Measuring the Effects of Social Innovations by Means of Time Series*

Donald T. Campbell *Lehigh University*

*Supported in part by NSF grant GS 1309X.

We live in an age of social reforms, of large-scale efforts to correct specific social problems. In the past most such efforts have not been adequately evaluated: usually there has been no scientifically valid evidence as to whether the problem was alleviated or not. Since there are always a variety of proposed solutions for any one problem, as well as numerous other problems calling for funds and attention, it becomes important that society be able to learn how effective any specific innovation has been.

From the statistician's point of view, the best designed experiments, whether in the laboratory or out in the community, involve setting up an *experimental group* and a *control group* similar in every way possible to the experimental group except that it does not receive the same experimental treatment. The statistician's way of achieving this all-purpose equivalence of experimental and control groups is randomization. Persons (or plots of land or other units) are assigned at random (as by the roll of dice) to either an experimental or a

control group. After the treatment, the two groups are compared, and the differences that are larger than chance would explain are attributed to the experimental treatment. This ideal procedure is beginning to be used in pilot tests of social policy, as in the New Jersey negative income-tax experiment (Kershaw and Fair, 1976; Watts and Rees, 1977) where several hundred low-income employed families who agreed to cooperate were randomly assigned to experimental groups (which received income supplements of differing sizes) and a control group (which received no financial aid). The effects of this aid on the amounts of other earnings, on health, family stability, and the like were then studied. (Joseph Newhouse's essay in this book describing an experiment in health insurance illustrates this approach.)

Unfortunately, while such experimental designs are ideal, they are not often feasible. They are impossible to use, for example, in evaluating any new program that is applied to all citizens at once, as most legal changes are. In these more common situations much less satisfactory modes of experimental inference must suffice. The *interrupted time series design,* on which this paper will concentrate, is one of the most useful of these quasi-experimental designs. The proper interpretation of such data presents complex statistical problems, some of which are not yet adequately solved. This essay will touch upon a number of these problems, in terms of words and graphs rather than mathematical symbols. The discussion will start by considering two actual cases.

THE CONNECTICUT CRACKDOWN ON SPEEDING

On December 23, 1955, Connecticut instituted an exceptionally severe and prolonged crackdown on speeding. Like most public reporting of program effectiveness, the results were reported in terms of simple before-and-after measures:

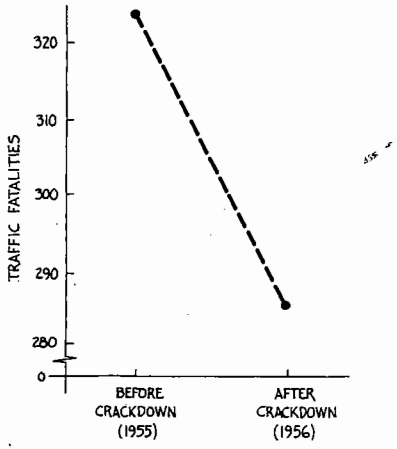

Figure 1 *Connecticut traffic fatalities, 1955–56. Source: Campbell and Ross &1968)*

a comparison of this year's figures with those of the year before. That is, the 1956 total of 284 traffic deaths was compared with the 1955 total of 324, and the governor stated, "With a saving of 40 lives in 1956 . . . we can say the program is definitely worthwhile." Figure 1 presents the data graphically. But this simple quasi-experimental design is very weak and deceptive. There are so many other possible explanations for the change from 324 to 284 highway fatalities. In attributing all of this change to his crackdown, the governor is making an implicit assumption that without the crackdown there would have been no change at all. A time series presentation, using the fatality records of several prior and subsequent years, adds greatly to the strength of the analysis. Figure 2 shows such data for the Connecticut crackdown. In this larger context the 1955–1956 drop looks trivial. We can see that the implicit assumption underlying the governor's statement was almost certainly wrong.

