a substantial treatment effect. Figure 13 shows a series of paired outcomes, those on the left to be interpreted as no effect, those in the center and on the right as effect. Note some particular cases. In instances of granting opportunity on the basis of merit, like 13a and b (and Figure 12), neglect of the background regression of pretest on posttest leads to optimistic pseudo-effects: in Figure 13a, those receiving the award do do better in later life, though not really because of the award. But in social ameliorative efforts, the setting is more apt to be like Figure 13d and e, where neglect of the background regression is apt to make the program look deleterious if no effect, or ineffective if there is a real effect.

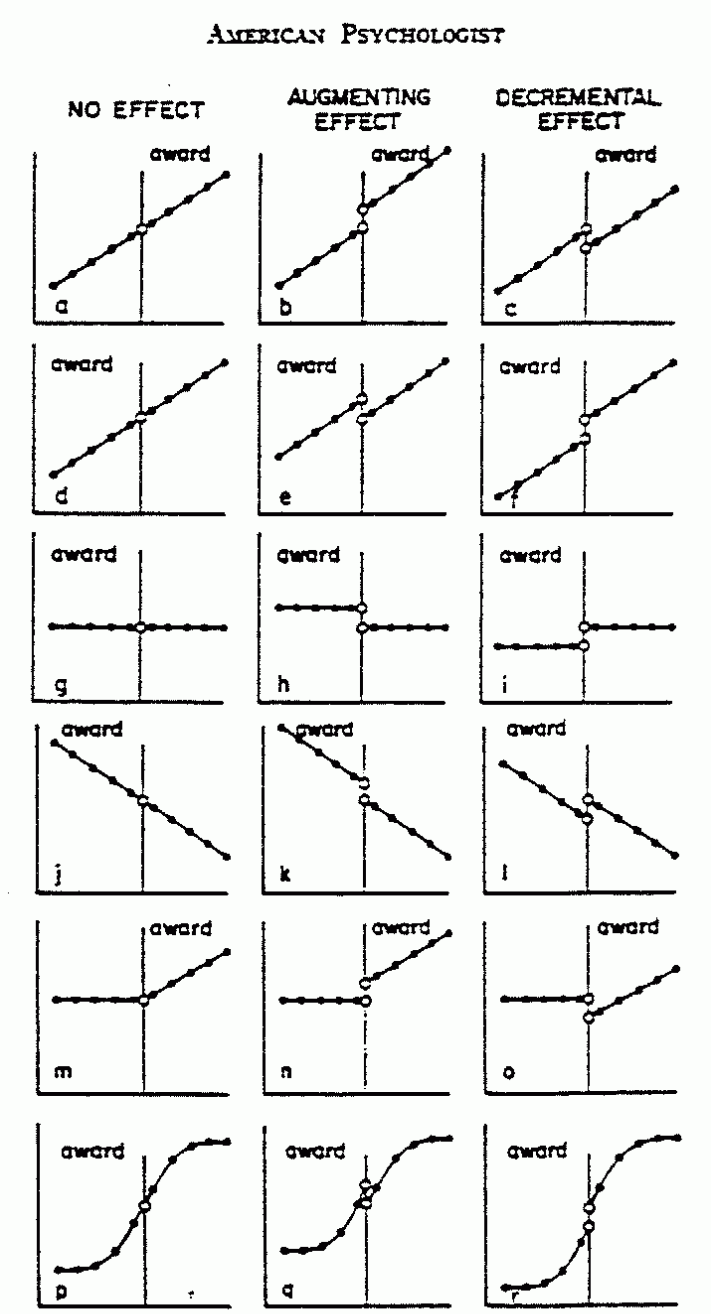

FIG. 13. Illustrative outcomes of regression discontinuity analyses.

The design will of course work just as well or better if the award dimension and the decision base, the pretest measure, are unrelated to the posttest dimension, if it is irrelevant or unfair, as instanced in Figure 13 g, h, and i. In such cases the decision base is the functional equivalent of randomization. Negative background relationships are obviously possible, as in Figure 13j, k, and l. In Figure 13, m, n, and o are included to emphasize that it is a jump in intercept at the cutting point that shows effect, and that differences in slope without differences at the cutting point are not acceptable as evidences of effect. This becomes more obvious if we remember that in cases like m, a tie-breaking randomization experiment would have shown no difference. Curvilinear background relationships as in Figure 13 p, q, and r, will provide added obstacles to clear inference in many instances, where sampling error could make Figure 13p look like 13b.

