

Interrupted Time Series Quasi-Experiments¹

Gene V Glass
Arizona State University

Researchers seek to establish causal relationships by conducting experiments. The standard for causal proof is what Campbell and Stanly (1963) called the "true experiment." Often, circumstances will not permit meeting all the conditions of a true experiment. Then, a quasi-experiment is chosen. Among the various quasi-experimental designs is one that rivals the true experiment: the interrupted time-series design. It has become the standard method of causal analysis in applied behavioral research.

Just what is a "cause" is a matter of deep philosophical debate. Perhaps I can safely ignore that debate and appeal to your intuitive understanding that renders meaningful such statements as "The nail caused the tire to go flat" or "Owning a car causes teenagers' grades to drop." If every relationship were causal, the world would be a simple place; but most relationships are not. In schools where teachers make above-average salaries, pupils score above average on achievement tests. It is not safe, however, to say that increasing teachers' salaries will cause an increase in pupils' achievement. Business executives who take long, expensive vacations make higher salaries than executives who don't. But will taking the summer off and touring Europe increase your salary? Try it and find out.

Relationships: Causal and Spurious

Relationships can fail to be causal relationships in two principal ways: because of (a) third variables and (b) an ambiguous direction of influence. The third-variable situation occurs when two things are related because each is causally related to a third variable, not because of any causal link between each other. The teachers' salaries and pupil achievement example is probably an instance of the third-variable situation. In this case, the third variable might be the wealth of the community; rich communities pay teachers more and have pupils who score higher on achievement tests for a host of reasons connected to family wealth but not to teachers' pay. Teachers are professionals who want to be paid well and deserve to be; but I doubt that, once the negotiations are finished, a teacher tries any harder to teach the pupils because of a few hundred dollars on the salary schedule. So the relationship of teachers' salaries and pupil achievement—a relationship that is an empirical fact, incidentally—is due to common relationships to a third variable.

The business executive's vacation is an example of ambiguous direction of influence. A travel agent might publish a graph in an advertisement that shows this relationship. However, the simple fact of the relationship leaves quite ambiguous whether long vacations cause higher salaries (presumably through improving morale and vitality and the like) or higher salaries cause long, expensive vacations. The truth is obvious in this case, and it is quite the opposite of the direction of influence that the travel agents wants people to believe. But many other examples are less clear. Does enhanced motivation cause pupils to learn successfully in school, or is it mainly the other way around: success in learning causes an increase in motivation to learn? The truth is probably some of each in unknown amounts, which goes to show how ill-

¹ Reprinted from Jaeger, R. M. (1997). *Complementary methods for research in education*. 2nd Edition. Pp. 589-608. Washington D. C.: American Educational Research Association.

advised one is to think of each relationship as if it were a causal relationship. Experimenters name relationships that would not stand the test of a true experiment "spurious."

Experimenters have devised a methodology that lays both of these problems to rest. They contrive two or more sets of circumstances that are alike in all respects except for the phenomenon that is being tested as a possible cause, and then they subsequently observe whether the expected effect ensues. For example, an experimenter might take a large sample of teachers and their pupils and divide them into two identical groups except that one group's teachers receive a \$1,500.00 raise and the other group's do not. Then a year later he or she measures the pupils' achievement to see whether it has been affected. By setting up two identical groups of teachers and their pupils, the experimenter ruled out all possible third variables as explanations of the eventual difference in pupil achievement. Can it be said that an achievement advantage for the pupils of the better paid teachers is not really due to increased pay since the better paid teachers might have had older pupils or smarter pupils or the better paid teachers might have had more experience? In other words, might not there be some third-variable problems here? No, because the teachers and pupils were equivalent in all respects at the beginning of the year. Can it be said that the direction of influence between salaries and achievement is ambiguous in this experiment? No, because the different salaries were set by the experimenters before pupil achievement was observed. Hence, the differences in achievement could not have caused the differences in teacher salaries; the only possible direction of influence is the other way around.

This style of experimental thinking has been around for well over 150 years. In its original conception (due primarily to J. S. Mill) it was relatively impractical because it held that the conditions compared in the experiment had to be identical in all respects except for the hypothesized cause; that is, all third variables were to be ruled out by ensuring that they did not vary between the conditions. But all the possible third variables can not be known and even if they could be, they couldn't possibly be equated. Imagine having to equate the high-paid and low-paid teachers on their age, sex, height, weight, IQ, experience, nationality and on and on.

Randomized or "True" Experiments

The experimental method received a big boost in the 1920s when a young Englishman named Ronald Fisher devised an ingenious, practical solution to the third-variable problem. Fisher reasoned that if, for example, chance alone was used to determine which teachers in the experiment were paid more and which less, then any of an infinite number of possible third variables would be equated between the two groups, not numerically equal, but equated within the limits of chance, or randomly equated as it has come to be known. If a coin flip determines which teachers enter the high-pay group and which the low-pay group, then with respect to any third variable you can imagine (eye color, shoe size, or whatever) the two groups will differ only by chance. Fisher then reasoned as follows: if, after the experiment, the only observed differences between the conditions are no larger than what chance might account for, then those differences might well be due to the chance differences on some third variables. But if the differences are much larger than what chance might produce (e.g., if all of the pupils of well-paid teachers learn much more than the pupils of poorly paid teachers), then chance differences in third variables could not account for this result (differences in teacher pay must be the cause of the large differences in pupil achievement). Because experimenters must calculate the size of differences that

chance is likely to produce and compare them with the differences they actually observe, they necessarily become involved with probability theory and its application to statistics.

Fisher's modern experimental methods were applied in agricultural research for 20 years or so before they began to be applied in psychology and eventually in education. In the early 1960s, a psychologist, Donald Campbell, and an educational researcher, Julian Stanley (Campbell & Stanley, 1963), published a paper that was quickly acknowledged to be a classic. They drew important distinctions between experiments of the type Fisher devised and many other designs and methods being employed by researchers with aspirations to experiments but failing to satisfy all of Fisher's conditions. Campbell and Stanley called the experiments that Fisher devised "true experiments." The methods that fell short of satisfying the conditions of true experiments they called "quasi-experiments," quasi meaning seemingly or apparently but not genuinely so.

True experiments satisfy three conditions: the experimenter sets up two or more conditions whose effects are to be evaluated subsequently; persons or groups of persons are then assigned strictly at random, that is, by chance, to the conditions; the eventual differences between the conditions on the measure of effect (for example, the pupils' achievement) are compared with differences of chance or random magnitude.

The most common substitute for a true experiment is what Campbell and Stanley labeled a "pre-experimental design." In the most common example, two groups—a treatment group and a control group—are compared on a dependent variable "after the fact" of their having found their way into these two groups through no intervention of the experimenter. This "ex post facto" experiment usually involves an attempt to correct statistically for differences between the experimental and control groups on a succession of "third variables"—through such devices as analysis of covariance. For example, a group of students who report watching over 100 hours of television per month is compared with a group that watches fewer than 20 hours; the dependent variable is success in school. The researcher might try to equate the groups after the experiment has run its course by adjusting school success averages for differences in intelligence and social-economic status. The adjusted differences in this pre-experimental design were long thought to be dubious, but never more so than after Cronbach (1982) explained the many and varied deficiencies of this method of testing causal claims. Nothing so recommends the techniques of quasi-experimentation presented below as does Cronbach's critique of the fatally flawed ex-post facto pre-experiment.

