Journal of Methods and Measurement in the Social Sciences
Vol.3, No. 1, 52-85, 2012

Speculations on Quasi-Experimental Design in HIV/AIDS Prevention Research

Donald T. Campbell and **Beatrice J. Krauss**
Lehigh University
City University of New York
School of Public Health at Hunter College

This paper provides a speculative discussion on what quasi-experimental designs might be useful in various aspects of HIV/AIDS research. The first author's expertise is in research design, not HIV, while the second author has been active in HIV prevention research. It is hoped that it may help the HIV/AIDS research community in discovering and inventing an expanded range of possibilities for valid causal inference.

Keywords**:** HIV, prevention, research design, quasi-experimental

Two types of HIV/AIDS research are considered. Most HIV/AIDS prevention educational efforts (HIV-Ed) are likely to be quasi-experimental in nature, since random assignment to treatment, and isolation of the designated experimental and control subjects from other HIV-Ed efforts are very difficult to achieve. For therapies directly addressing the presence of the Human Immunodeficiency Virus and therapies for HIV+ persons designed to prophylactically prevent or treat AIDS-associated conditions, random assignment and isolation from rival therapies is already being achieved on a very impressive scale. Nonetheless, we consider some quasi-experimental designs that can provide valid causal inference without a "deprived" control group equally deserving and as needy as those in the experimental group(s).

We have two models of "true" experiments with which to compare "quasi" experimentation. First is the older model of experimental "isolation" and laboratory "control." Second is the randomized assignment to treatment. We social scientists who have worked on quasi-experimental design come out of the randomized assignment tradition, and share with it the out-in-the-real world lack of isolation and control. We often end up using our analyses of the implausible assumptions required by specific quasi-experimental designs applied to specific problems to argue for random assignment experiments. (The title of Campbell and Boruch, 1975, illustrates this: "Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects.") But we also end up recommending specific quasi-experimental designs for specific problems and providing advice on how to increase their clarity of causal inference.

In this paper, we want to distinguish the type of quasi-experimental designs we focus on from multivariate correlational approaches, causal modeling, path analysis, LISREL, etc. This contrast is made clearer if one recognizes that quasi-experimental research designs also have a special kinship with the older laboratory-experimentation tradition. For both, the rival causal influences to be "controlled" have to be specified, measured, and worked on (or assumed away) one by one. (Random assignment, in contrast, offers mass-produced control of "all" rival sources of change, including unknown and unspecified ones.) This kinship can be epitomized by the need to seek out "natural laboratories" for quasi-experimentation, recognizing that these will be few and far between, and that most observational data will be uninterpretable as to causal impact.

In line with this search for "natural laboratories" goes a strategy that puts clarity of causal inference ahead of representativeness. The strategy is to first achieve some clear-cut instances of preventive programs effective in unique locales, and then later worry about the other locales to which this effect might be generalized, exploring these others by theory-guided cross-validation. We recommend rejecting the notion that the efficacy of a prevention program can be ascertained everywhere the program is in place, and that therefore we should do this estimation on a representative sample of applications. On the contrary, only rarely will these be "natural laboratories" in which the causal impact of a prevention program can be estimated. Pilot studies of prevention programs should be done in these few settings that foster clarity, as also should our retrospective quasi-experimental evaluations of prevention efforts already in place.

While this paper focuses on seeking control through the design of data collection (e.g., collecting data on comparison groups not receiving the experimental treatment, or using pre-experimental time series to forecast what the data would have been like without the intervention, compared to with the intervention), we do want to recommend LISREL and EQS (Bentler, 1990) "Measurement Modeling" over the more traditional regression approaches which are too often used to misleadingly try to improve such quasi-experimental data sets. Partial correlation, covariance adjustments, causal modeling using measured variables (rather than latent "causes" measured with unreliability and reliable irrelevant variance) matching on pretests and other so-called "independent" variables, all neglect the bias resulting from "error in independent variables," all produce "regression artifacts" masquerading as causal effects. Either graph the data in its raw natural units, leaving pre-treatment and post-treatment data in a comparable metric, or use adjustments which recognize and attempt to avoid the bias due to error in variables.

While the focus of the newer econometric methods associated with Heckman's name tends to be on the statistical procedures employed, as represented by Moffitt (1991) at least, there is also an emphasis on "design" issues, or "natural laboratories." One of their approaches is to locate variables of type "Z" (in their notation) "that affect the availability of the treatment to different individuals but not their behavior directly" (Moffitt, 1991, p. 343). Random assignment to treatments provides such a Z, but so also could arbitrary factors leading to differences in city or state provision of a treatment program without that “Z” also being correlated with levels of the outcome variable. Moffitt (1991, p. 361) says, "Indeed, to use the language of economics, it is probably not possible to locate appropriate “Z” variables on the 'demand' side of the market - that is, among individuals who are availing themselves of the programs - and it would be more fruitful to look on the 'supply' side, where availability of programs is determined in the first place." That is, on the demand side, the latent variables determining individual differences in taking advantage of programs are, in Moffitt's judgment, almost impossible to model. However, random invitation to intervention designs can be put into effect combined with measurement of who does and does not take advantage of available interventions (Kessler, 1993) to address demand biases as noted later in section V. of this manuscript. If we can judge all the Heckman methods from Moffitt's overview, they differ from LISREL Measurement Modeling in that they make no effort to model and control for such bias. Instead, in the "Z" approaches, they seek out special subsets of data that constitute "natural experiments," or "unbiased" comparisons, an approach very much in the spirit of the quasi-experimental design tradition.

I. Impact Assessment in Time Series of Administrative Records

Where there are (or can be set up) administrative records which are useful as outcome variables, these often make possible often powerful quasi-experiments. Figures 1 through 7 provide illustrations of effective and ineffective programs (from Campbell, 1976). The discussion of threats to validity (i.e., plausible alternative explanations of the outcomes) will be omitted for these figures, in favor of raising such issues in the HIV/AIDS context. Visual presentation of the time series and the point of intervention is recommended, and should be used to seek out rival explanations of the ups-and-downs from persons knowledgeable about the specific situation and time period. Tests of significance in the tradition of Box and Tiao (1975; McCleary & Hay, 1980) are available, and occasionally will find significant effects where visual inspection might not recognize it (e.g., Figure 7, below). On the other hand, visibly plausible evidence of effect may not prove to be statistically significant by the Box and Tiao tests, due to too few measurement occasions, for example (e.g., Figure 2), and should nonetheless be presented as supporting evidence.

IA. Time-Series of HIV/AIDS Related Treatments in Clinical Trials

In the last 25 years, the norms for testing new therapies have completely shifted to randomized trials. Overall, this shift is to be applauded, but it has been so overdone that credible evidence from nonrandomized clinical trials is now being neglected. It is now time to devote methodological attention to improving such clinical trials, making them more effective in clarity of causal inference. Randomized trials are costly and awkward to implement. Innumerable therapeutic packages need exploring (see Volberding, 1990). Interpretable clinical trials with encouraging effects should be used to pilot-test therapies, discovering those promising enough to warrant expensive randomized trials. In addition, even though new medicines are always initially in short supply (so that most of the appropriate needy are going to go untreated anyway, so that randomized trials serve to decrease the number of untreated, not increase it), designated untreated control groups feel deprived, and vocal opposition to such experiments is generated (see Folker, 2009, for activist Martin Delaney's role in fast-tracking HIV drug approval).

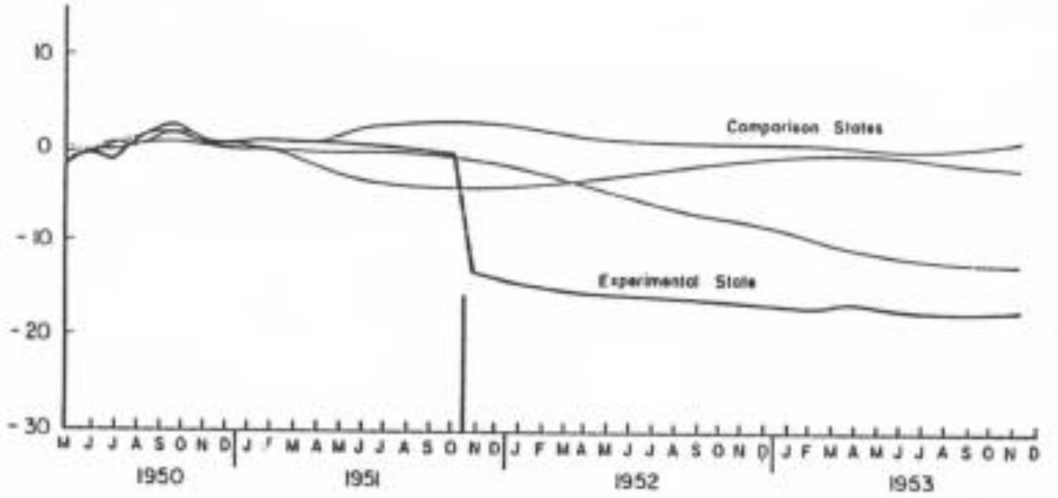


Figure 1. Effect of introducing a law in the Experimental State requiring repayment of welfare costs from the deceased recipient's estate on the old age assistance caseloads. Monthly data have all values expressed as a percentage of the caseload 18 months prior to the change of the law. (Modified from Baldus, 1973, p. 204, Figure 1.)