As further illustration, Figure 14 provides computer-simulated data, showing individual observations and fitted regression lines, in a fuller version of the no-effect outcomes of Figure 13a. Figure 15 shows an outcome with effect. These have been generated[7] by assigning to each individual a weighted normal random number as a "true score" to which is added a weighted independent "error" to generate the "pretest". The "true score" plus another independent "error" produces the "posttest" in no-effect cases such as Figure 14. In treatment-effect simulations, as in Figure 15, there are added into the posttest "effect points" for all "treated" cases, that is, those above the cutting point on the pretest score.[8]

[7] J. Sween & D.T. Campbell, Computer programs for simulating and analyzing sharp and fuzzy regression-discontinuity experiments. In preparation.

[8] While at least one workable test of significance is available, it may prove quite difficult to achieve a test that preserves the imagery of extrapolating to a hypothetical tie-breaking randomization experiment. Initially, following Walker and Lev (1953, p. 400; Sween & Campbell, 1965, p. 7), we tested the significance of the difference of the cutting-point intercepts of two regression lines (of posttest on pretest), one fitted to the observations below the cutting-point and one fitted to the observations above. In computer simulations of no-effect cases, "significant" pseudo-effects were repeatedly found. It turns out that this is one of those settings in which the least-squares solution is biased. The nature of the bias can perhaps be communicated by considering what would happen if both the regression lines of pretest-on posttest and of posttest-on-pretest were to be plotted for the distribution as a whole. These two regression lines will cross at the center of the distribution (i.e., at the cutting point in symmetrical examples such as in Figures 14 and 15) and will fan apart at the ends. When these two regressions are fitted instead to the half distributions, they will cross in the center of each half and fan apart at the cutting point. In an example such as Figure 14, the regression of posttest on pretest will be the lower one at the cutting point for the no-treatment half, and the higher one for the treatment half. This pseudo-effect does not appear for plots of actual column means, as visually inspected, and Figures 14, 15, 16, and 17 should have been drawn with actual column means plotted instead of the fitted straight lines. The amount of this bias is a function of the pretest-posttest correlation and if this can be properly estimated, corrected cutting-point intercept estimates could be computed. However, the whole distribution cannot be used to estimate this correlation, for a true effect will cause part of the correlation. An estimate might be based upon the correlations computed separately above and below the cutting point, correcting for restricted range. Maximum likelihood estimation procedures may also be found.

At the moment, the best suggestion seems to be that provided by Robert P. Abelson. The posttest-on-pretest regression is fitted for a body of data extending above and below the cutting point in equal degrees. Column means are expressed as departures from that regression. A *t* test is then used to compare the columns just below and just above the cutting point. To increase the actuarial base, wider column definitions can be explored. This test unfortunately loses the analogy to the tie-breaking true experiment, an analogy that the present writer has been dependent upon for conceptual clarification.

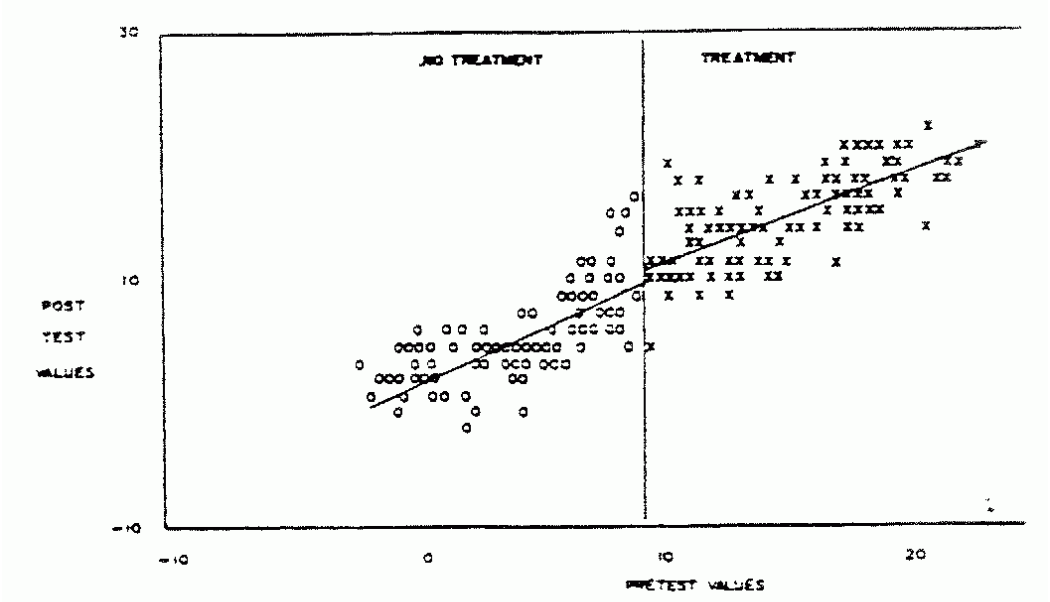

FIG. 14. Regression discontinuity design: No effect.