A Quasi-Experimental Design: The Interrupted Time Series

Quasi-experiments cannot be so easily described. There are many varieties of them, and the various species arise from different ways of attempting to control for third variables without actually using random assignment. One of the most promising quasi-experimental designs is known as the interrupted time-series experiment.

In Figure 1 appears a graph of the number of depressed thoughts (the solid black line) recorded for 49 consecutive days by a mildly depressive young woman. On the first day, she counted about 70 depressed thoughts; on the second day, about 55. After 14 days of recording (after the first solid vertical line in Figure 1) the woman was placed on a behavioral modification therapy designed to eliminate depression. The therapy was continued for 21 days and then terminated (at the second vertical line), and the recording of depressed thoughts continued for an additional 14 posttreatment days.

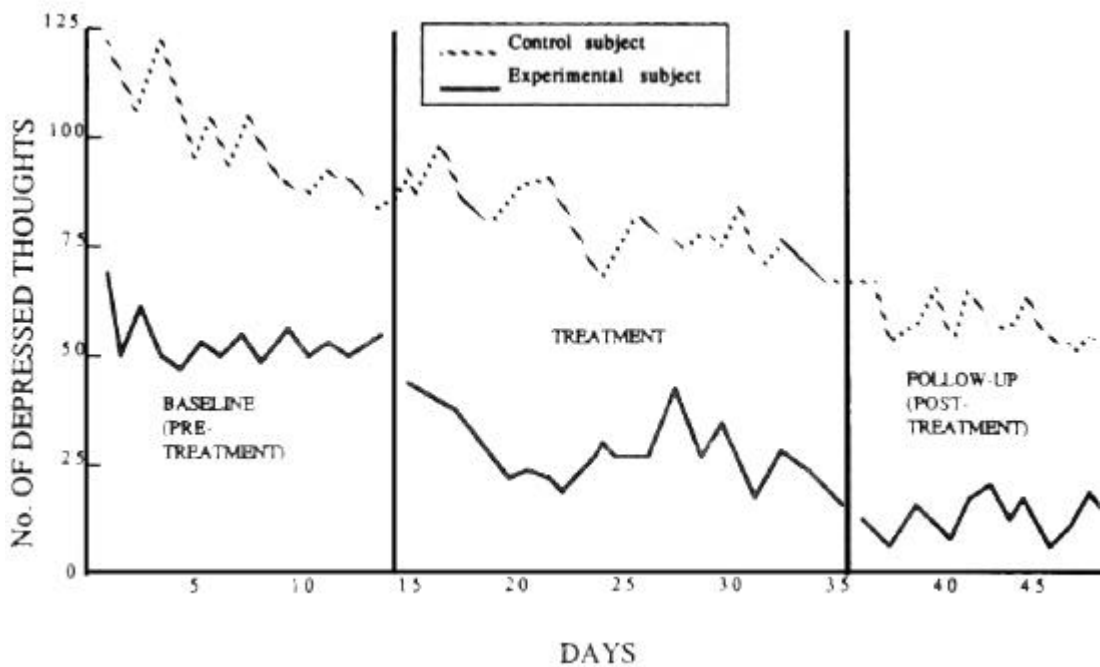


Figure 1. Incidence of depressed thoughts for an experimental subject receiving behavioral therapy and for a control subject.

By studying the pattern of the graph, paying particular attention to what happens to the series precisely at the point of shifting from no-treatment to treatment or vice versa, one hopes to learn something about the effect of the therapy on depressed thoughts. In general terms, the therapy is called a treatment or an intervention and depressed thoughts are the outcome or dependent variable.

You might often find that a time-series experiment is a workable alternative when the conditions of a true experiment can not be met; for example, when you have only one or two units (persons or classrooms) instead of a substantial number, or when different units can not be treated in different ways (methods A and B) at the same time. But the time-series experiment imposes its own requirements, the foremost of which is that it sometimes requires that data be recorded for many consecutive points in time before and after a treatment is introduced. How many is "many"? We'll return to this important question later.

Another set of circumstances in which one might use a time-series experiment to assess the effect of a treatment is with data archives and naturally occurring interventions. In this case, the intervention is made by someone other than the researcher and it is not normally made for experimental purposes, although the researcher makes use of it for causal analysis. The data to evaluate the impact of the intervention come from archives: collections of data gathered routinely across time for administrative purposes. An example of this type of archival time-series experiment appears as Figure 2. Here you see a graph of the traffic fatality rate in Great Britain, by month, from 1961 through 1970. The dependent variable is the number of traffic deaths in a month divided by the number of miles driven in Great Britain, in units of 100,000,000. For example, there were about 12 deaths per 100,000,000 miles driven in January 1961; that is, if the total mileage driven in Great Britain in January 1961 was 300,000,000, then there were about $3 \times 12 = 36$ traffic deaths that month. In October 1967, the British Road Safety Act of 1967 instituted a variety of measures designed to reduce traffic accidents, foremost among which was setting up road blocks and

administering breath tests to discover drunken drivers. This intervention is indicated by the vertical line in Figure 2.

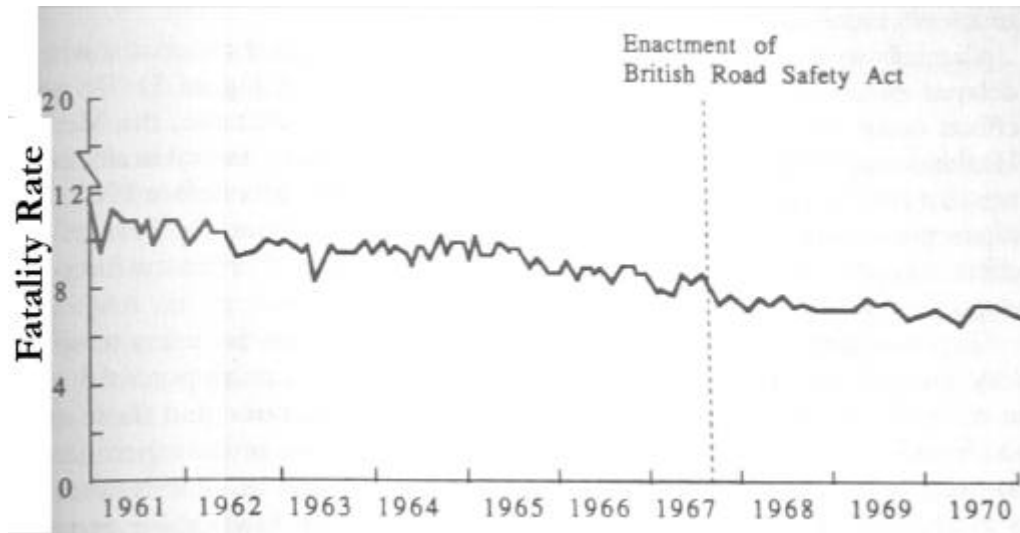


Figure 2. British fatality rate corrected for miles driven and with seasonal variations removed. (Source: Ross, Campbell & Glass, 1970).