To explore this more fully, turn to Figure 3, which presents in a stylized manner how an identical shift in values before and after a treatment can in some instances be clear-cut evidence of an effect and in others be no evidence at all of a change. Thus with only 1955 and 1956 data to go on, that drop of 40 traffic fatalities shown in Figure 1 might have been a part of a steady annual drop already in progress (the reverse of the steady rise in line F of Figure 3), or of an unstable zigzag (as in line G of Figure 3). Figure 2 shows that in Connecticut the unstable zigzag is the case. The 1955–1956 drop is about the same size as the drops of 1951–1952, 1953–1954, and 1957–1958, times when no crackdowns were present to explain them. Furthermore, the 1955–1956 drop is only half the size of the 1954–1955 rise. Thus with all this previous instability in full graphic view, one would be unlikely to claim all of any year-to-year change as due to a crackdown, as the governor seemed to do.

Later we shall examine Figure 2 again to raise a more difficult problem of inference. But before we do this, let's spend more time on the stability issue, with the help of an illustration from a reform that even a skeptical methodologist can believe was successful.

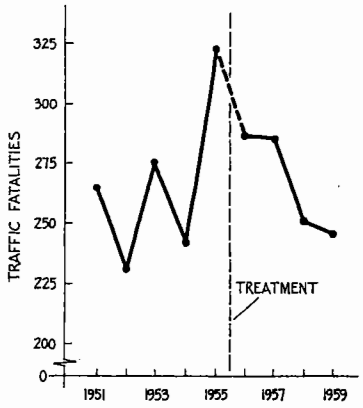

Figure 2 *Connecticut traffic fatalities, 1951–59. Source: Campbell and Ross (1968)*

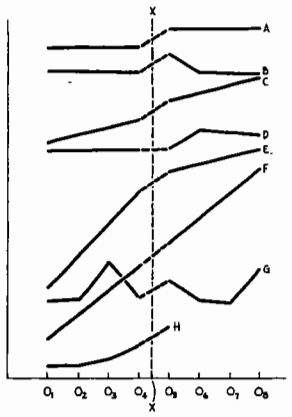

Figure 3 *Some possible outcome patterns from the introduction of a treatment at point X into a time series of measurements, O_1 to O_8. The O_4 to O_5 gain is the same for all time series, except for D, while the legitimacy of inferring an effect varies widely, being strongest in A and B, and totally unjustified in F, G, and H. Source: Campbell and Stanley (1963)*

THE BRITISH BREATHALYSER CRACKDOWN OF 1967

In September 1967 the British government started a new program of enforcement with regard to drunken driving. It took its popular name from a device for ascertaining the degree of intoxication from a sample of a person's breath. Police administered this simple test to drivers stopped on suspicion, and if it showed intoxication, then took them into the police station for more thorough tests. This new testing procedure was accompanied by more stringent punishment, including suspension of license. Figure 4 shows the effect of this crackdown on Friday and Saturday night casualties (fatalities plus serious injuries). The effect is dramatically clear. There is an immediate drop of around 40% and a leveling off at a level that seems some 30% below the precrackdown rate,

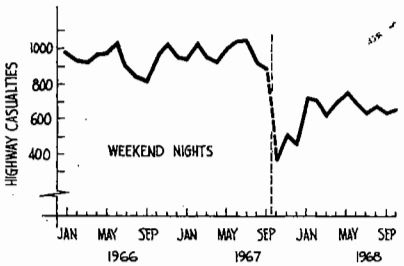

Figure 4 *Effects of the September 1967 English Breathalyser crackdown on drunken driving. Fatalities plus serious injuries, Fridays and Saturdays, 10:00 P.M. to 4:00 A.M., by month. Source: Ross, Campbell, and Glass (1970)*

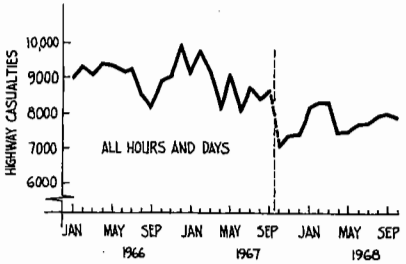

Figure 5 *Effects of the September 1967 English Breathalyser crackdown. Fatalities plus serious injuries, all hours and days, by month. Source: Ross, Campbell, and Glass (1970)*

although this is hard to tell for sure since we don't know what changes time would have brought in the casualty rate without the crackdown.