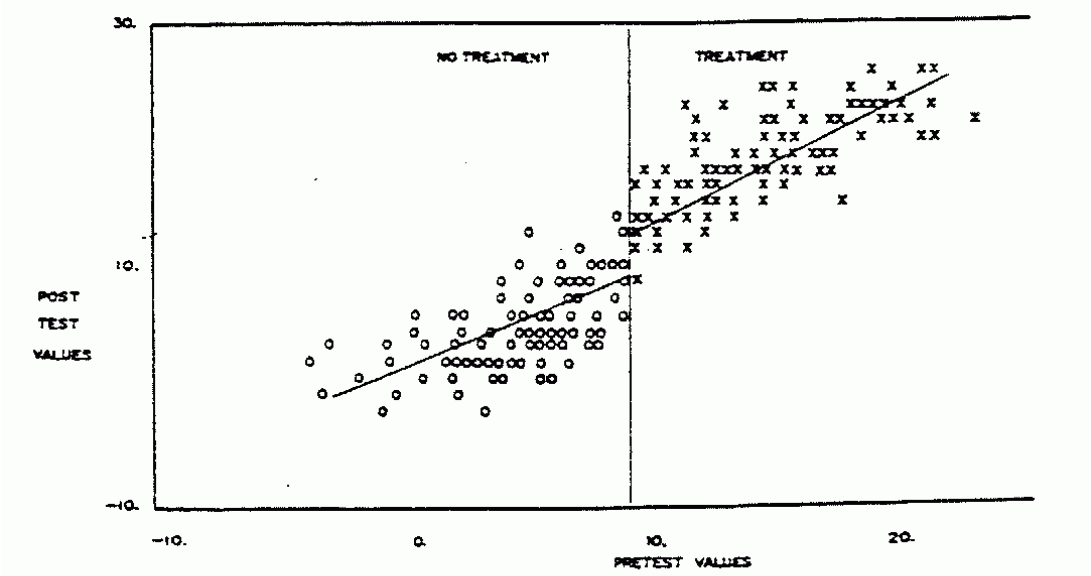

FIG. 15. Regression discontinuity design: Genuine effect.

This design could be used in a number of settings. Consider Job Training Corps applicants, in larger number than the program can accommodate, with eligibility determined by need. The setting would be as in Figure 13d and e. The base-line decision dimension could be per capita family income, with those at below the cut-off getting training. The outcome dimension could be the amount of withholding tax withheld two years later, or the percentage drawing unemployment insurance, these follow-up figures being provided from the National Data Bank in response to categorized social security numbers fed in, without individual anonymity being breached, without any real invasion of privacy – for it is the program that is being examined, via regularities of aggregates of persons. While the plotted points could be named, there is no need that they be named. In a classic field experiment on tax compliance, Richard Schwartz and the Bureau of Internal Revenue have managed to put together sets of personally identified interviews and tax-return data so that

statistical analyses such as these can be done, without the separate custodians of either interview or tax returns learning the corresponding data for specific persons (Schwartz & Orleans, 1967; see also Schwartz & Skilnick, 1963). Manniche and Hayes (1957) have spelled out how a broker can be used in a two-staged matching of doubly coded data. Kaysen (1967) and Sawyer and Schechter (1968) have wise discussions of the more general problem.

What is required of the administrator of a scarce ameliorative commodity to use this design? Most essential is a sharp cut-off point on a decision-criterion dimension, on which several other qualitatively similar analytic cut-offs can be made both above and below the award cut. Let me explain this better by explaining why National Merit scholarships were unable to use the design for their actual fellowship decision (although it has been used for their Certificate of Merit). In their operation, diverse committees make small numbers of award decisions by considering a group of candidates and then picking from them the *N* best to which to award the *N* fellowships allocated them. This provides one cutting point on an unspecified pooled decision base, but fails to provide analogous potential cutting points above and below. What could be done is for each committee to collectively rank its group of 20 or so candidates. The top *N* would then receive the award. Pooling cases across committees, cases could be classified according to number of ranks above and below the cutting point, these other ranks being analogous to the award-nonaward cutting point as far as regression onto posttreatment measures was concerned. Such group ranking would be costly of committee time. An equally good procedure, if committees agreed, would be to have each member, after full discussion and freedom to revise, give each candidate a grade, A+, A, A-, B+, B, etc., and to award the fellowships to the *N* candidates averaging best on these ratings, with no revisions allowed after the averaging process. These ranking or rating units, even if not comparable from committee to committee in range of talent, in number of persons ranked, or in cutting point, could be pooled without bias as far as a regression discontinuity is concerned, for that range of units above and below the cutting point in which all committees were represented.