The simple logic of the time-series experiment is this: if the graph of the dependent variable shows an abrupt shift in level or direction precisely at the point of intervention, then the intervention is a cause of the effect on the dependent variable. In the two examples presented thus far, the particular causal questions are these: "Did the behavioral psychotherapy reduce the rate of depressed thoughts?" and "Did the British Road Safety Act of 1967 reduce the traffic fatality rate?"

The effect of an intervention on an outcome variable can assume a variety of forms. Some of these are depicted in Figure 3. Having detailed expectations about how the graph should change in response to an intervention is the best protection against deluding oneself and being led astray by extraneous influences and chance occurrences. For example, suppose that a highway safety law required a car tire safety check and replacement of worn tires. Because not all tires would have worn treads at the moment the law was enacted, and because perhaps 10% of the worn tires might be detected and replaced monthly for about the first year, an intervention effect like that in case E in Figure 3 might be anticipated. If an abrupt level change—case A—were observed in the data instead, the chain of causal reasoning from tire inspection to reduced accidents would be weakened, even though the possibility of an effect due to some unknown cause remained strong.

Carefully spelled out expectations for the form of an effect are crucial where delayed effects might exist (for example, cases B and F in Figure 3). Delayed effects occur frequently and are perfectly plausible. For example, the Mental Health Act of 1965 authorized construction of community mental health centers that couldn't possibly have begun to function significantly before 1967. An experimenter who does not have definite expectations about the form of an effect may interpret as an effect every little ephemeral and irrelevant jump or wiggle in the curve.

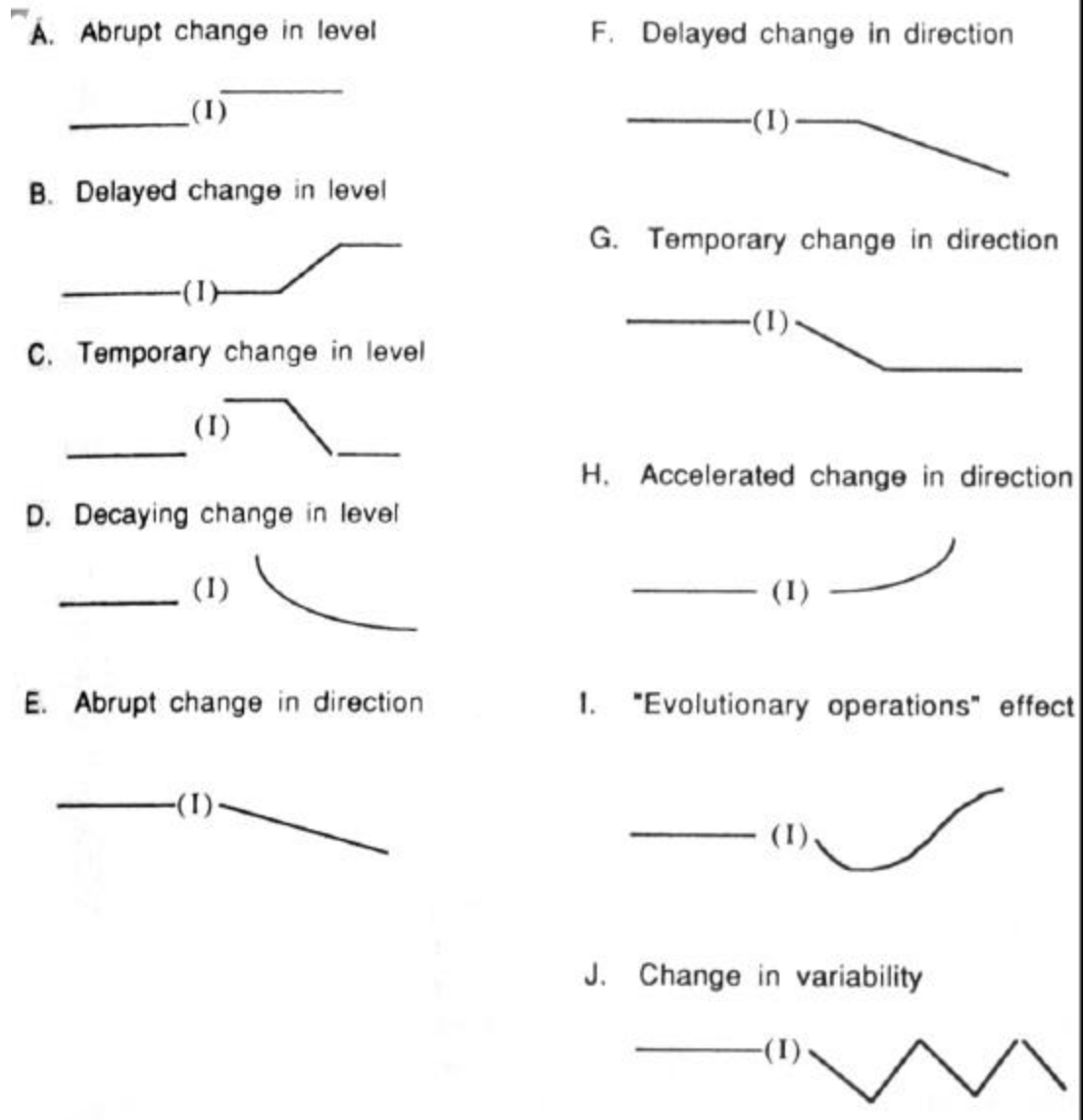


Figure 3. Varieties of intervention effects in the time-series experiment.
 [(I) denotes the time at which an intervention is introduced.]

Graphs of time-series data can change level and direction for many reasons, some related to the intervention and others not. Separating potential reasons for effects into those essentially related to the intervention and those only accidentally related is the principal task in analyzing time-series experiments. You might recognize the time-series experimental design as an extension in both directions (the past and the future) of the old and disreputable pretest-posttest experiment. In the pretest-posttest design, a group of persons is observed, a treatment is applied to them, and then they are observed a second time. An increase (or a decrease) of the group average from the first observation to the second is taken to be the effect of the independent variable (the treatment) on the dependent variable that was observed. The pretest-posttest design has such a bad reputation because so many influences other than the treatment can account for a change in scores from pre to post. The person or persons in the experiment could grow tired, smarter, or less cooperative from the single pre-measure to the postmeasure. Likewise, two less-than-identical tests might be given before and after,

or the experimental units could have been chosen because they fell so far from the average on the pretest thus greatly increasing the odds of their regression (i.e., movement) toward the mean of the posttest. Each of these influences can cause increases or decreases in scores that are confounded with whatever effects the intervention or treatment might have. The pretest-posttest design deserves its bad press.