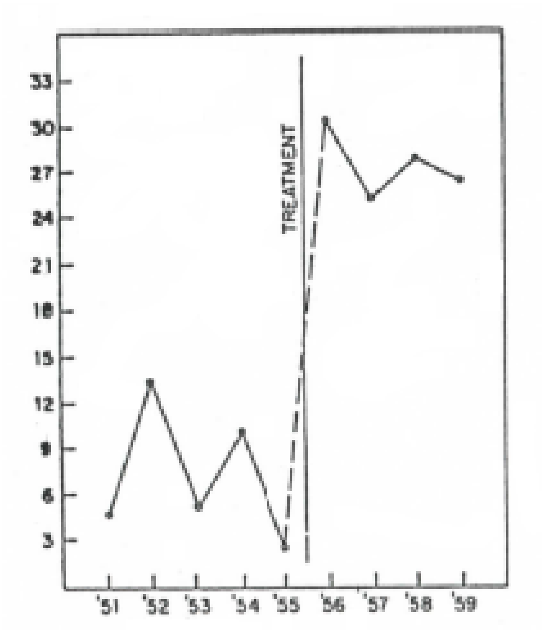


Figure 2. Suspensions of licenses for speeding, as a percentage of all suspensions before and after the Connecticut crackdown on speeding. (Campbell & Ross, 1968).

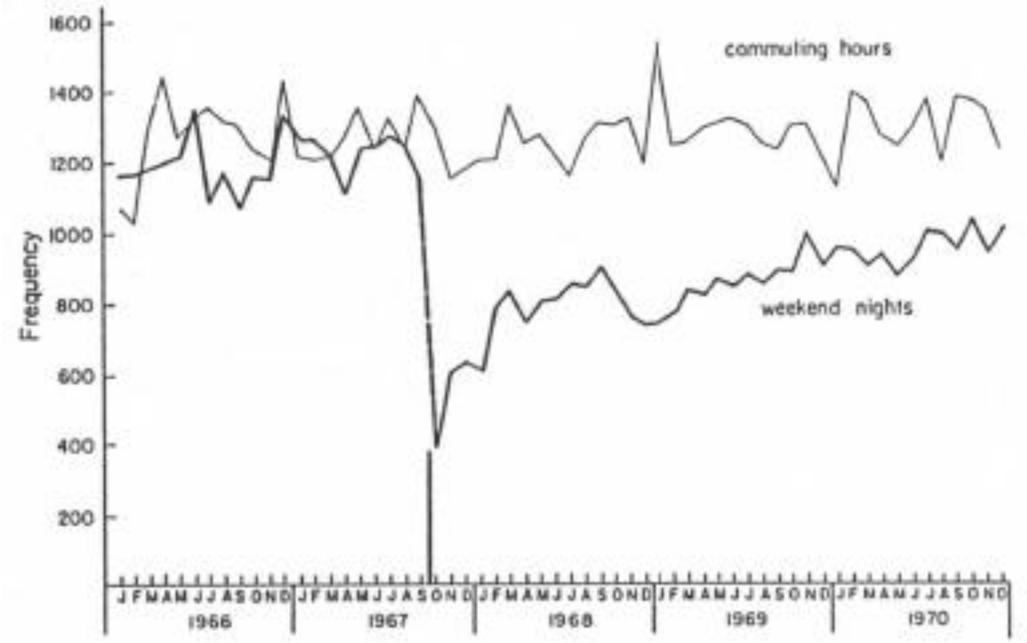


Figure 3. British traffic casualties (fatalities plus serious injuries) before and after the British Breathalyser crackdown of October 1967, seasonally

adjusted (Ross, 1973, Figures 10 and 11 combined). Bars were closed prior to and during commuting hours.

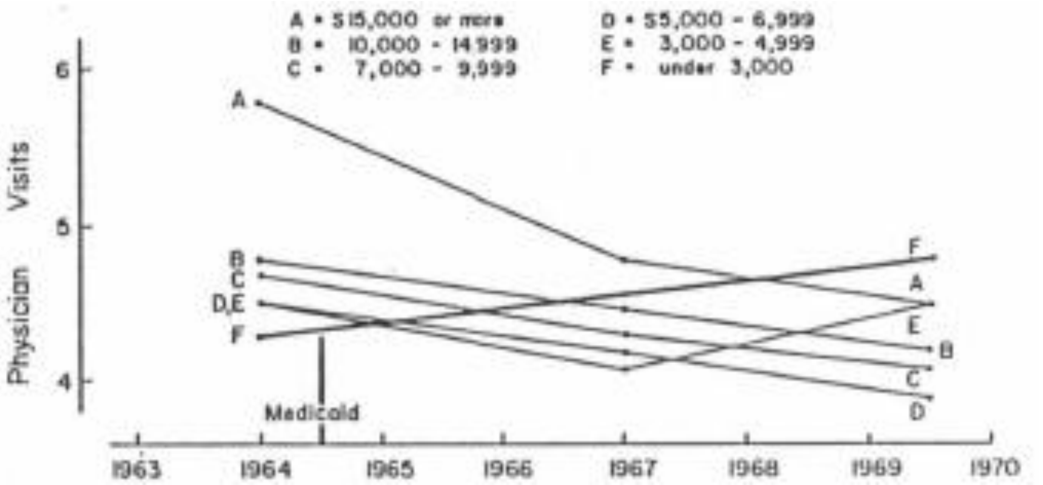


Figure 4. Possible evidence for the effect of Medicaid on contacts with doctors by persons in low-income families. The first data set is based on weekly surveys carried out between July 1963 and June 1964. The second set come from July 1966-June 1967. Eligibility was extended to Group E in 1968.The third wave is entirely within 1969. (Wilder, 1972, p. 5, Table II.)

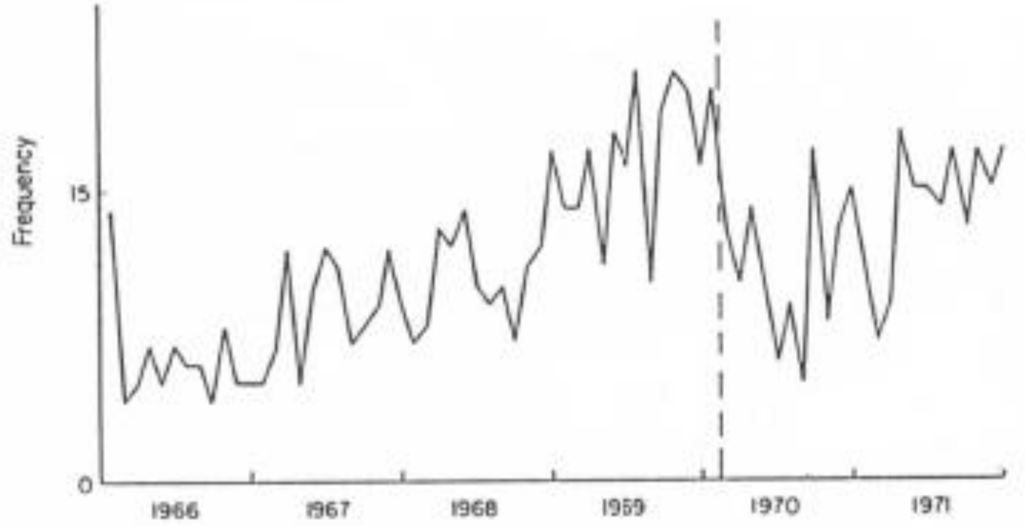


Figure 5. Gun homicides by month, Washington, D.C., 1966-1971. 'Operation Disarm the Criminal' operated January to June, 1970 (Zimring, 1975, p.189).