It is the dimensionality and sharpness of the decision criterion that is at issue, not its components or validity. The ratings could be based upon nepotism, whimsy, and superstition and still serve. As has been stated, if the decision criterion is utterly invalid we approach the pure randomness of a true experiment. Thus the weakness of subjective committee decisions is not their subjectivity, but the fact that they provide only the one cutting point on their net subjective dimension. Even in the form of average ratings the recommended procedures probably represent some slight increase in committee work load. But this could be justified to the decision committees by the fact that through refusals, etc., it cannot be known at the time of the committee meeting the exact number to whom the fellowship can be offered. Other costs at the planning time are likewise minimal. The primary additional burden is in keeping as good records on the nonawardees as on the awardees. Thus at a low cost, an experimental administrator can lay the groundwork for later scientific follow-ups, the budgets for which need not yet be in sight.

Our present situation is more apt to be one where our pre-treatment measures, aptitude measures, reference ratings, etc., can be combined via multiple correlation into an index that correlates highly but not perfectly with the award decision. For this dimension there is a fuzzy cut-off point. Can the design be used in this case? Probably not. Figure 16 shows the pseudo-effect possible if the award decision contributes any valid variance to the quantified pretest evidence, as it usually will. The award regression rides above the nonaward regression just because of that valid variance in

this simulated case, there being no true award effect at all. (In simulating this case, the award decision has been based upon a composite of true score plus an independent award error.) Figure 17 shows a fuzzy cutting point plus a genuine award effect.[9] The recommendation to the administrator is clear: aim for a sharp cutting point on a quantified decision criterion. If there are complex rules for eligibility, only one of which is quantified, seek out for follow-up that subset of persons for whom the quantitative dimension was determinate. If political patronage necessitates some decisions inconsistent with a sharp cutoff, record these cases under the heading "qualitative decision rule" and keep them out of your experimental analysis.

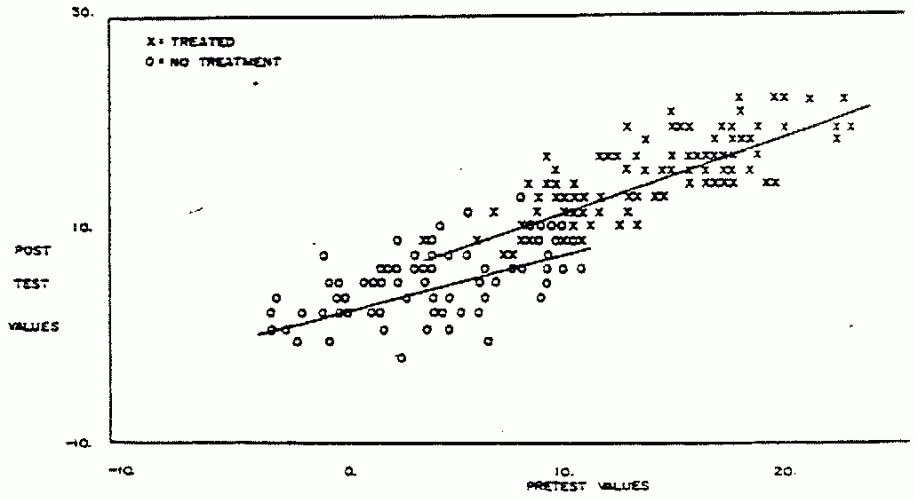

FIG. 16. Regression discontinuity design: Fuzzy cutting point, pseudo treatment effect only.

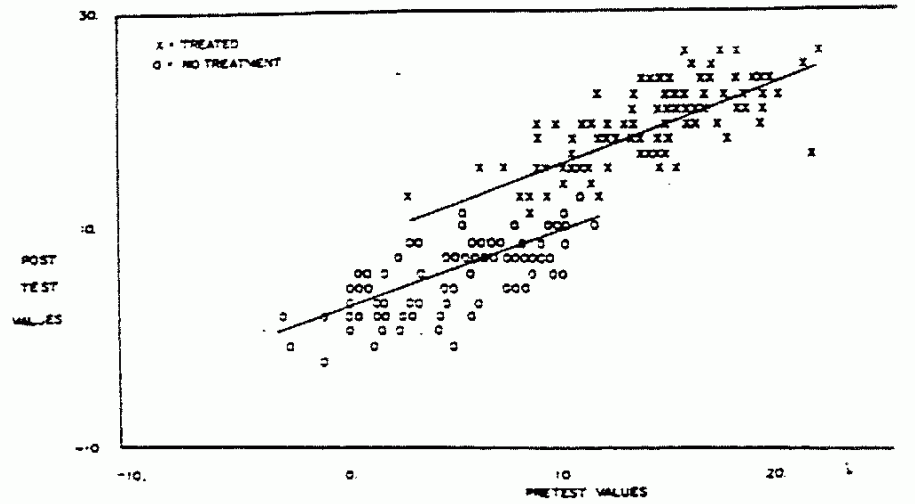

FIG. 17. Regression discontinuity design: Fuzzy cutting point, with real treatment plus pseudo treatment effects.