The time-series experiment adds many preintervention and postintervention observations and thus permits separating real intervention effects from other long-term trends in a time series. Consider Figure 1 again. The series of depressed thoughts for the control subject shows a gradual, regular decline from Day 1 to Day 49. The course that the graph follows shows no abrupt shifts at those points in time when the experimental subject, represented by the solid black line, is put on treatment and taken off. There is no evidence of an intervention effect on the control subject's graph, as there ought not to be, since this subject was given no therapy but merely recorded the rate of depressed thoughts for 49 days. But notice how, if a single premeasure had been taken on the control subject around Day 5 and a single postmeasure around Day 25, it would have appeared as though this subject had experienced a sizable reduction in depressed thoughts due to treatment. To be sure the control subject's graph shows a decline across time, but the decline could be due to many extraneous factors, including becoming bored with counting depressed thoughts, a shifting standard for what constitutes a depressed thought, hormonal changes and the like.

There are many different ways to arrange experimental units and to plan the intervention of one or several interventions. Some ways may be cheaper or safer than others. In Figure 4, several possible alternative time-series designs are sketched. They are special purpose designs, each created for reasons of economy or validity. The Os in the diagrams stand for observations or measurements; they are taken repeatedly over time. The Is stand for interventions, that is, the introduction of a treatment. The "1" and the "2" subscripts are added to the Is to indicate that two different treatments were introduced into the series.

Time-series experiments differ in important ways depending on whether the "experimental unit" is a single person or a group of persons. If the unit is a large group of persons, one can either measure each person in the group at each point in time or subsample different persons at each point. For example, in Figure 5 appear the results of a time-series experiment on the effects of Outward Bound—a wilderness camping experience—on participants' self-confidence. Several hundred participants were available for study, but it would have been unnecessary and too costly to test each person each week of the 11 month study. By randomly sampling a new group of 20 persons each week, no participant was burdened by the testing and yet an accurate picture was obtained of the group's response to treatment.

Statistical Analysis of the Time-Series Experiment

Although operant psychologists made extensive use of the time-series design for years before methodologists christened it as such, they have remained resistant to most considerations of statistical analysis, preferring to rely on visual inspection of the graphs to judge the presence of an effect. Even though the methods of analyzing time-series experiments have been in the literature for nearly three decades, those who could benefit most from their application refuse to employ them. Indeed, one can scan volumes of applied operant literature (e.g., the *Journal of Applied Behavior Analysis*) and never encounter statistics more complex than means and graphs. And yet, the statistical problems of interpreting time-series experiments are sufficiently complex that merely eye-balling graphs and drawing conclusions is not enough.

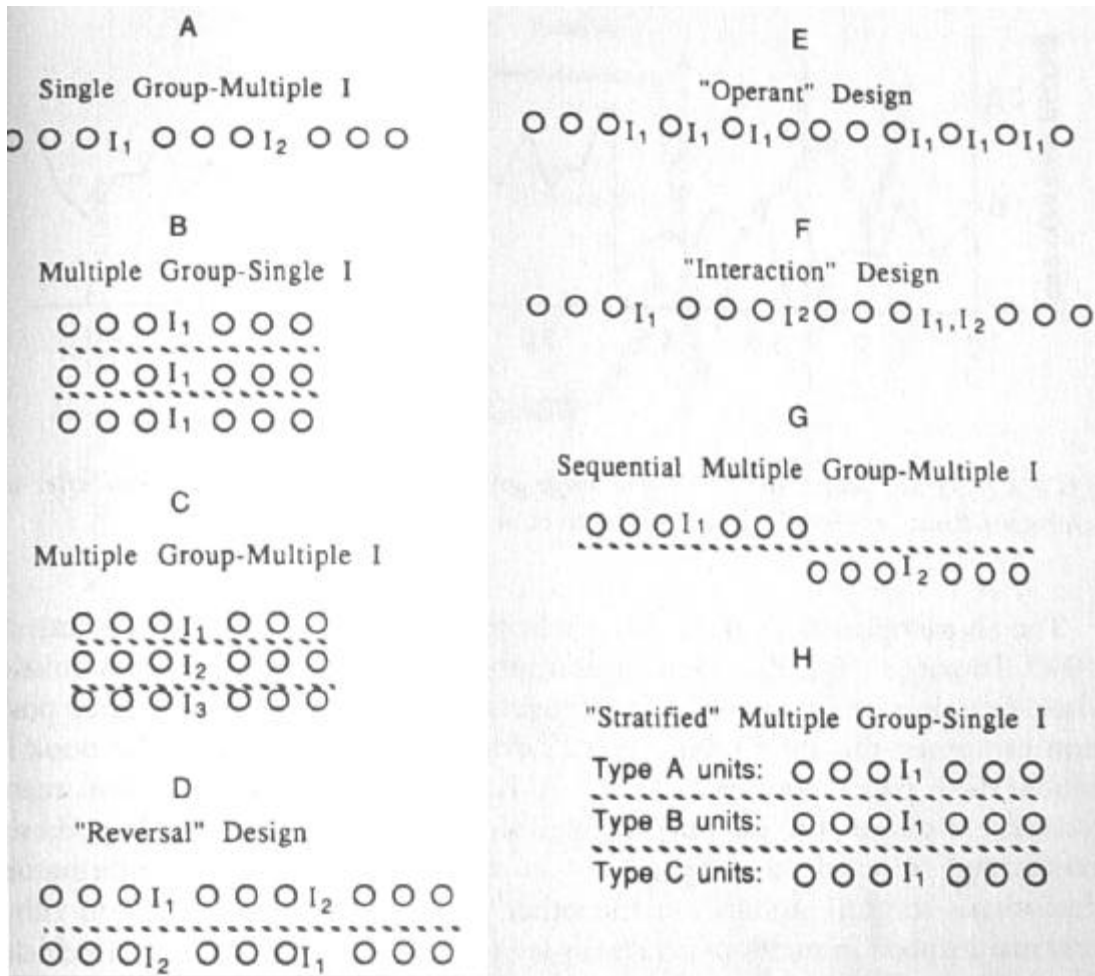


Figure 4. Variations on the basic time-series experimental design. (O represents an observation or measurement of the dependent variable; I represents an intervention.)

The operant psychologists' position on the application of statistics to time-series experiments was possibly never stated more emphatically and directly than it once was by an anonymous reviewer of a manuscript submitted to an applied behavioral analysis journal 25 years ago:

Applied behavior analysis doesn't need sensitive statistics to salvage small effects. The field needs basic effects, that is, effects that are large, reliable, powerful, clear, and durable. Such basic effects will give us strong principles. The field doesn't need dubious, small, and sometime but statistically significant results, because these would result in the development of more and weaker principles. Behavior modification became a technology by ignoring small effects. The sort of eyeball analysis that has been done has acted as a filter, weeding out the small effects. ... In the reviewer's opinion, the development of sensitive statistics appropriate for the research designs of applied behavior analysis would be bad for the field.