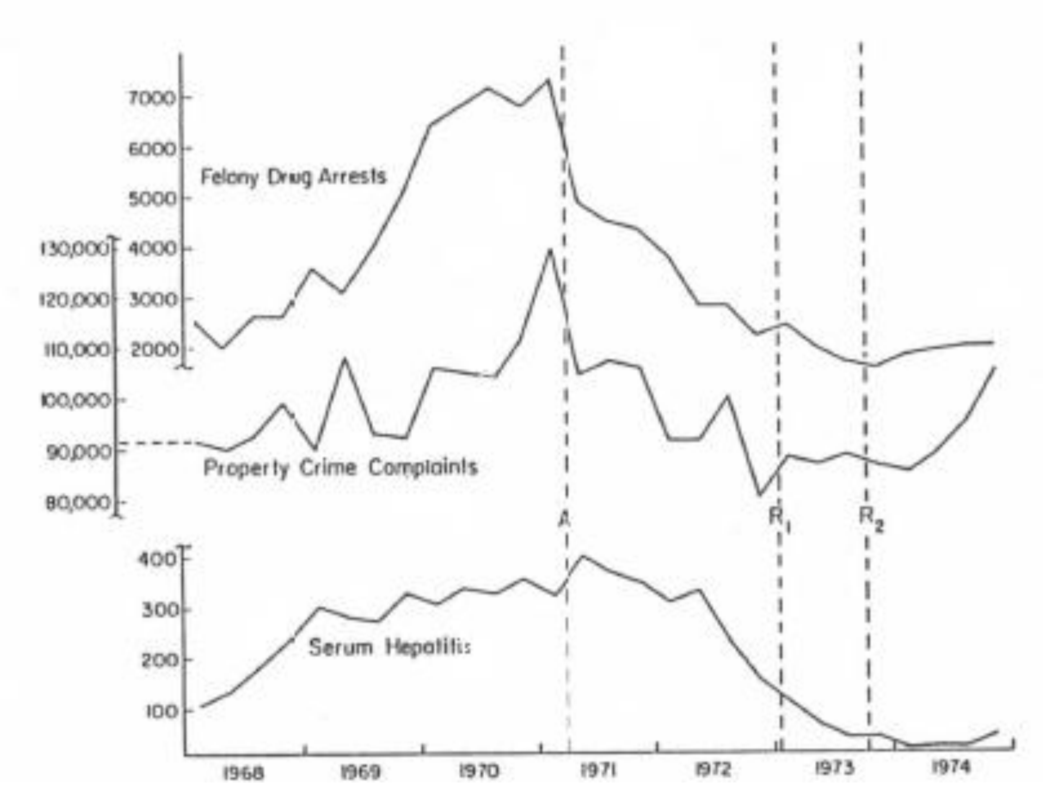


Figure 6. New York City records of possible relevance to drug abuse (courtesy of Anthony F. Japha, Drug Law Evaluation Project, Association of the Bar of The City of New York and Drug Abuse Council, 1973).

A. In March of 1971, the Police Department initiated a major change of policing effort, increasing attention to drug suppliers, decreasing attention to drug users.

R1. January 1973, Governor Rockefeller announces proposed new stricter drug laws and criminal penalties.

R2. September, 1973. Rockefeller drug laws take effect.

Felony Drug Arrests. (Quarterly frequency.) The sharp drop after A presumably shows the anticipated change in police arrest activity rather than a drop in drug use necessarily.

Property Crime Complaints. (Quarterly, seasonally adjusted.) Commonly regarded as an indicator of drug addiction. Is the drop at point A evidence of impact of the police campaign against suppliers? Comparison data from Hoboken, Boston, and Philadelphia would help.

Serum Hepatitis. (Quarterly. Excludes transfusion-based and infectious hepatitis.) This type of hepatitis is spread at least in part through hypodermic needles used in shooting drugs. Is the decline a delayed effect of policy change at A? Or is it due to a decrease in vigilance in reporting hepatitis cases (as the parallel drop for infectious hepatitis, not shown, might indicate)?

At the present time, drug abuse indicators are probably too much influenced by police effort and unknown forces, and reform programs are too delayed in their effects, or too weak, to produce clear-cut evidence of program impact. But the data are encouraging enough to justify the refinement of indicators, the search for new more direct measures less influenced by extraneous forces, and the use of comparison series from jurisdictions not impacted by the program under study.

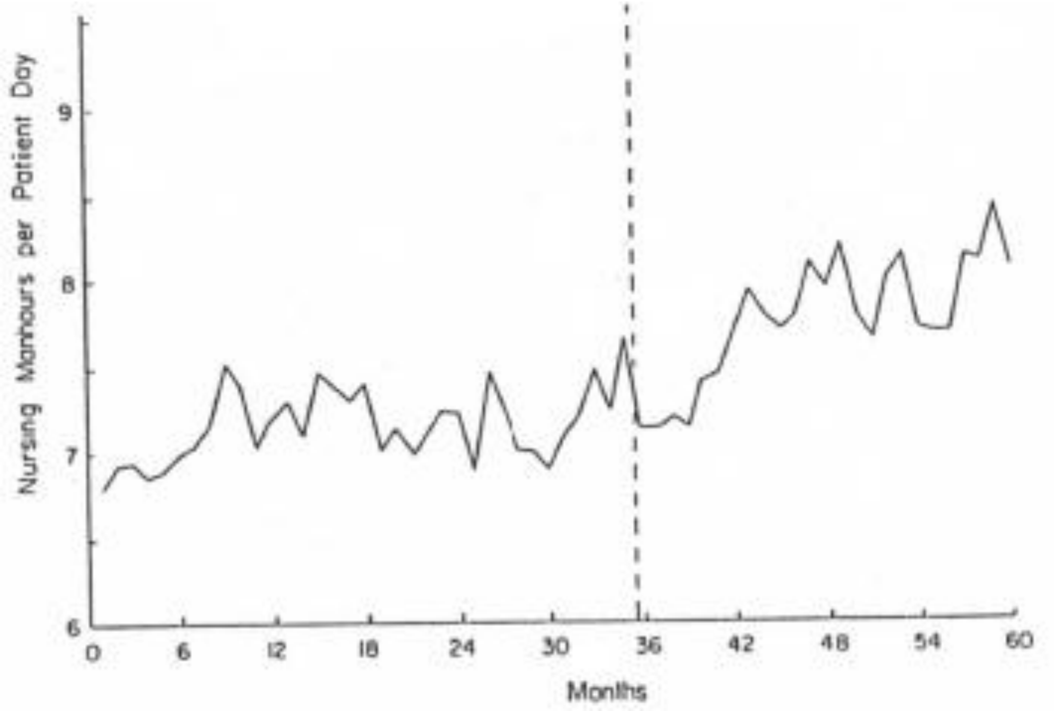


Figure 7. Effect of hospital merger on nursing work hours per patient day, seasonally adjusted (from Whittaker, 1974). The merger on the 35th month produced a significant *increase* in hospital costs, using the Box and Tiao (1965) statistics. Similar increases were also shown on other indicators using both monetary and work time units. Note that these findings go directly against the hopes and indeed the official reports from merger programs, where analysis of this detail have not been made. Roos (1973) reports similar findings, and points out that, with comparison hospitals added, this is the best evaluation method available, randomized assignment of hospitals to mergers being out of the question.

As a methodological model, we might look into the research establishing penicillin, for example the 5 day cure of syphilis introduced around 1940. Our unfootnoted understanding is that no randomized trials were used. Instead, the clinical trials (quasi-experiments) were totally convincing. Patients whose blood tests had regularly shown spirochetes for years tested free of them after one such treatment. Were we to graph the results for a single patient, it would probably have looked like Figure 8.

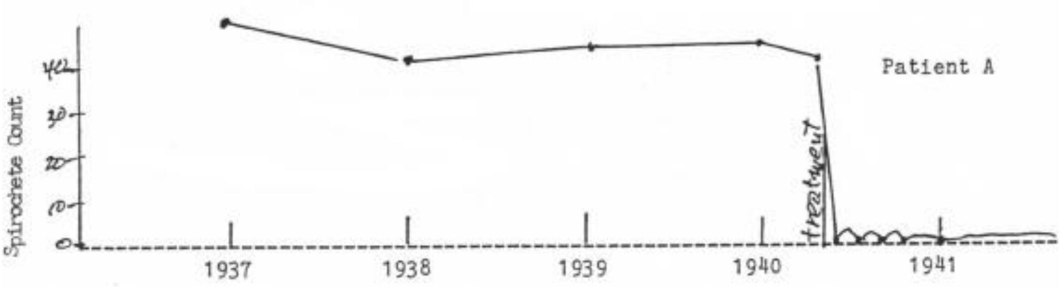


Figure 8. Hypothetical time series of spirochete counts for a single patient.

Replicated over dozens of clinical trials, and rarely refuted in later applications, the clinical and scientific communities were completely convinced, and rightfully so. Randomized trials would have been superfluous. Probably such graphs were not prepared, nor had there been any methodological emphasis on keeping the method of assay comparable across times for a single patient and across patients. For less dramatic cures, clinical practice needs to be improved for clarity of causal inference (Campbell, 1985) Repeated changes to new methods of assay may reduce this clarity. When they are introduced, the older method often should be continued as well, just to keep the time-series interpretable.

This method is most appropriate for long-standing pathology indicators that have been repeatedly measured prior to the onset of treatment. It is not appropriate for acute flair-ups. For reasons to be discussed below, it would be inappropriate for testing cures for fevers. (Every cure would be found effective.)