Does the effect show up when casualties at all hours of all days are totaled? Figure 5 shows such data. While the effect is probably still there, it is certainly less clear, the crackdown drop being not much larger than the unexplained instability of other time periods. (The crackdown drop is, however, the largest month-to-month change, not only during the plotted period but also for a longer period going back to 1961, for data from which the seasonal fluctuations have been removed.)

THE STATISTICAL ANALYSIS OF INSTABILITY

The problem of the statistician is to formalize the grounds for inference that we have used informally or intuitively in our judgments from these graphs. It is clear that the more unstable the line is before the policy change or treatment point, the bigger the difference has to be to impress us as a real effect. One approach of statisticians is to assume that the time series is a result of a general trend plus specific random deviations at each time period. The theory of this type of analysis is well worked out when the random deviations at each point are completely independent of deviations at other points. But in real-life situations the sources of deviation or perturbation at any one point are apt to be similar for adjacent and near points in time, and dissimilar for more remote points. This creates deceptive situations both for statistical tests of significance and for visual interpretation. Figures 6 and 7 illustrate this with a computer-simulated time series. For each point in time, times 1 to 40, there is a *true score.* These true scores, if plotted, would make a straight diagonal line from a lower left score of 0 to an upper right score of 40, with no bump whatsoever at the hypothetical treatment point. These true scores are the same for Figures 6 and 7. To each true score a randomly chosen deviation has been added or subtracted. In Figure 6 the deviation at each time point is drawn independently of every other deviation. This is a simulation of the case in which the hypothetical treatment introduced between time periods 20 and 21 has no effect at all. Occa-

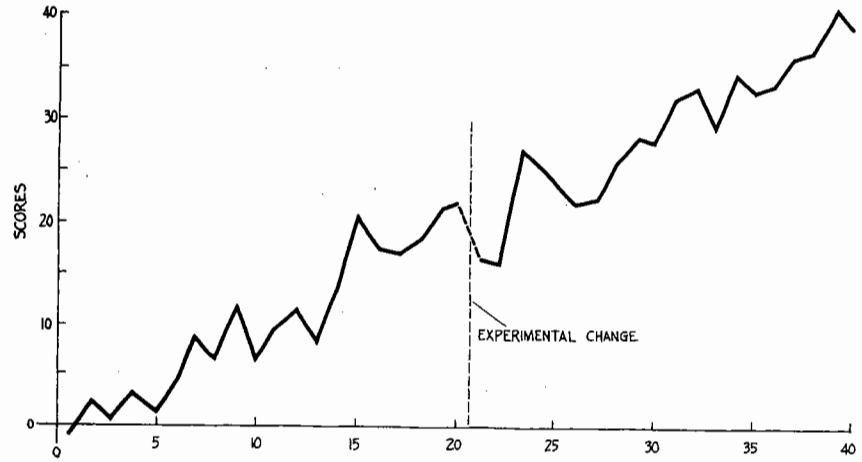

Figure 6 *Simulated time series with independent error. Source: Ross, Campbell, and Glass (1970)*