The above opinion is not merely an ancient relic from a discarded past. As recently as 1992, Parsonson and Baer were arguing for the superiority of impressionistic, visual examination of time-series experiments over statistical analysis. Their position prompted this mild rebuke from Levin (1992, p. 221), who edited the

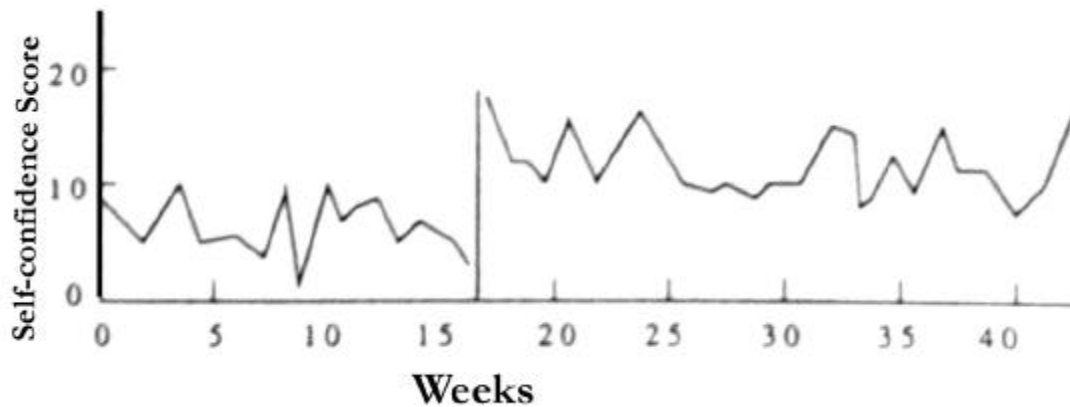


Figure 5. Self-confidence scores on a large group of participants before and after an Outward Bound experience. (Source: Smith *et al.*, 1976)

book in which their paper appeared: "... it is the author's opinion that many researchers cannot (or do not) distinguish between exploratory (hypothesis-generating) research investigations, on the one hand and confirmatory (hypothesis-testing) studies, on the other. Whereas visual analysis and other informal inspection methods represent legitimate exploratory-research vehicles for communicating what 'is interesting to the eye' or what 'merits further study,' corroborating predicted outcomes on the basis of formal statistical analysis is often mandatory for confirmatory-research studies."

Of course, everyone would like "large, reliable, powerful, clear and durable effects." But perusal of the *Journal of Applied Behavior Analysis* will show that its authors are more than willing to investigate and report small, weak and ephemeral effects, or effects that are entirely illusory. Such perusal will also reveal that the "eyeball" test is applied very differently by different persons—some seeing effects, some seeing none at all. Consider the study published by Gettinger (1993) on the effects of two methods of spelling instruction in a regular second grade classroom. Four boys, two above average spellers and two below average spellers, were alternately taught to spell a targeted set of words by either the Direct Instruction method or the Invented Spelling method (wherein words are spelled phonetically at first in context then eventually replaced with conventional orthography). The different methods were applied in a multiple-group ABAB design. The results of the study appear as Figure 6, in which the series for the two above average spellers have been averaged and likewise for the two below average spellers. Gettinger (1993, p. 287) wrote this about the data on which Figure 6 is based: "As predicted, children spelled more of the six words correctly when they received direct instruction and practice on words in isolation than during the invented spelling condition." I know of no legitimate statistical analysis that would support the conclusion that the ABAB design in Figure 6 shows a differential effect for Direct Instruction versus Invented Spelling. Judge for yourself.

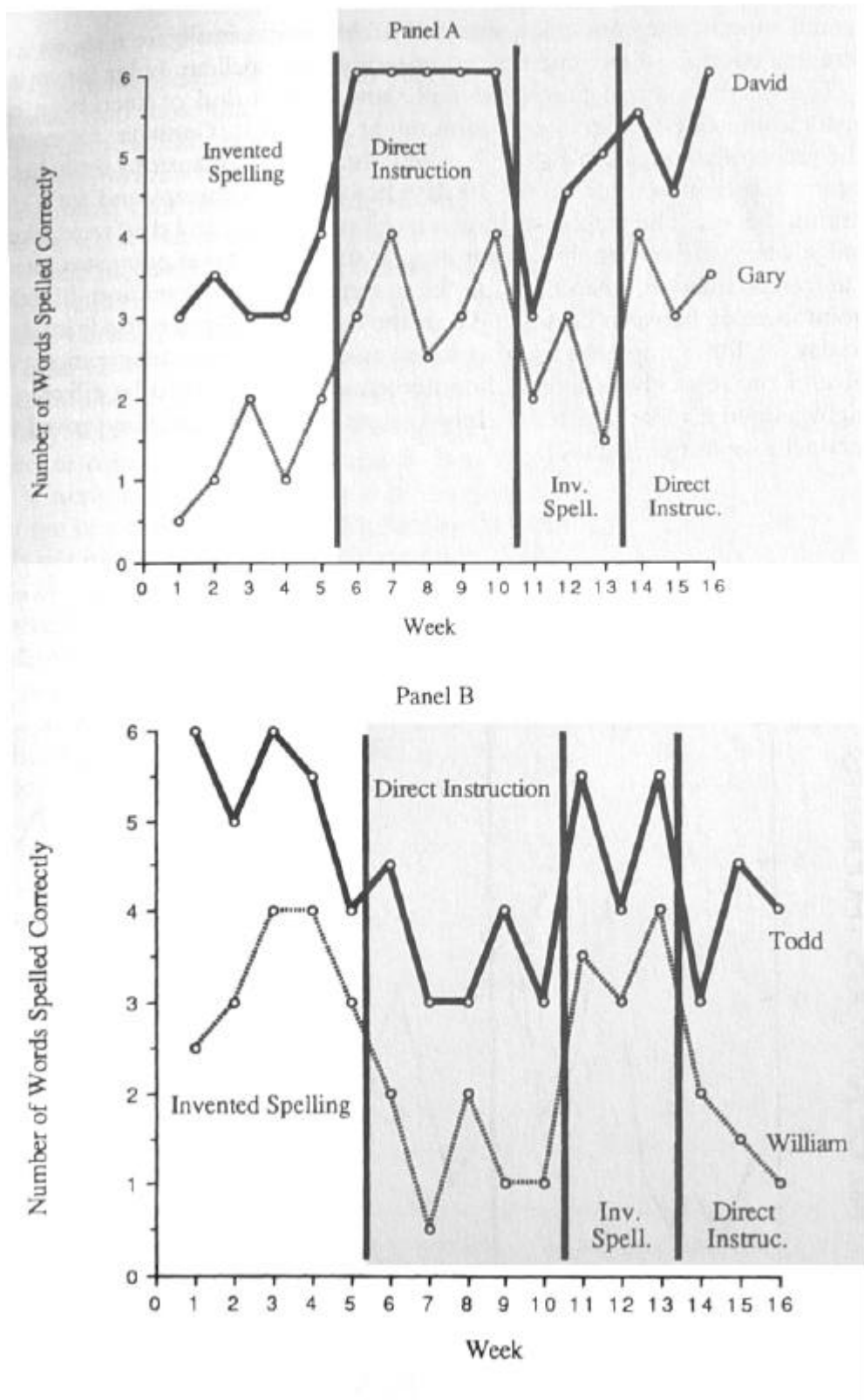


Figure 6. Number of words spelled correctly by four boys (two good spellers—Panel A; two poor spellers—Panel B) under two methods of spelling instruction (after Gettinger, 1993).