To make the discussion more concrete, consider the controversy over "Compound Q" at the Sixth International Conference on AIDS in San Francisco. For this draft, we based our discussion on The New York Times (National Ed., page 24), Saturday, June 23, 1990, entitled "Tests of New AIDS Drug Assailed at Parley."

> "Martin Delaney, who heads Project Inform, the San Francisco AIDS organization that is coordinating the trials, said the 46 patients in the experiment had improved significantly over the first four months of taking the drug. ... Before participants in the trial started taking Compound Q, they were losing immune system cells called CD-4 cells at an average rate of one cell every three days. While taking the drug, he said, they gained an average of two cells every three days. ... But there is no control group in the Project Inform experiments; each patient's condition is simply compared to his condition before he

> started taking the drug. Many researchers are extremely critical of this approach. ... Mr. Delaney has said his organization's unconventional trials were necessary because people with AIDS were already taking the drug in larger doses on their own.
>
> "Mr. Delaney's announcement today was attacked immediately by Dr. Arnold Relman, the editor of The New England Journal of Medicine. The two were on a panel on clinical trials. 'You don't know and we don't know whether this is just a flash in the pan,' Dr. Relman said. While he approves the expansion of clinical tests to get drugs to fatally ill patients sooner, Dr. Relman said, he is 'opposed to irrational and uncontrolled experiments!' Other researchers at the conference said these data were not enough to make the case that the drug had been helpful, and criticized Mr. Delaney for not providing more information. Dr. Relman said it was wrong for Mr. Delaney to give selective bits of data to the public very early in the experiment, before review of the data from independent researchers."

We have not examined the details of Delaney's data. In the long run, they did not justify initial claims, but did lead the way toward later, well-controlled fast-track drug research. But they also could have been compelling, as was the penicillin-for-syphilis case, *without a control group*. Let's assume 1.) that he had repeated measures on the CD-4 T-cell level for five or so months prior to the introduction of 2 therapy (or X therapy, as in Figures 9 and 10), and for a similar number of months afterward; 2.) that the introduction of therapy was *not* timed as a response to a particularly low measure (in fact because of the variability of CD-4 counts within individuals and between laboratories, it has been recommended that repeated counts below a clinical cut-off occur before therapy is initiated, Maini et al., 1996; Sax, Boswell, White-Guthro & Hirsch, 1995) ; 3.) that during the 10 month period no other therapy was introduced. (A *constant* background of other remedies during this 10 month period would not be invalidating.) The outcome of single patients might look as in Figure 9A. Such an outcome is compelling insofar as there are no plausible rival explanations for the change in slope. Most of laboratory experimentation in the physical and biological sciences similarly lacks a control group.

In Figure 9A, the treatment seems to have been successful in slowing the rate of decline or reversing it, but the results are not compelling for single cases (in contrast with the penicillin/syphilis example). Combining those of Delaney's 46 cases for which there are a sufficient number of pre- and post-measures and producing an average time series (aligned in terms of months before and months after, rather than calendar time) could produce a smooth and convincing plot, such as in Figure 9B.

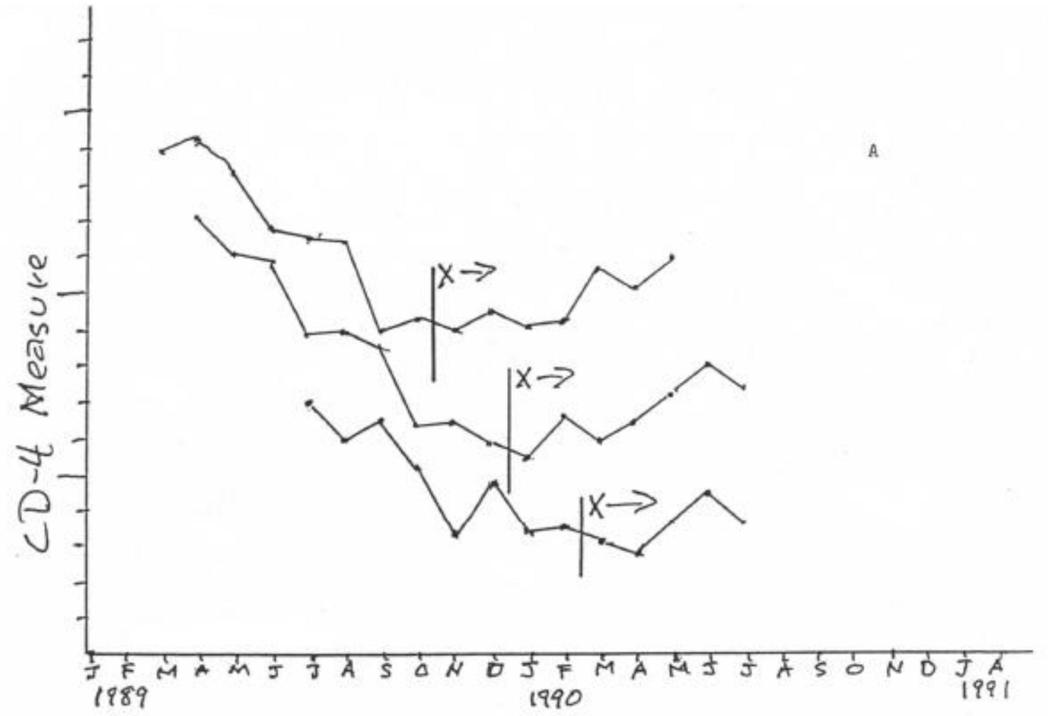


Figure 9A. Hypothetical individual time-series of a CD-4 measure.

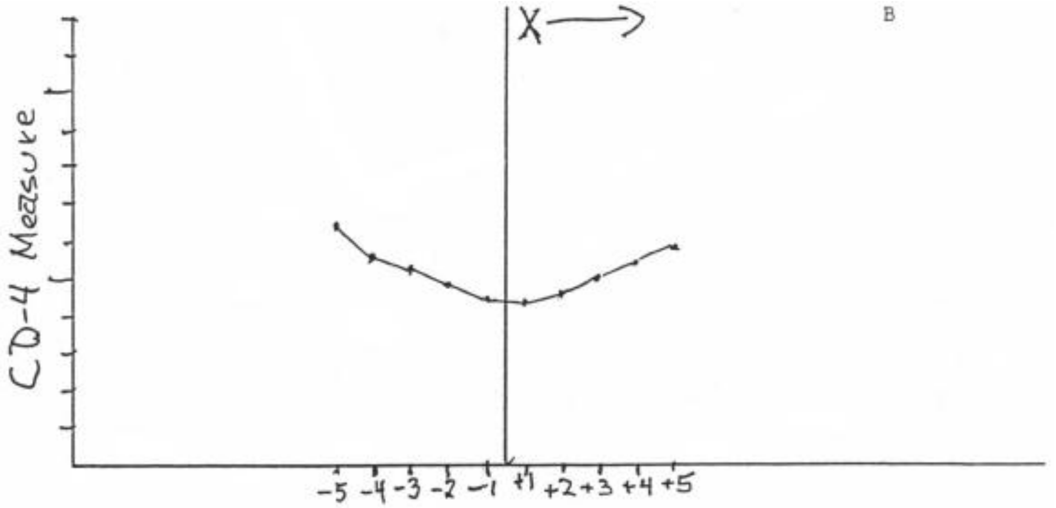


Figure 9B. Average of many individual time-series of a CD-4 measure, reorganized before averaging into months before and after introduction of Therapy X. (Hypothetical demonstration of an effective therapy, without a control group.)

No doubt we already know a great deal about the frailties of CD-4 T-Cell tests, often used for decisions about initiation of therapy. Do doctors with HIV+ patients obtain frequent enough blood tests or account for factors influencing variability? It might be that usable data can be obtained from existing patient records. More likely, the Feds (e.g. NIAID, NCHS, CDC, etc.) should provide supplemental funding to several thousand clinicians with HIV+ patients so that such time-series are available on a number of indicators against which to clinically test new therapies. It would also be necessary to collect patient data on other therapies being tried from other suppliers.

The most likely source of a pseudo-effect in a case such as Delaney's Q comes from a combination of an erratic time-series of measures and the initiation of treatment in response to an extreme measure (Campbell, 1969, pp. 412-414; 1984). If the CD-4 T-Cell measure showed the sort of instability illustrated by the three patients in Figure 10A, and if treatment was always introduced right after an extremely low measure, then on the average, the immediately following measures would show a less extreme departure from the general trend, even if the treatment had no effect.