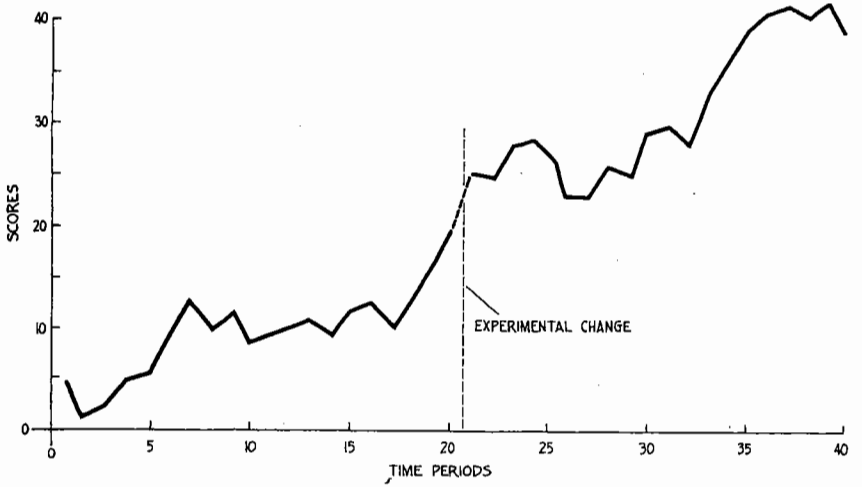

Figure 7 *Simulated time series with correlated (lagged) error. Source: Campbell (1969)*

sionally, by chance, random deviations will occur in such a pattern so as to make it look as though the treatment had an effect, as perhaps in Figure 6. It is the task of *tests of significance* to estimate when the difference from before treatment to after treatment is more than such random deviations could account for. Statistical formulas have been worked out that do this well in the case of independent deviations such as Figure 6 illustrates.

Figure 7 is based upon the same straight diagonal line as Figure 6. It has the same magnitude of deviations added. But the deviations are no longer independent. Instead, four smaller deviations have been added at each point, in a

staggered or lagged pattern. A new deviation is introduced at each time period and persists for three subsequent periods. As a result, each point shares three such deviations with the period immediately prior and three with the period immediately following. It shares two deviations with periods two steps away in either direction, and one deviation with periods three steps away. For periods four or more steps away, the deviations are independent. While Figure 7, like Figure 6, is a straight line distorted by random error, note how much more dynamic and cyclical it seems. Such nonindependent deviations mislead both visual judgments of effect and tests of significance that assume independence, through producing judgments of statistically significant effect much too frequently. (To emphasize the lack of any true or systematic departures, let it be emphasized that were one to repeat each simulation 1,000 times, and to average the results, each average would approximate the perfectly straight diagonal line of the true scores. That is, Figures 6 and 7 impose an underlying linearity, which will usually be inappropriate.) There are a variety of ways in which statisticians are attempting to get appropriate tests of significance for the real-life time series in which nonlinear general trends and nonindependent deviations are characteristic. (Probably most appropriate are the procedures of Box and Tiao, 1975; Cook and Campbell, 1979; McCleary and Hay, 1980.)

REGRESSION ARTIFACTS

We have moved from simple problems of inference to more complex ones. We will return soon to some more easily understood problems. But before doing that, let us attempt to understand a final difficult problem, known in one statistical tradition as *regression artifacts.* If we can be sure that the policy change took place independently of the ups and downs of the previous time periods, there is no worry. But if the timing of the policy change was chosen just because of an extreme value immediately prior, then a regression artifact will be sufficient to explain the occurrence of subsequent less extreme values. To see if a regression artifact might be at work in the Connecticut case, let us return to Figure 2. Here we can note that the most dramatic change in the whole series is the 1954–1955 increase. By studying the newspapers and the governor's pronouncements, we can tell that it was this striking increase that caused him to initiate the crackdown. Thus the treatment came when it did because of the 1955 high point.

In any unstable time series, after any point that is an extreme departure from the general trend, the subsequent points will on the average be nearer the general trend. Try this out on Figure 6. Move your eye from left to right, noting each point that is "the highest so far." For most of these, the next point is lower, or has *regressed* toward the general trend. Such regression subsequent to points selected for their extremity is an automatic feature of the very fact of instability and should not be given a causal interpretation. Applied to Figure 2, this means that even with no true effect from the crackdown at all, we would expect 1956 to be lower than the extreme of 1955.