The graph of a real time-series will show a good deal of fluctuation quite apart from any effect an intervention might have on it. Consider, for example, the graph of the series in Figure 7, where the number of anxious thoughts of a neurotic patient is recorded for 10 days before psychotherapy and for 21 days during therapy. The graph in Figure 7 was showed to 13 researchers who were asked to judge whether the roughly 7-point drop at the intervention point was merely a random fluctuation, unaccountable for in terms of the intervention (like the 9-point increase between days 5 and 6 or the more than 9-point rise from day 29 to day 30, for example) or whether it was evidence of a true intervention effect. Five of the researchers judged the intervention of therapy to be effective and eight judged it to be ineffective. Analysis by visual inspection and by jury proved to be inconclusive in this instance.

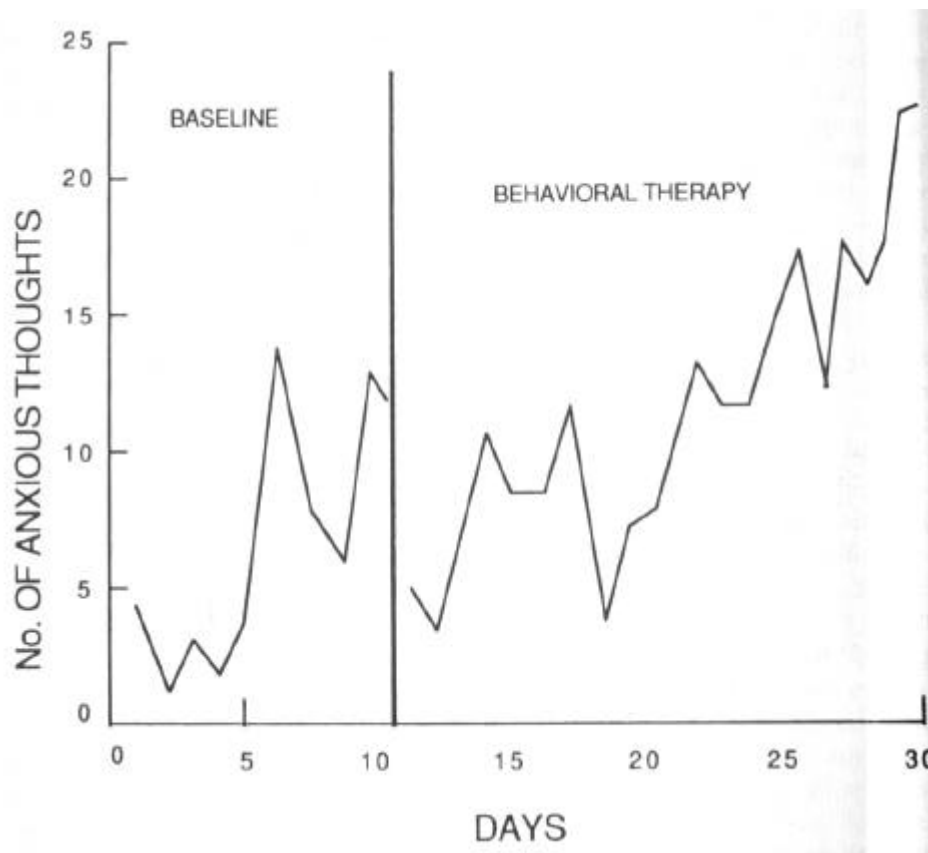


Figure 7. Incidence of anxious thoughts for a single subject before and during behavioral therapy.

Stationary vs Non-stationary Series.

A key distinction has to be drawn between two types of series: those that are stationary and those that are nonstationary. The distinction makes all the difference in the world when it comes to inspecting graphs and analyzing whether or not an intervention had an effect. The difference between stationary and nonstationary time series is the difference between graphs that fluctuate around some fixed levels across time and graphs that wander about, changing their level and region haphazardly. The graph of British traffic fatalities in Figure 2 is reasonably stationary, with most of the preintervention observations contained in a region between 8 and 11.

It is far easier to detect an intervention effect in a stationary process than in a nonstationary one. Intervention effects are seen as displacements of the curve, and as you can see, nonstationary processes are, by definition, curves that are subject to

displacements of level and slope at random points in time. The trick with nonstationary time series is to distinguish the random displacements from the one true deterministic displacement caused by the intervention. And that's not easy.

If a process is basically nonstationary, one needs to establish that fact and, furthermore, be able to estimate the degree of nonstationarity before you can decide how much displacement of the curve precisely at the point of intervention is convincing evidence of an effect. Behavioral psychologists have long (since Skinner) recognized this fact in their advice to establish a stable baseline for the organism before intervening; unfortunately, as one moves out of the laboratory into the world, one often encounters baselines that are not so accommodating and would never settle into stationary no matter how long they are observed. Nonstationarity is a fact of life in the world. It must be dealt with, not wished away.

How can one establish the fact and character of nonstationarity? By patiently watching the series for a long time before intervening. This is why the researchers who looked at Figure 7 and tried to guess whether the intervention was really effective or not, did such a bad job. The preintervention period, the "baseline" if you will, simply wasn't long enough to establish the character of the process; i.e., it wasn't long enough to give one a good idea of how the series should behave in the absence of any intervention. The circumstances in Figure 2 are quite different. The British fatality rate can be expected to be about 8 to 10 after mid-1967 as it was for several months before.

Once again, the statistician must act as a messenger of the doleful news: the single most important thing to remember about time-series experiments is that they require long baseline periods for establishing whether the process is stationary or nonstationary.

How long is "long"? Once again, the statistician must, in honesty, give a vague answer where a precise one is expected. It is difficult to say exactly how long a baseline must be for a time-series experiment to be safe. The answer depends in part on how much one knows in advance about the process being studied. If a researcher has studied extensively a particular phenomenon—like out-of-seat time" for pupils in a classroom—and has never seen much nonstationarity, it's conceivable that a satisfactory baseline can be established in relatively short order in a time-series experiment, a dozen or so points in time, say. But when a series is quite nonstationary, 40 or 50 baseline time points may be needed before the process is sufficiently understood that one would attempt to assess the effects of an intervention. We have found that time-series of observations of a single individual are often stationary, particularly if you discard the first few points during which the subject may be acclimating to the equipment or observation procedures. On the other hand, time-series based on large groups of persons—like classrooms, cities, or the population of an entire nation—are often nonstationary. I hasten to point out that this generalization isn't always true even in the examples presented here.