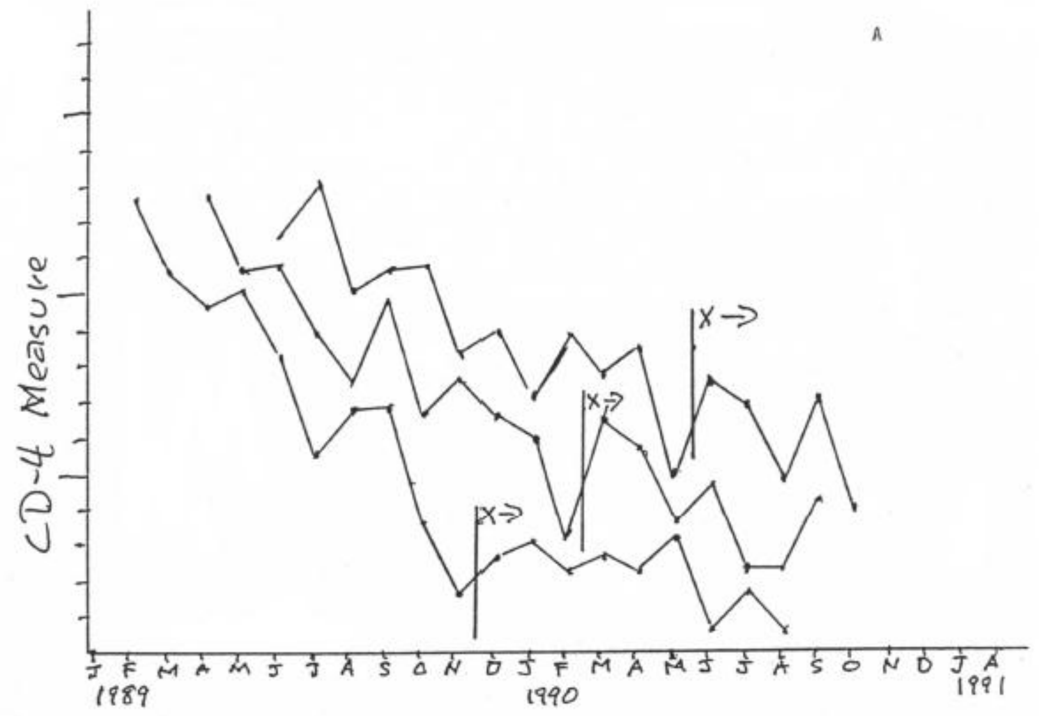


Figure 10A. Hypothetical individual time-series in which there is no true reversal of trend, but in which Therapy X is always introduced after an erratic low point.

Figure 10B illustrates this for an average of many cases, realigned after the initiation of X. (The shape of this pseudo-effect curve can be estimated from a time series correlogram of autocorrelations of differing lags based upon records in which no treatment was introduced. In Figure 10B, a first order Markov process has been assumed, with a coefficient of approximately .5. That is, from the extreme point just before treatment was introduced, the adjacent points before and after are half way back to the basic trend line, those two points away, for which r= .5x .5 =.25, the regression back to the trend has been 75%, for three points away, r = .5 x .5 x .5 = .125 or 87.5% regression toward trend, etc.) From Delaney's case records we should be able to decide whether the onset of treatment was, in a given case, precipitated by an extreme measure. From his and other records, we should be able to estimate typical CD-4 T-Cell trends for HIV positive patients in the absence of treatment, and the autocorrelation coefficients.

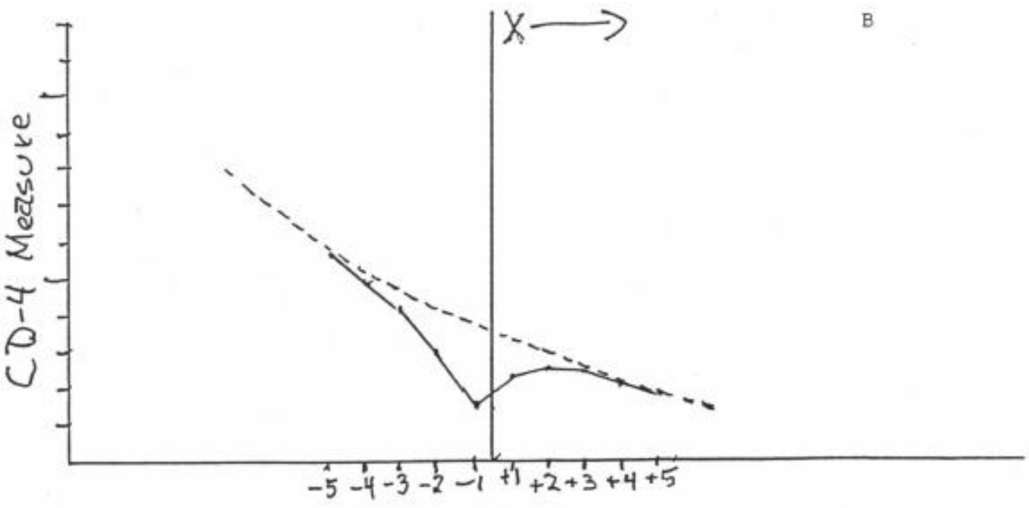


Figure 10B. Hypothetical example of the average of many individual time-series of a CD-4 measure (reorganized before averaging into months before and months after the introduction of Therapy X) in which a pseudo-remission result is produced by introducing the therapy after extreme low points in the series. The dashed line represents the “true” trend of the measure.

For prospective quasi-experiments of this type, there are precautions that could be taken. After the decision to introduce Therapy X, one could routinely wait several measurement periods before starting it. Or one could introduce such a delay only in cases in which the decision was made after an extreme measure, as judged by the expected trend for such cases and that patient's own measurement series.

IB. Administrative Time Series Evaluation of a Public School HIV/AIDS Prevention Program

Let's assume, for example, that we have lucked into a natural laboratory with these features: There are two reasonably comparable large high schools, with differing school boards and administrators, one of which introduces an intense three-year HIV/AIDS prevention educational program several years ahead of the other. More luck: They are both in a state that requires HIV testing before marriage, and/or they are both served by hospitals that routinely do HIV blood tests of mothers giving birth to children, and by clinics that give HIV blood tests for all abortion patients.

We should regard settings providing such regularized blood testing as precious natural labs. Such hospital practice is widespread. We understand that the State of Illinois for a while included HIV in its premarital blood tests, but has since removed this requirement because of the low rate of positives. The HIV/AIDS research community should have testified for retaining it. We can handle rare events (their standard errors are very small), the rate was bound to increase, and in those communities still with virgin women getting married, the tests would have prevented some important, if rare, tragedies. We should concentrate our prevention program pilot studies in such natural laboratories (Campbell, 1987, pp. 415-420, 425-426). It is not impossible that such interpretable quasi-experiments already exist, waiting to be analyzed. We should, of course, do what we can to increase their number, and to increase their potential clarity of causal inference.

In this fancied retrospective study, we find out what high school and what years attended for each HIV positive case, and for a representative sample of HIV negatives (several times as large, since the added precision is cheap). We hope that all HIV positives were notified, and were asked what their likely route of infection was. As a part of that inquiry, background information including schools attended might be asked. Or long lists of persons tested by the hospitals (both HIV+ and HIV-) could be compared with attendance names for local schools. This can be done without informing school-record custodians of the HIV tests (Campbell, et al., 1977) or their outcome. While not all of the 95% of HIV- outcomes should be included in the search list (for reasons of economy), the HIV-cases searched should be much more numerous (e.g., 5 times as many) as the HIV+ cases, all of which are searched. (It is very important for HIV/AIDS research to devise ways in which names can be connected with HIV+ test results without harm to the tested individuals, and to obtain approval of such procedures.)

For a conspicuously effective preventive education program, we might get an outcome like such is shown in Figure 11.

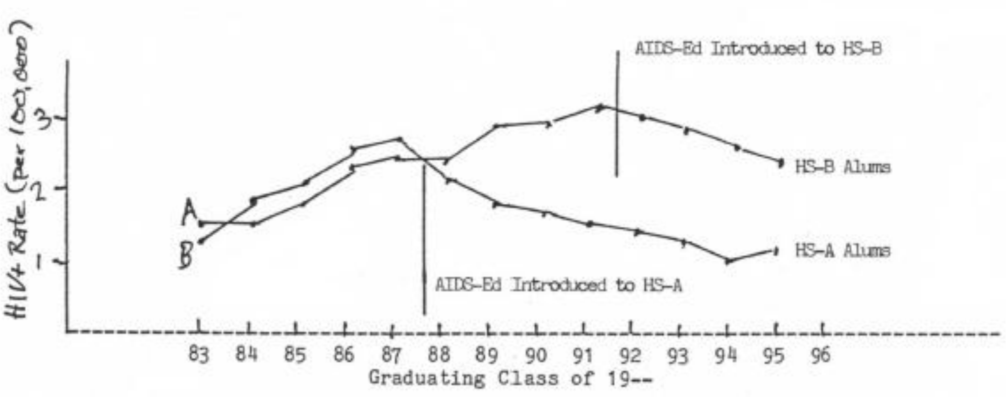


Figure 11. Hypothetical HIV+ rates for two high schools, by graduating class.