[9] There are some subtle statistical clues that might distinguish these two instances if one had enough cases. There should be increased pooled column variance in the mixed columns for a true effects case. If the data are arbitrarily treated as though there had been a sharp cutting point located in the middle of the overlap area, then there should be no discontinuity in the no-effect case, and some discontinuity in the case of a real effect, albeit an underestimated discontinuity, since there are untreated cases above the cutting point and treated ones below, dampening the apparent effect. The degree of such dampening should be estimable, and correctable, perhaps by iterative procedures. But these are hopes for the future.

Almost all of our ameliorative programs designed for the disadvantaged could be studied via this design, and so too some major governmental actions affecting the lives of citizens in ways we do not think of as experimental. For example, for a considerable period, quantitative test scores have been used to call up for military service or reject as unfit at the lower ability range. If these cutting points, test scores, names, and social security numbers have been recorded for a number of steps both above and below the cutting point, we could make elegant studies of the effect of military service on later withholding taxes, mortality, number of dependents, etc. Unfortunately for this purpose, the nobly experimental "Operation 100,000" (Office of the Secretary of Defense, 1967) is fuzzing up this cutting point, but there are several years of earlier Vietnam experience all ready for analysis.

This illustration points to one of the threats to external validity of this design, or of the tie-breaking experiment. The effect of the treatment has only been studied for that narrow range of talent near the cutting point, and generalization of the effects of military service, for example, from this low ability level to the careers of the most able would be hazardous in the extreme. But in the draft laws and the requirements of the military services there may be other sharp cutting points on a quantitative criterion that could also be used. For example, those over 6 feet 6 inches are excluded from service. Imagine a five-year-later follow-up of draftees grouped by inch in the 6 feet 1 inch to 6 feet 5 inches range, and a group of their counterparts who would have been drafted except for their heights, 6 feet 6 inches to 6 feet 10 inches. (The fact that the other grounds of deferment might not have been examined by the draft board would be a problem here, but probably not insurmountable.) That we should not expect height in this range to have any relation to later-life variables is not at all a weakness of this design, and if we have indeed a subpopulation for which there is a sharp numerical cutting point, an internally valid measure of effects would result. Deferment under the present system is an unquantified committee decision. But just as the sense of justice of United States soldiers was quantified through paired comparisons of cases into an acceptable Demobilization Points system at the end of World War II (Guttman, 1946; Stouffer, 1949), so a quantified composite index of deferment priority could be achieved and applied as uniform justice across the nation, providing another numerical cutting point.

In addition to the National Data Bank type of indicators, there will be occasions in which new data collections as by interview or questionnaire are needed. For these there is the special problem of uneven cooperation that would be classified as instrumentation error. In our traditional mode of thinking, completeness of description is valued more highly than comparability. Thus if, in a fellowship study, a follow-up mailed out from the fellowship office would bring a higher return from past winners, this might seem desirable even if the nonawardees' rate of response was much lower. From the point of view of quasi-experimentation, however, it would be better to use an independent survey agency and a disguised purpose, achieving equally low response rates from both awardees and nonawardees, and avoiding a regression discontinuity in cooperation rate that might be misinterpreted as a discontinuity in more important effects.

Randomized Control Group Experiments

Experiments with randomization tend to be limited to the laboratory and agricultural experiment station. But this certainly need not be so. The randomization

unit may be persons, families, precincts, or larger administrative units. For statistical purposes the randomization units should be numerous, and hence ideally small. But for reasons of external validity, including reactive arrangements, the randomization units should be selected on the basis of the units of administrative access. Where policies are administered through individual client contacts, randomization at the person level may be often inconspicuously achieved, with the clients unaware that different ones of them are getting different treatments. But for most social reforms, larger administrative units will be involved, such as classrooms, schools, cities, counties, or states. We need to develop the political postures and ideologies that make randomization at these levels possible.

"Pilot project" is a useful term already in our political vocabulary. It designates a trial program that, if it works, will be spread to other areas. By modifying actual practice in this regard, without going outside of the popular understanding of the term, a valuable experimental ideology could be developed. How are areas selected for pilot projects? If the public worries about this, it probably assumes a lobbying process in which the greater needs of some areas are only one consideration, political power and expediency being others. Without violating the public tolerance or intent, one could probably devise a system in which the usual lobbying decided upon the areas eligible for a formal public lottery that would make final choices between match pairs. Such decision procedures as the drawing of lots have had a justly esteemed position since time immemorial (e.g., Aubert, 1959). At the present time, record keeping for pilot projects tends to be limited to the experimental group only. In the experimental ideology, comparable data would be collected on designated controls. (There are of course exceptions, as in the heroic Public Health Service fluoridation experiments, in which the teeth of Oak Park children were examined year after year as controls for the Evanston experimentals [Blayney & Hill, 1967].)