OTHER REASONS FOR SHIFTS IN TIME SERIES

It is going to be important for administrators, legislators, the voting public, and other groups of nonstatisticians to be able to draw conclusions from time series data on important public programs. For this reason, two further points will be made that are less directly statistical. First note that there are many reasons for abrupt shifts in time series other than the introduction of a program change. One very deceptive reason is a shift in recordkeeping procedures. Such shifts are apt to be made at the same time as other policy changes. For example, a major change in Chicago's police system came in 1959 when Professor Orlando Wilson was brought in from the University of California to reform a corrupt police department. Figure 8 shows his apparent effect on thefts—a dramatic *increase.* This turns out to be due to his reform of the recordkeeping system, and the rise was anticipated for that reason.

In a real situation, unlike in an insulated laboratory, many other causes may be operating at the same time as the experimental policy change. Thus in Connecticut or in England, a drop in traffic casualties might have been due to especially dry weather, or fewer cars on the road, or to new safety devices, or to a multitude of other factors. If one had been able to design an experiment with the randomized control groups discussed at the beginning of this essay, such explanations would have been ruled out statistically. However, setting up such an experiment would have been impossible in these two situations. We must instead try to rule out these rival explanations of an effect in other ways. One useful approach is to look at newspaper records of rainfall, changes in traffic density, and other possible causes of the shift.

Another approach is to look for some control comparison that should show the effects of these other causes, if they are operating, but where the specific reform treatment was not applied. For Connecticut, the data from four nearby states are relevant, as shown in Figure 9. All of these states should have been affected by changes in weather, new safety features in cars, and so forth. While these data support the notion that the 1956 Connecticut fatalities would have been lower than 1955 even without the crackdown, the persisting decline throughout 1957, 1958, and 1959 is steepest in Connecticut and may well indicate a genuine effect of the prolonged crackdown. While visually we have little

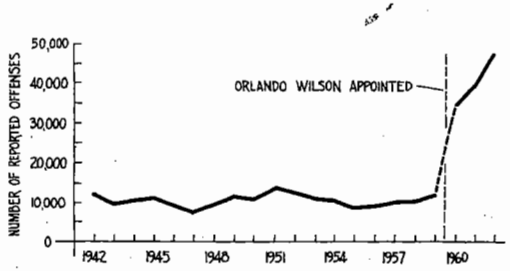

Figure 8 *Reported larcenies under $50 in Chicago from 1942–1962. Source: Campbell (1969). Data from Uniform Crime Reports for the United States, 1942–62.*

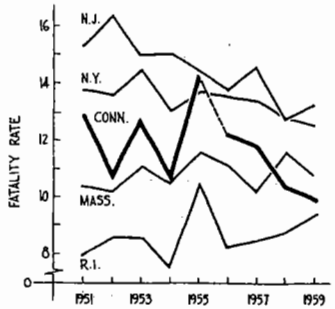

Figure 9 *Traffic fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons). Source: Campbell (1969)*

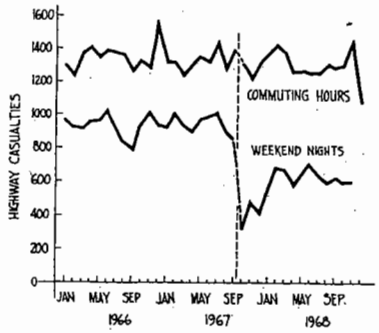

Figure 10 *A comparison of casualties during closed hours (commuting hours) and weekend nights in the English Breathalyser crackdown. Source: Ross, Campbell, and Glass (1970)*

difficulty in using these supplementary data, the statistician has many real problems in combining them all in an appropriate test of significance.