Few statisticians would insist on more than 100 points in time, but some of the more rigid ones might not accept fewer. Fifty time points (25 pre and 25 post) is a good round number, provided you make a definite hypothesis about the form of the intervention effect and stick with it; that is, provided you don't succumb to the temptation of fishing around in the data after you see it and taking second and third guesses at the form of the intervention effect. Fifteen preintervention points is certainly cutting things as thin, and even then so short a baseline is only safe if you know a lot in advance about the probable statistical character of the series. If you're held to under 10 preintervention points, then you shouldn't be performing a time-series experiment unless you have an intervention so potent that it will work huge effects, in which case you'll probably know it without an experiment. If you attempt time-series experiments

with too short baselines you're likely to end up with data like those in Figure 8. There the incidence of child molestation is graphed for a 12-year period around the enactment of three regulations that made pornography easily available to adults. The time series is simply too short to permit any kind of conclusion.



Figure 8. Child molestation (offenses against girls) in the city of Copenhagen. (Broken vertical lines represent successive liberalizations of Danish pornography laws.) (Source: Kutchinsky, 1975)

Time series nearly always evidence a form of statistical correlation that is quite unlike what we are accustomed to in working with random samples. For many reasons, today's score is more highly correlated with either yesterday's or tomorrow's score than it is with either the score from a week ago or a week hence. This condition of correlation within the series wreaks havoc with ordinary statistical techniques; it's about the only violation of standard statistical assumptions (normality, equal variance, and independence) that is worth worrying about. The proper means are known of taking account of this form of correlation, but the methods are complex by elementary textbook standards. At the end of this paper, I've given references in which you can pursue these more technical questions on your own. Suffice it to say that proper statistical analysis involves very complex methods now relegated to computer programs. Among the best of these for analyzing time-series experiments is a series of computer programs produced by Quantitative Micro Software (Quantitative Micro Software, 1994). For an explanation of the state-of-the-art analysis methods on which such programs as the Micro TSP program is based, see McCleary and Welsh (1992).

Investigation and Detective Work

In some time-series experiments, it frequently happens that subsections of a large data pool reveal intervention effects that are not apparent in the complete body of data. Consider the data in Figure 2, for example. If there is any effect of the British Road Safety Act of 1967 on the traffic fatality rate, it certainly isn't very apparent in the graph in Figure 2. But look at the graph in Figure 9 to see what happened when the fatalities on weekend nights were singled out of the total body of data. There we see a huge, more than 50%, reduction in fatalities coincident with the implementation of the Road Safety Act in October 1967. We can move from the equivocal results of Figure 2 to the clear certainty of Figure 9, merely by separating the larger body of data.

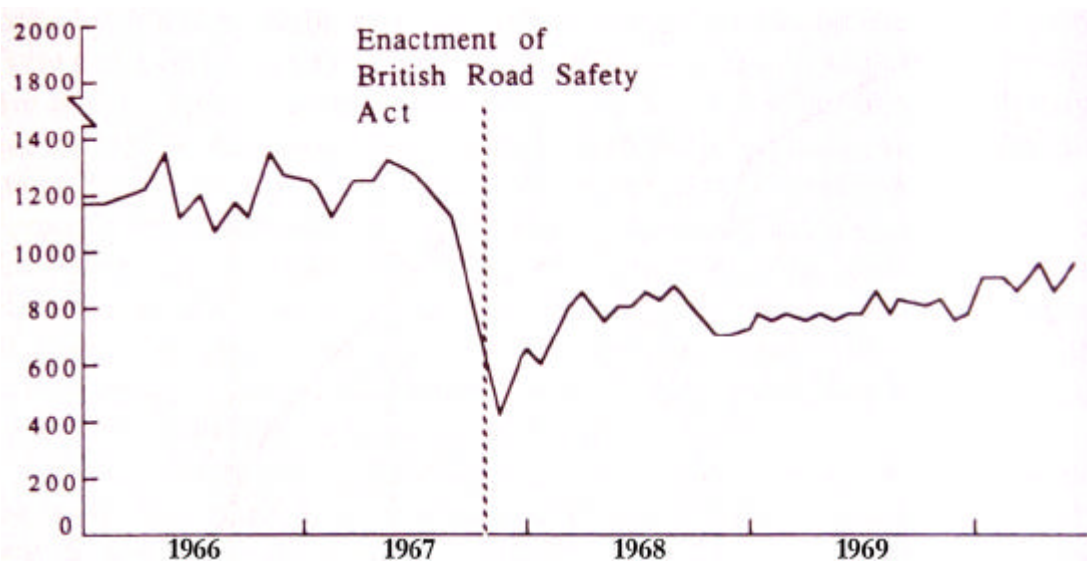


Figure 9. Fatalities for Friday nights, 10 p.m. to midnight; Saturday mornings, midnight to 4 a.m.; Saturday nights, 10 p.m. to midnight; and Sunday mornings, midnight to 4 a.m., corrected for weekend days per month, seasonal variations removed. Broken vertical line represents implementation of the British Road Safety Act. (Source: Ross, Campbell & Glass, 1970).

Why is the effect so apparent in Figure 9 when it was barely discernible (or indeed, not discernible at all) in Figure 2? As was mentioned earlier, an essential feature of the Road Safety Act was a program of roadblocks where drivers were tested for blood alcohol level. And what time is better for finding drunks on the road than weekend nights? The picture is completed in Figure 10 when one inspects the fatalities curve for the hours commuting to and from work in the morning and late afternoon when there are few drunken drivers. Sorting the data in Figure 2 into two different series in Figures 9 and 10 not only has revealed the intervention effect, but has illuminated the whole question of how the Road Safety Act worked its effect.



Figure 10. Fatalities for Mondays through Fridays, 8 a.m. to 10 a.m. and 4 p.m. to 5 p.m. corrected for weekdays per month, seasonal variation removed. Broken vertical line represents implementation of the British Road Safety Act. (Source: Ross, Campbell & Glass, 1970).

Opponents of the anti-drunk-driver features of the Road Safety Act argued that other features of the law were actually responsible for the decrease in fatalities: for example, the construction of more and better traffic signals, a tire inspection program, and a reduction in the numbers of motorcycles and mopeds on the road. The pattern of the time series in Figures 9 and 10 refutes these counterclaims, as each of the counterclaims would lead one to expect an effect that was initially small then grew, and was of equal size for commuting hours and weekend nights.

Consider Figure 11 as a final example where the answers aren't so clear. The figure is a graph of the enrollment of the Denver Public Schools from 1928 to 1975, and the experimental question concerns forced racial integration and white flight." In 1969, in the Federal District court, Judge Doyle rendered a decision in the *Keys vs. Denver Public Schools* case that represented the first forced integration decision in a major northern city. The question is, "Was there a resulting flight of white families to the suburbs to escape racially integrated schools?"