This graph uses only the HIV+ alumni. The HIV- ones could be used to check on changes in the representation of the high schools in the record-keeping catchments. Rates of HIV+ could be computed using other records on high school alumni numbers, or using the HIV- controls since HIV/AIDS-Education is apt to also reduce the number of pregnancies and abortions. School B, late in introducing the HIV/AIDS prevention training, provides a useful control for School A during 1988-91, and a cross validation after 1992. This controls for locally shared trends, and for shared shifts in HIV testing procedures. The analysis would be worth doing even if delayed treatment in School B was not available. It would be worth doing if only the HIV tests for births were available. If several HIV test result series were available, these should be examined separately as well as combined.

For a quasi-experimental design (and for a randomized design, especially since in the HIV/AIDS setting they rapidly become quasi [e.g. Turner, Miller & Moses, 1989, Chapters 5 & 6]) "laboratory notes" should be kept, systematically and extensively. These characterize the classic "laboratory control" tradition in the physical and biological sciences, but are rare in applied social science, even though much more needed. Project historians or ethnographers with treatment delivery rival change agents and measurement delivery as a part of their agenda, plus logs kept by well placed observers among the staff are essential (Campbell, 1987, pages 422-425). When graphic results are in, we should check with a wide range of well placed local observers for explanations of the ups and downs and differences (even though we know that they, like the stock market experts, will over-interpret "random" fluctuations). (In the example of Figure 11, an epidemic of needle-delivered drugs would be a significant plausible rival hypothesis for differential changes, or a crack epidemic through the use of prostitution to pay for the crack.)

We should brainstorm about other time-series that may be retrievable, such as local drug store sales of condoms. Military units requiring periodic HIV testing would provide useful natural laboratories. In fact, stored military records provided some of the first descriptions of the natural history of HIV progression without treatment (Goedert & Blattner, 1988). For such institutions, should we try to influence their frequency of blood sampling? Where there are legally required periodic blood testing for drugs for airline pilots, railroad engineers and brakemen, bus drivers, etc., do these provide settings in which HIV blood testing could be added, creating a lab for HIV/AIDS prevention experiments? (Random spot checks are unlikely to be useful.) If the HIV/AIDS prevention educational package were to be delivered by mail and telephone, we would no doubt use random assignment. An institutionalized framework of periodic measurement is useful for randomized experiments, too.

Note that these time series are about institutionally defined cohorts and institutional units for HIV/AIDS prevention program delivery. No individuals are being retested. (We might recommend that the intensive educational program be instituted in grades 6 through 12 at once, even though we have designated them "high school.") The rare cases picked up by the partial catchment net are used as indicators for their institutional cohort. They are clearly not "representative" in a representative sampling sense. They would be tapping most from those high school classmates who did not go on to college, for example, and who started childbearing most promptly (although the HIV tests used to plot a given high school cohort year could have been made over a several-year time span, reassembled by cohort year). This lack of statistical representativeness does not make them unusable as indicators. But shifts in their selection bias from year to year are plausible rival explanations of ups and downs in the graphs and their possibility should be carefully considered. Note that for the HIV rates, in the high school setting, pretests would probably be impossible for any design.

II. Before and After Measures on an Experimental and a Non-Randomly Assigned Comparison Group

In the educational-psychology methodological tradition, this is Design 10 of Campbell and Stanley (1963/66) and Chapter 3 of Cook and Campbell (1979), but in considering such designs, we will mix our presentation with consideration of random assignment possibilities.

Continuing to consider the natural lab of HIV testing for births and high school (or grade 6-12) education programs: If the available regular

HIV testing precluded multiple pre-intervention years, we would recommend recruiting high schools willing to cooperate, selecting matched pairs from these, and introducing the educational program in a randomly selected one of each pair, one or two years prior to its introduction in the other (the comparison, or control), and then following up on both for a number of years. If one had 10 such matched (blocked) and then randomized pairs, one would begin to get the statistical benefits of randomization for these institutions as statistical individuals. But even with only one pair, or two, one will have reduced the plausibility that the treated institution was selected just because of its extremity. Such a two-group comparison (or even one based upon several matched and then randomized pairs) is clearly a quasi-experiment. But well worth doing en route to adding more pairs.

Do we have any chance of doing experiments that use HIV tests and employ pretests and posttests? In the spirit of speculating about possible interpretable settings where HIV testing is routine, let's consider a first pregnancy HIV measure as a pretest, second pregnancy as a posttest. To a random subsample of mothers all HIV-, we offer a videocassette preventive education program. We keep the names of these and a randomly equivalent control group (perhaps much larger) as screening lists for births in later years in our catchment area hospitals. One of the likely outcomes will be fewer births in the experimental group as well as fewer HIV+ cases overall.

But the maternity ward that will let in researchers for the HIV/AIDS Prevention Education videotape will probably want it used for all of its mothers. Thus we move again to a quasi-experimental framework, with the hospital as the treatment unit. And the hospitals that will allow such training may be the ones that already have relatively high HIV+ rates. Thus experimental hospitals and comparison hospitals are likely to differ systematically. We believe that such studies would still be worth doing, but it would be very helpful to have the prior year's HIV+ rates for all hospitals, perhaps separately for first and for second births, since the latter is our outcome measure. We would also want trends on total births as another outcome measure.

III. The Regression Discontinuity Design

Trochim (1984) has presented this design most thoroughly (see also Campbell, 1969, pp. 419-425; Cook & Campbell, 1979, pp. 137-143). Like random assignment, a treatment is applied by explicit rule (on a measured variable rather than a latent one). Its advantage over random assignment is that there is no equally deserving control group that is being deprived of the potential boon. Instead, eligibility (need, deservingness) is made explicit in a quantified score, on which assignment is made with sharp cutting points where an abrupt difference in treatment strength occurs. Outcome measures are required not only for those receiving the treatment, but also those ruled ineligible. These outcomes are plotted in terms of the eligibility score. In the hypothetical example of Figure 12, the CD-4 percentage for HIV+ patients at the start of the study has been used as the eligibility score for the assignment of a potential therapy X (AZT, Q, etc.). In this example, those in the healthy range (50% to 70%) have been deemed ineligible due to lack of need. Those below 20% have been deemed too ill to profit from the treatment, and ineligible for that reason. The design presumes that treatment X is in short supply, that many HIV+ persons are going to go without it, and that the supply available should go to those most eligible. The non-eligible in January 1991 are kept without it for the whole year, even if their eligibility status changes, and even if the supply increases. It thus shares some of the ethical problems of the untreated control group in randomized trials, except only that those who get the treatment are more eligible.

The eligibility score may be based upon many variables (including subjective ratings), but these must be combined into a single continuous quantitative measure which is to be the sole decision criterion for those included in the study. (Those getting or being denied the treatment regardless of the score are to be excluded from the study.)

In the hypothetical case of Figure 12, there are two sharp cutoffs. In many implementations, only one, the not-needy, would be used. Within the eligible range, all get full strength treatment. A dosage level proportional to need would lack the sharp corners to show up on the outcome measure. This arbitrary abruptness provides the distinct effect-shadow on the outcome measure. A covariance analysis of the outcome measure, using dosage level as the covariate, is usually appropriate. (Where there is a latent assignment variable imperfectly reflected in any measured variable, a covariance analysis produces pseudo-effects akin to regression artifacts.)

In the illustration in Figure 12A, the outcome measure is the same as the decision criterion measure, but this is not a requirement of the method. In Figure 12B, a severity of HIV score is substituted as the outcome measure. (Longevity, or other long-term outcome measures, make implausible illustrations because of the implausibility of long-term exclusions of other therapies for both treated and untreated.)

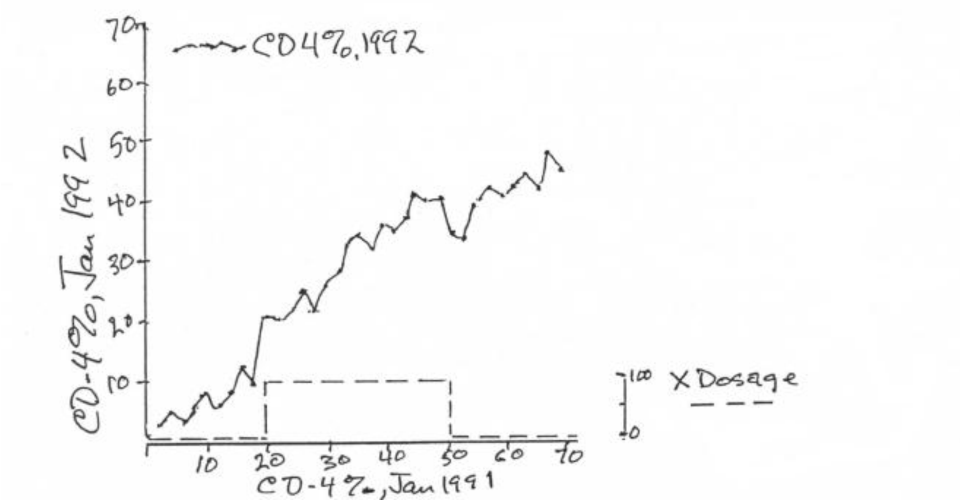


Figure 12A. Hypothetical regression discontinuity design demonstrating the effectiveness of Therapy X; see text for details.