For England, there were no appropriate comparison nations available. But British pubs are closed before and during commuting hours, so casualties from such hours provide a kind of comparison base, as shown in Figure 10. Unfortunately there is a lot of instability in these data so they do not enable us to estimate with much confidence the degree to which the initial crackdown effects are persisting.

A FINAL NOTE

There are several general lessons to be learned from these brief illustrations. The first involves the distinction between *true experiments,* in which experimental and control groups are assigned by randomization, and *quasi-experiments.* True experiments, when they are possible, offer much greater power and precision of inference than do quasi-experiments. The administrators of social innovations, in consultation with statisticians, should attempt to use

such designs where possible. Where true experiments are not possible, or have not been used, there are some quasi-experimental designs, such as the interrupted time series, that can be very useful in evaluating policy changes. These too require statistical skill to avoid misleading conclusions. Evaluation of social innovations is an important and challenging area of application for modern statistics.

PROBLEMS

1. Why is the design on which this essay concentrates called an *interrupted* time series design?
2. How does Figure 2 change your perception of Figure 1?
3. Consider Figure 3. Suppose that X is a currency devaluation that results in a price increase between 0_4 and 0_5. What effect will X have on the price of Product A? Product C? Product F?
4. Consider Figure 3. Why does the author say that one can most legitimately infer an effect in A and B while one would be totally unjustified in inferring one in F, G, and H?
5. Consider Figures 4 and 5. Is the effect of the Breathalyser crackdown more pronounced in one of the figures? If so, which one?
6. a. Explain the difference between independent and lagged error.
   b. How would you characterize the effects of the two types of error on the plots of data in Figures 6 and 7?
7. What does the author mean when he says Figure 7 is more "dynamic and cyclical" than Figure 6?
8. Explain what the author means by the term *regression artifact.*
9. Refer to Figure 8. Explain the sharp increase in the number of reported offenses when Orlando Wilson was appointed.
10. Should the 1955–1956 traffic fatality decline in Rhode Island (Figure 9) be attributed to the speeding crackdown in neighboring Connecticut? Explain your answer.
11. What is a *true experiment*? What is a *quasi-experiment*?
12. Comment on this statement: Quasi-experiments are used in evaluating policy changes because it is virtually impossible to apply true experimental design to social situations.

REFERENCES

G. E. P. Box and G. C. Tiao. 1975. "Intervention Analysis with Applications to Economic and Environmental Problems." *Journal of the American Statistical Association* 70: 70–79.

D. T. Campbell. 1969. "Reforms as Experiments." *American Psychologist* 24(4): 409–429.

D. T. Campbell and H. L. Ross. 1968. "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis." *Law & Society Review* 3(1): 33–53.

D. T. Campbell and J. C. Stanley. 1963. "Experimental and Quasi-Experimental Designs for Research on Teaching." In *Handbook of Research on Teaching,* N. L. Gage, ed. Chicago: Rand-McNally, pp. 171–246. (Reprinted as *Experimental and Quasi-Experimental Designs for Research.* 1966. Chicago: Rand-McNally.)

T. D. Cook and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand-McNally.

D. Kershaw and J. Fair, eds. 1976. "Operations, surveys and administration," Vol. 1: *The New Jersey Income-Maintenance Experiment.* New York: Academic Press.

R. McCleary and R. A. Hay, Jr. 1980. *Applied Time Series Analysis for the Social Sciences.* Beverly Hills: Sage.

H. L. Ross, D. T. Campbell, and G. V. Glass. 1970. "Determining the Social Effects of a Legal Reform: The British 'Breathalyser' Crackdown of 1967." *American Behavioral Scientist* 15(1): 110–113.

H. W. Watts and A. Rees, eds. 1977. "Expenditures, health, and social behavior; and the quality of the evidence," Vol. 2 and 3: *The New Jersey Income-Maintenance Experiment.* New York: Academic Press.