Another general political stance making possible experimental social amelioration is that of *staged innovation*. Even though by intent a new reform is to be put into effect in all units, the logistics of the situation usually dictate that simultaneous introduction is not possible. What results is a haphazard sequence of convenience. Under the program of staged innovation, the introduction of the program would be deliberately spread out, and those units selected to be first and last would be randomly assigned (perhaps randomization from matched pairs), so that during the transition period the first recipients could be analyzed as experimental units, the last recipients as controls. A third ideology making possible true experiments has already been discussed: randomization as the democratic means of allocating scarce resources.

This article will not give true experimentation equal space with quasi-experimentation only because excellent discussions of, and statistical consultation on, true experimentation are readily available. True experiments should almost always be preferred to quasi-experiments where both are available. Only occasionally are the threats to external validity so much greater for the true experiment that one would prefer a quasi-experiment. The uneven allocation of space here should not be read as indicating otherwise.

More Advice for Trapped Administrators

But the competition is not really between the fairly interpretable quasi-experiments here reviewed and "true" experiments. Both stand together as rare excellencies in contrast with a morass of obfuscation and self-deception. Both to emphasize this contrast, and again as guidelines for the benefit of those trapped

administrators whose political predicament will not allow the risk of failure, some of these alternatives should be mentioned.

Grateful testimonials. Human courtesy and gratitude being what it is, the most dependable means of assuring a favorable evaluation is to use voluntary testimonials from those who have had the treatment. If the spontaneously produced testimonials are in short supply, these should be solicited from the recipients with whom the program is still in contact. The rosy glow resulting is analogous to the professor's impression of his teaching success when it is based solely upon the comments of those students who come up and talk with him after class. In many programs, as in psychotherapy, the recipient, as well as the agency, has devoted much time and effort to the program and it is dissonance reducing for himself, as well as common courtesy to his therapist, to report improvement. These grateful testimonials can come in the language of letters and conversation, or be framed as answers to multiple-item "tests" in which a recurrent theme of "I am sick", "I am well", "I am happy", "I am sad" recurs. Probably the testimonials will be more favorable as: (*a*) the more the evaluative meaning of the response measure is clear to the recipient – it is completely clear in most personality, adjustment, morale, and attitude tests; (*b*) the more directly the recipient is identified with his answer; (*c*) the more the recipient gives the answer directly to the therapist or agent of reform; (*d*) the more the agent will continue to be influential in the recipient's life in the future; (*e*) the more the answers deal with feelings and evaluations rather than with verifiable facts; and (*f*) the more the recipients participating in the evaluation are a small and self-selected or agent-selected subset of all recipients. Properly designed, the grateful testimonial method can involve pretests as well as posttests, and randomized control groups as well as experimentals, for there are usually no placebo treatments, and the recipients know when they have had the boon.

Confounding selection and treatment. Another dependable tactic bound to give favorable outcomes is to confound selection and treatment, so that in the published comparison those receiving the treatment are also the more able and well placed. The often-cited evidence of the dollar value of a college education is of this nature – all careful studies show that most of the effect, and of the superior effect of superior colleges, is explainable in terms of superior talents and family connections, rather than in terms of what is learned or even the prestige of the degree. Matching techniques and statistical partialings generally undermatch and do not fully control for the selection differences – they introduce regression artifacts confusable as treatment effects.

There are two types of situations that must be distinguished. First, there are those treatments that are given to the most promising, treatments like a college education which are regularly given to those who need it least. For these, the later concomitants of the grounds of selection operate in the same direction as the treatment: those most likely to achieve anyway get into the college most likely to produce later achievement. For these settings, the trapped administrator should use the pooled mean of all those treated, comparing it with the mean of all untreated, although in this setting almost any comparison any administrator might hit upon would be biased in his favor.

At the other end of the talent continuum are those remedial treatments given to those who need it most. Here the later concomitants of the grounds of selection are poorer success. In the Job Training Corps example, casual comparisons of the later unemployment rate of those who received the training with those who did not are in general biased against showing an advantage to the training. Here the trapped

administrator must be careful to seek out those few special comparisons biasing selection in his favor. For training programs such as Operation Head Start and tutoring programs, a useful solution is to compare the later success of those who completed the training program with those who were invited but never showed plus those who came a few times and dropped out. By regarding only those who completed the program as "trained" and using the others as controls, one is selecting for conscientiousness, stable and supporting family backgrounds, enjoyment of training activity, ability, determination to get ahead in the world – all factors promising well for future achievement even if the remedial program is valueless. To apply this tactic effectively in the Job Training Corps, one might have to eliminate from the so-called control group all those who quite the training program because they had found a job – but this would seem a reasonable practice and would not blemish the reception of a glowing progress report.