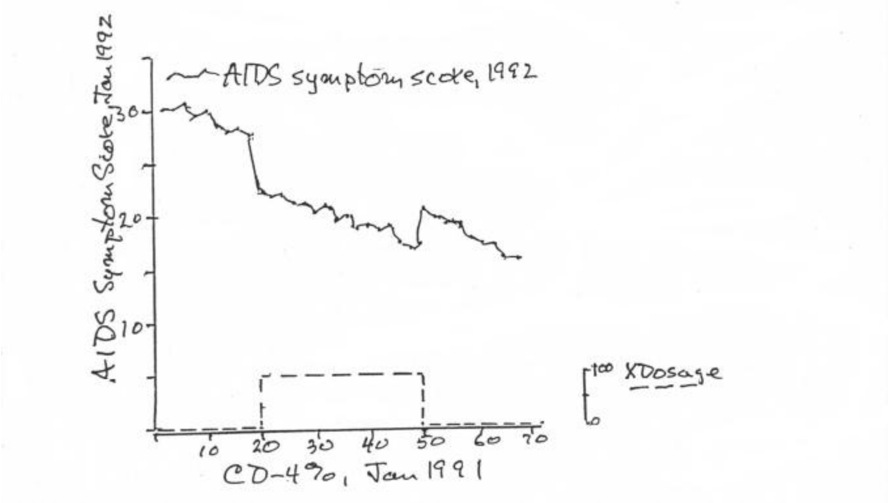


Figure 12B. Effectiveness demonstrated with an outcome measure dissimilar to the eligibility criterion.

The Regression Discontinuity Design is appropriately applied to samples of communities or institutions for some types of treatment variables, as when census data are used to determine community eligibility for a program, or hospital size determines eligibility for governmental subsidy for special equipment, etc., and a sharp cutoff is used. In our consideration of HIV/AIDS-Education program possibilities, we have not come up with a plausible example, in a setting in which insulation from other treatments could be provided.

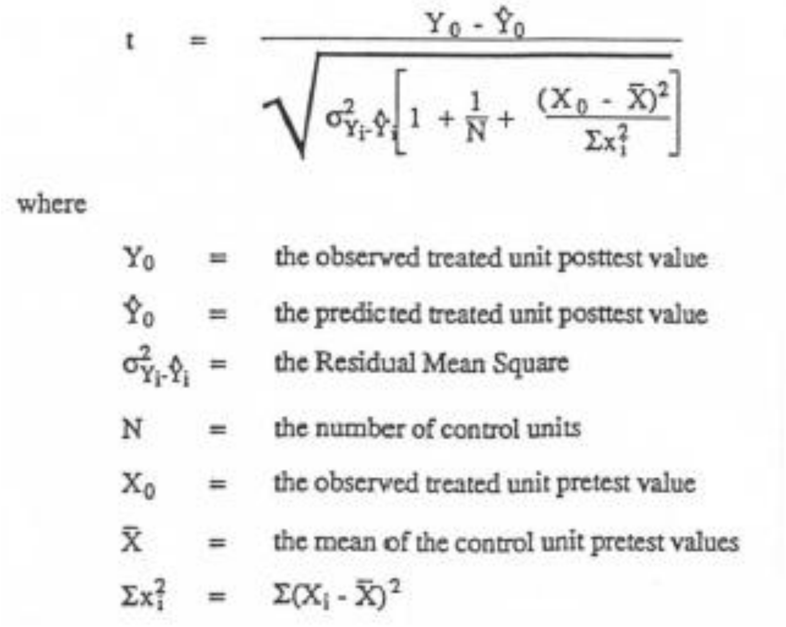
IV. The Regression Point Displacement Design

"Pilot studies" for community-wide HIV/AIDS-Education programs need as powerful a test of significance as possible. Perhaps this is it. Where a substantial number of social units (metropolitan areas in Figure 13) are measured periodically, and where one (or few) of them is (are) targeted for an intensive HIV/AIDS-Education program the design is available.

The design has been introduced into the HIV/AIDS research methods literature by Coyle, Boruch, and Turner (1991, pp. 149-159) inspired in part by the unpublished Campbell (1990), who in turn had been inspired by Fleiss and Tanur (1973), and had introduced illustrations of it in Riecken, Boruch, et al. (1974, pp. 115-116) and Cook and Campbell (1979, pp. 143-146). The presentation that follows is primarily based upon Trochim and Campbell (in preparation).

Coyle et al. (1991) present the design in concert with the Regression Discontinuity Design (RDD). Some applications of the Regression Point Displacement Design (RPDD) might be regarded as a degenerate version of RDD, as when the most extreme group on the pretest or decision measure is selected for the Pilot Study just for that reason. But Trochim and Campbell argue that it is applicable no matter what the eligibility (or pretest) score is, and no matter why chosen (although the reasons why chosen provide alternatives to the pilot program [or other treatment] as rival explanations of a significant effect).

For Figure 13, there has been a token effort to make the hypothetical illustration more realistic by using real data as a background. The pretest measure is the 1989 rate per 100,000 of new AIDS cases (actually December 1988 through November 1989). The outcome measure is the subsequent year's new cases (designated 1990). The data are from 96 metropolitan areas in the United States larger than 500,000 (HIV/AIDS Surveillance Report, December 1990, pp. 6-7, Center for Disease Control, Atlanta, GA 30333). For 95 of these areas, a linear regression has been fitted between the 1989 and 1990 rates. The data for one city (Miami) has been modified for 1990 to simulate an effect. (The real rates for Miami are 53.2 in 1989, 58.4 in 1990. The substituted rate for 1990 in Figure 13 is 48.4.) This hypothetical effect (plotted as x) departs from the regression line based on the 95 other areas with a t-value of 39.7, p<.0001, based upon the following formula:



$$t = \frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2_{Y_i-\hat{Y}_i}\left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]}}$$

where

Y_0 = the observed treated unit posttest value

$\hat{Y}_0$ = the predicted treated unit posttest value

$\sigma^2_{Y_i-\hat{Y}_i}$ = the Residual Mean Square

N = the number of control units

X_0 = the observed treated unit pretest value

$\bar{X}$ = the mean of the control unit pretest values

Σx_i^2 = $\Sigma(X_i - \bar{X})^2$

This t-value can be tested in the usual way assuming df=N-2. Note that the term $(X_O - \overline{X})^{2/} \sum x^{2}_{i}$ has the effect of producing a larger error term for experimental units lying at or beyond the extremes of the distribution. The power is greatest, and mistaken inferences due to assuming the wrong curve (linear, quadratic, etc.) least for an experimental unit lying in the middle of the distribution.

There are lots of things wrong with this example. On the statistical side, the single extreme city (San Francisco) extends too great an influence on the curve. (When computed on log transformed rates, the t-value drops to 23.7, p<.0001.) AIDS rates are dependent upon contagion events too far in the past to be affected by a one-year program between 1989 and 1990. HIV incidence rates would be better, and even for them a 3 year gap between pretest and posttest would be more appropriate.

As with the Regression Discontinuity Design, the RPDD design may be used with an entirely dissimilar outcome measure, and indeed the dramatic illustrations of the effect of Medicaid used in Riecken and Boruch, et al. (1974, p. 115) and Cook and Campbell (1979, p. 144) used income class levels and physician visits per year. (As graphed, Trochim & Campbell find a t-value of 21.9, p<.0002 for this example, with only 5 untreated and one treated group. Riecken, Boruch, et al. and Cook and Campbell did not report a test of significance.)

For all the implausibility of this example, where single site demonstrations are involved, and measures on many untreated sites are available, this may well be the analysis of choice. Consider its strengths compared with selecting a single comparison city (as in Section II, above). Here there is the statistical power of many comparison cities. Moreover, there is not the worry about comparability that occurs in a single site control. No exact matching or statistical adjustment for pre-treatment in equality is needed.