The total enrollment of the Denver Public Schools dropped from 1970 on, but it had been dropping ever since 1960. The question is whether the enrollment dropped faster after 1969 than before, and it appears to, but the situation is hardly unequivocal. The enrollment grew as fast during the 1950s as it fell in the early 1970s, and no one implicated racial strife or harmony in the earlier acceleration. Moreover, the national birthrate was falling sharply in the late 1960s, exactly 5 years ahead of what looks like a sharp decline in Denver School enrollment in the 1970s; and for all that Judge Doyle's decision might have done, it is stretching things to believe that it could have affected the birthrate 5 years before it was handed down.

The Denver case in Figure 11 has to remain uncertain. But I suspect that it could be resolved fairly conclusively by breaking down and plotting in several alternative ways the total enrollment series in Figure 11. Breaking the enrollment data down by grade might cast a little light on things. If it's really white flight that is causing the decline, one might expect a larger decline at the elementary grades than at the

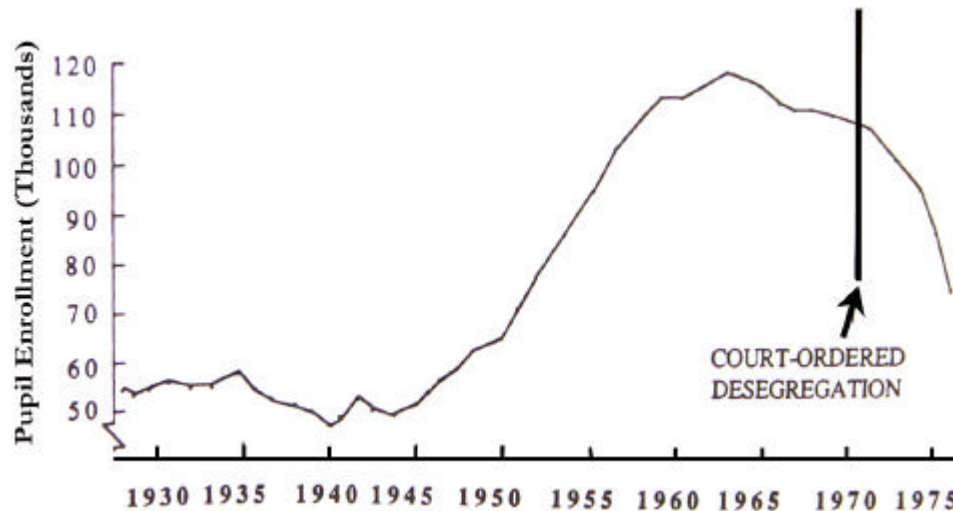


Figure 11. Enrollment in Denver Public Schools from 1928 to 1975.

secondary grades, particularly grades 11 and 12 where parents would likely decide to stick it out for the short run. If enrollment data existed separately for different ethnic groups, these time series would provide a very revealing test. If they showed roughly equal declines across all ethnic groups, the "white flight" hypothesis would suffer a major setback. Data on enrollment that could be separated by individual school, neighborhood, or census tract would be exceptionally valuable. These various units could be ranked prior to looking at the data on their susceptibility to white flight. Such a ranking could be based on variables like pre-1969 ethnic mixture," percentage of change in ethnic-mixture of the school under the desegregation plan," or mobility of families based on percentage of housing values mortgaged or amount of disposable income." If the large enrollment declines fell in the highly susceptible regions, the pattern would constitute some degree of support for the white flight hypothesis.

Conclusion

At its easiest, time-series experiments require little more than good graphing skills, a skeptical attitude toward one's pet hypotheses and the capacity to subdivide the data to locate hidden effects. At its most complex, it involves complicated statistical analyses to separate the unaccountable variation of indices across time from the determinant effects of planned interventions.

References

- 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.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Gettinger, M. (1993). Effects of invented spelling and direct instruction on spelling performance of second-grade boys. *Journal of applied analysis of behavior*, 26, 281-291.
- Glass, G. V, Willson, V. L., & Gottman, I. M. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.
- Komechak, M. G. (1974). Thought detection as a precipitator of anxiety responses. Doctoral dissertation, North Texas State University.
- Kutchinsky, B. (1973). The effect of easy availability of pornography on the incidence of sex crimes: The Danish experience. *Journal of Social Issues*, 29, 163-181.
- Levin, J.R. (1992). Single-case research design and analysis: Comments and concerns. Chp 8 (pp. 213-224) In Kratochwill, T. R. & Levin, J.R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. N.Y.: Lawrence Erlbaum.
- McCleary, R. & Welsh, W. N. (1992). Philosophical and statistical foundations of time-series experiments. Chp 3 (pp. 41-91) in Kratochwill, T. R. & Levin, J.R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. N.Y.: Lawrence Erlbaum.
- Parsonson, B.S. & Baer, D.M. (1992). The visual analysis of data, and current research into the stimuli controlling it. Chp. 2 (pp. 15-40) in In Kratochwill, T. R. & Levin, J.R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. N.Y.: Lawrence Erlbaum.
- Quantitative Micro Software (1994). Micro-TSP: Version 7. Irvine, CA: Quantitative Micro Software (4521 Campus Drive, Suite 336, Irvine, CA 92715).
- Ross, H. L., Campbell, D. T., & Glass, G. V (1970). Determining the social effects of a legal reform. In S. S. Nagel (Ed.), *Law and social change* (pp. 15-32). Beverly Hills, CA: SAGE.
- Smith, M. L. et al. (1976). Evaluation of the effects of Outward Bound. In G. V Glass (Ed.), *Evaluation studies review annual: Vol. 1* (Chapter 19). Beverly Hills, CA: SAGE.

Suggestions for Further Reading

The following references deal with general questions of the design of time-series experiments and their application to specific fields of study.

Campbell, D. T. (1969). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 212-242). Madison: University of Wisconsin Press.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409-429. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Glass, G. V, Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Kratochwill, T. R. & Levin, J.R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. N.Y.: Lawrence Erlbaum.

The following references deal with problems in the inferential statistical analysis of time-series experiments. Unfortunately, they are not simple. Each requires some background in statistics and a good deal of dedication to extracting its message.

Box, G. E. P., & Tiao, G. C. (1965). A change in level of non-stationary time-series. *Biometrika*, 52, 181-192. (The first rigorous treatment of the intervention problem. Provides the accepted solution for the most commonly encountered nonstationary process.)

Glass, G. V, Wilson V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press. (Presents a solution to the intervention analysis problem in the general case.)

Gottman, J. M., & Glass, G. V (1979). Analysis of the interrupted time-series experiment. In T. R. Kratochwill (Ed.), *Strategies to evaluate change in single subject research*. New York: Academic Press. (Further applications of the techniques derived in Glass, Willson, and Gottman (1975).)

McDowall, D., McCleary, Meidinger, E. E., & Hays, R. A. (1980). *Interrupted time series analysis*. Beverly Hills, CA: SAGE.