These are but two more samples of well-tried modes of analysis for the trapped administrator who cannot afford an honest evaluation of the social reform he directs. They remind us again that we must help create a political climate that demands more rigorous and less self-deceptive reality testing. We must provide political stances that permit true experiments, or good quasi-experiment. Of the several suggestions toward this end that are contained in this article, the most important is probably the initial theme: Administrators and parties must advocate the importance of the problem rather than the importance of the answer. They must advocate experimental sequences of reforms, rather than one certain cure-all, advocating Reform A with Alternative B available to try next should an honest evaluation of A prove it worthless or harmful.

Multiple Replication in Enactment

Too many social scientists expect single experiments to settle issues once and for all. This may be a mistaken generalization from the history of great crucial experiments in physics and chemistry. In actuality the significant experiments in the physical sciences are replicated thousands of times, not only in deliberate replication efforts, but also as inevitable incidentals in successive experimentation and in utilizations of those many measurement devices (such as the galvanometer) that in their own operation embody the principles of classic experiments. Because we social scientists have less ability to achieve "experimental isolation", because we have good reason to expect our treatment effects to interact significantly with a wide variety of social factors many of which we have not yet mapped, we have much greater needs for replication experiments than do the physical sciences.

The implications are clear. We should not only do hard-headed reality testing in the initial pilot testing and choosing of which reform to make general law; but once it has been decided that the reform is to be adopted as standard practice in all administrative units, we should experimentally evaluate it in each of its implementations (Campbell, 1967).

CONCLUSIONS

Trapped Administrators have so committed themselves in advance to the efficacy of the reform that they cannot afford honest evaluation. For them favorably biased analyses are recommended, including capitalizing on regression, grateful testimonials, and confounding selection and treatment. *Experimental administrators* have justified the reform on the basis of the importance of the problem, not the certainty of their

answer, and are committed to going on to other potential solutions if the one first tried fails. They are therefore not threatened by a hard-headed analysis of the reform. For such, proper administrative decisions can lay the base for useful experimental or quasi-experimental analyses. Through the ideology of allocating scarce resources by lottery, through the use of staged innovation, and through the pilot project, true experiments with randomly assigned control groups can be achieved. If the reform must be introduced across the board, the interrupted time-series design is available. If there are similar units under independent administration, a control series design adds strength. If a scarce boon must be given to the most needy or to the most deserving, quantifying this need or merit makes possible the regression discontinuity analysis.

REFERENCES

AUBERT, V. (1959). Chance in social affairs. *Inquiry*, 2, 1-24.

BAUER, R.M. (1966). *Social indicators*. Cambridge, Mass: M.I.T. Press.

BLAYNEY, J.R. & HILL, I.N. (1967). Fluorine and dental caries. *The Journal of the American Dental Association* (Special Issue), 74, 233-302.

BOX, G.E.P. & TIAO, G.C. (1965). A change in level of a non-stationary time series. *Biometrika*, 52, 181-192.

CAMPBELL, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.

CAMPBELL, D.T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C.W. Harris (Ed.) *Problems in measuring change*. Madison: University of Wisconsin Press.

CAMPBELL, D.T. (1967). Administrative experimentation, institutional records, and nonreactive measures. In J.C. Stanley (Ed.) *Improving experimental design and statistical analysis*. Chicago: Rand McNally.

CAMPBELL, D.T. (1968). Quasi-experimental design. In D.L. Sills (Ed.) *International Encyclopedia of the Social Sciences*. New York: Macmillan and Free Press, Vol. 5, 259-263.

CAMPBELL, D.T. & FISKE, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

CAMPBELL, D.T. & ROSS, H.L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review*, 3(1), 33-53.

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

CHAPIN, F.S. (1947). *Experimental design in sociological research*. New York: Harper.

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

ETZIONI, A. & LEHMAN, E.W. (1967). Some dangers in "valid" social measurement. *Annals of the American Academy of Political and Social Science*, 373, 1-15.

GALTUNG, J. (1967). *Theory and methods of social research*. Oslo: Universitetsforloget; London: Allen & Unwin; New York: Columbia University Press.

GLASS, G.V. (1968). Analysis of data on the Connecticut speeding crackdown as a time-series experiment. *Law and Society Review*, 3(1), 55-76.

GLASS, G.V., TIAO, G.C., & MAGUIRE, T.O. (1969). Analysis of data on the 1900 revision of the German divorce laws as a quasi-experiment. *Law and Society Review*.

GREENWOOD, E. (1945). *Experimental sociology: A study in method*. New York: King's Crown Press.

GROSS, B.M. (1966). *The state of the nation: Social system accounting*. London: Tavistock Publications. (Also in R.M. Bauer (1966) *Social indicators*. Cambridge, Mass: M.I.T. Press.

GROSS, B.M. (1967). Social goals and indicators. *Annals of the American Academy of Political and Social Science*, 371, Part 1, May, pp. i-iii and 1-177; Part 2, September, pp. i-iii and 1-218.

GUTTMAN, L. (1946) An approach for quantifying paired comparisons and rank order. *Annals of Mathematical Statistics*, 17, 144-163.

HYMAN, H.H. & WRIGHT, C.R. (1967). Evaluating social action programs. In P.F. Lazarsfeld, W.H. Sewell, & H.L. Wilensky (Eds.) *The uses of sociology*. New York: Basic Books.

KAMISAR, Y. (1964). The tactics of police-persecution oriented critics of the courts. *Cornell Law Quarterly*, 4, 458-471.

KAYSEN, C. (1967). Data banks and dossiers. *The Public Interest*, 7, 52-60.

MANNICHE, E. & HAYES, D.P. (1957). Respondent anonymity and data matching. *Public Opinion Quarterly*, 21(3), 384-388.

OFFICE OF THE SECRETARY OF DEFENSE, Assistant Secretary of Defense (Manpower). (1967). Guidance paper: Project One Hundred Thousand. Washington, D.C., March 31, (Multilith).

POLANYI, M. (1966). A society of explorers. In *The tacit dimension*, (Chapter 3), New York: Doubleday.

POLANYI, M. (1967). The growth of science in society. *Minerva*, 5, 533-545.

POPPER, K.R. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul; New York: Basic Books.

RHEINSTEIN, M. (1959). Divorce and the law in Germany: A review. *American Journal of Sociology*, 65, 489-493.

ROSE, A.M. (1952). Needed research on the mediation of labor disputes. *Personnel Psychology*, 5, 157-200.

ROSS, H.L. & CAMPBELL, D.T. (1963). The Connecticut speed crackdown: A study of the effects of legal change. In H.L. Ross (Ed.), *Perspectives on the social order: Readings in sociology*. New York: McGraw-Hill.

SAWYER, J. & SCHECHTER, H. (1968). Computers, privacy, and the National Data Center: The responsibility of social scientists. *American Psychologist*, 23, 810-818.

SCHANCK, R.L. & GOODMAN, C. (1939). Reactions to propaganda on both sides of a controversial issue. *Public Opinion Quarterly*, 3, 107-112.

SCHWARTZ, R.D. (1961). Field experimentation in sociological research. *Journal of Legal Education*, 13, 401-410.

SCHWARTZ, R.D. & ORLEANS, S. (1967). On legal sanctions. *University of Chicago Law Review*, 34, 274-300.

SCHWARTZ, R.D. & SKOLNICK, J.H. (1963). Televised communication and income tax compliance. In L. Arons & M. May (Eds.), *Television and human behavior*. New York: Appleton-Century-Crofts.

SELVIN, H. (1957). A critique of tests of significance in survey research. *American Sociological Review,* 22, 519-527.

SIMON, J.L. (1966). The price elasticity of liquor in the U.S. and a simple method of determination. *Econometrica*34, 193 -205.

SOLOMON, R.W. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137-150.

STIEBER, J.W. (1949). *Ten years of the Minnesota Labor Relations Act*. Minneapolis: Industrial Relations Center, University of Minnesota.

STOUFFER, S.A. (1949). The point system for redeployment and discharge. In S.A. Stouffer et al, *The American Soldier. Vol. 2, Combat and its aftermath*. Princeton: Princeton University Press.

SUCHMAN, E.A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage.

SWEEN, J. & CAMPBELL, D.T. (1965). A study of the effect of proximally auto-correlated error on tests of significance for the interrupted time-series quasi-experimental design. Available from the author. (Multilith)

THISTLETHWAITE, D.L. & CAMPBELL, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309-317.

WALKER, H.M. & LEV, J. (1953). *Statistical inference*. New York: Holt.

WEBB, E.J., CAMPBELL, D.T., SCHWARTZ, R.D. & SECHREST, L.B. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.

WOLF, E., LUKE, G. & HAX, H. (1959). *Scheidung und Scheidungsrecht: Grundfrägen der Ehrescheidung in Deutschland*. Tübigen: J.C.B. Mohr.