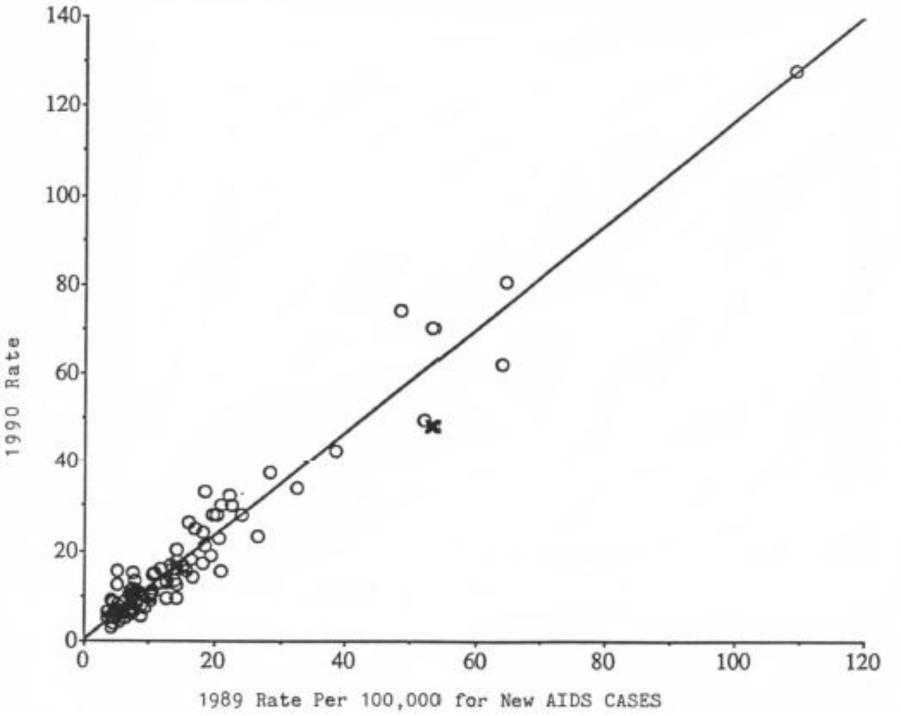


Figure 13. Hypothetical example of a RPDD study of the effects of a massive city-wide HIV/AIDS education program in Miami, indicated by the X. The circles represent the 95 metropolitan areas not receiving this treatment, and upon which the regression line is based. See text for explanation.

Where a significant effect is found, one, of course, should remember that the intense pilot program is only one of the possible explanations for it. Had Miami been chosen *just because* its 1989 value was a great increase over 1988 (as in the Connecticut crackdown on speeding example [Campbell, 1969]), then the significant departure of the 1990 value might be merely a return to the normal trend for Miami. That is, the exceptional value might be the 1989 rate, not the 1990. If the choice of Miami was only one aspect of the many other measures it was taking, over and above the HIV/AIDS-Education campaign under study, then the other measures, individually and collectively, may be the true cause. City-specific changes in the record keeping between 1989 and 1990 could cause statistically significant effects. Were one to choose a pilot city by a lottery, there would not be enough "random assignments" to provide statistical confidence in pretreatment equality, but it would render less plausible rival explanations of a significant effect, other than the pilot program.

Trochim and Campbell also provide a discussion of the much more stringent t-values required for a strategy that would start with an exceptional outlier and then ask what that city is doing to cause the significant departure. In Figure 13, the control city overlapping the X (Miami) is Newark. (We have not checked to see what its t-value is, or would have to be for significance at the 1/95x100 or 1/95x20 p-values.) The most eccentric value above the regression line is Fort Lauderdale (48.5 in 1989, 74.5 in 1990), possibly the result of better record keeping.

V. Randomized Invitations to Treatment with a High Rate of Turndown

A version of randomized pilot experiment fitting into voluntaristic norms for social and political participation is one in which a new ameliorative treatment is made available to a random subset of eligibles (let us call these E, the experimental group) some of whom accept (T, treated) others decline (U, untreated). Another random subset of eligibles is selected as a control group (C). All 3 groups, T, U, C, are measured after the treatment period. Let us assume no pretests are available. Experimental tutoring in a first year calculus course could be an example, with the shared outcome measure the final grade. The randomization provides a null expectation that the means of E and C be equal, $M_E = M_C$. But the voluntaristic division of E into U and T makes systematic selection likely, e.g., those accepting being those that give the course more importance, or have more time for schoolwork anyway, etc. So $M_T \neq M_C$, $M_T \neq M_U$ and $M_U \neq M_C$ for the three groups at hand. Figure 14 provides a simple graphic recapitulation.

If subtle questions or measures are part of the recruitment or baseline measurement, indeed some indication of the nature of selection biases can be documented [1] as long as these questions do not precondition response to the intervention (e.g., Krauss, Goldsamt, Bula, Godfrey, Yee & Palij, 2000).

[1] Krauss's first National Institute of Mental Health project (MH53834) was funded in 1994. It was an intervention designed to increase parents' abilities to prevent HIV in their pre-adolescent children in high HIV-seroprevalence housing projects (PATH). Don Campbell convinced me to use a random invitation design along with recruitment from randomly selected apartments in 10 large housing projects in New York City. "Otherwise, you will never know what kind of parent comes forward, and people will make all kinds of assumptions about them." His insight was profound. Fully 76% of eligible families participated, that is, signed informed consent with both parent and eligible child completing baseline measures. The only two predictors of coming into the project, assessed during recruitment, were 1) physical distance from the storefront where the intervention occurred (participation dropped off after a half mile), and 2) a negative response to the statement: "My child already knows enough to protect him/herself from HIV."

Visually, one intuits that the availability of the C distribution gives one a great advantage over just having the U and T distributions, as in E or E' alone. In an unpublished proposal (Campbell & Boruch, 1971) several modes of analysis avoiding a spurious M_C, or M_U, comparison with M_T were considered. The standard statistician's recommendations, "analyze 'em as you randomize 'em," compares M_E with M_C, accepting the dilution of estimated impact coming from including the U's as though they were treated. Other suggestions tried to make that dilution as small as possible without reintroducing a systematic selection. An "upper edge test" suggested a t-test on an arbitrarily truncated version of C and E. If 35% of those invited accepted, compare the mean of the top scoring 35% of the C group with the mean of the E group, including both U and T cases that fall within that range. For hypothetical cases such as E and E' in the graph, such truncation improved precision over the whole E, C comparison, even though there were fewer degrees of freedom. A 4-fold Chi-squared test, based upon the same cutting point, also produced higher significance levels. Another recommended tactic was to reduce U by eliminating from both E and C those thought particularly likely to decline (such as students with full time jobs and students who did not need the course for their major, etc.).

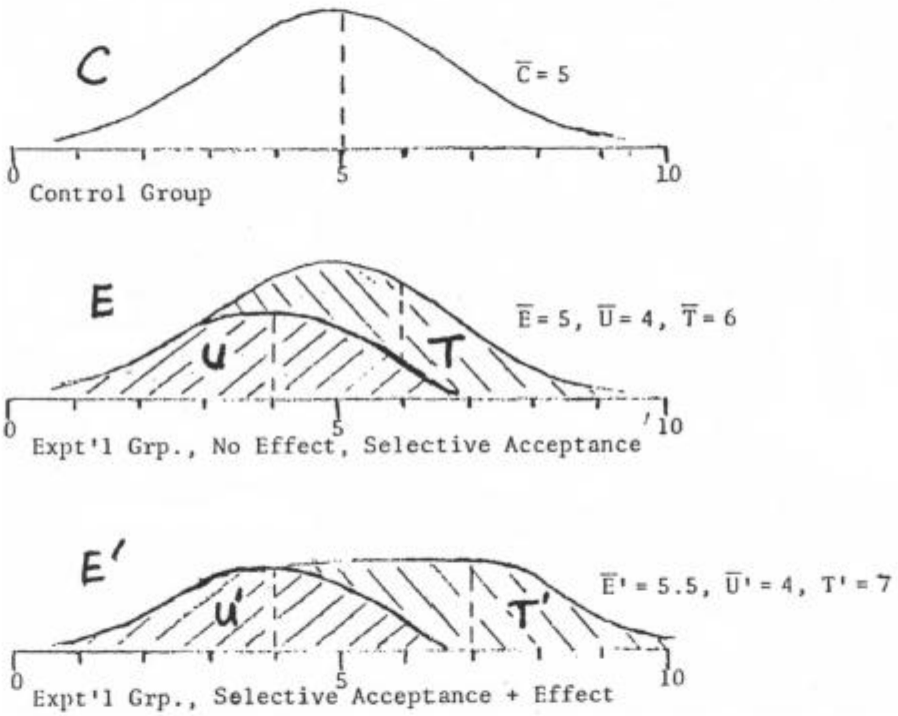


Figure 14. Distinguishing selective acceptance of a randomized invitation from the treatment effect.

If one assumes that U group members have not been affected by the invitation-declension process, then an estimate of the expected value of M_{T} under null conditions (*T, italicized to distinguish it from* T) can be estimated from observing M_{C} and M_{U}.

Randomization legitimates the expectation $M_{\mathrm{C}} = M_{\mathrm{E}}$ under no-effect conditions. E is a product of its components U and T. Working with Means, M_{E} is a weighted mean of M_{U} and M_{T}. Where n = number of persons in U and m =number of persons in T,

$$M_{\mathrm{E}} = \frac{nM_{\mathrm{U}} + mM_{\mathrm{T}}}{n+m}$$

Under null expectations $M_{\mathrm{E}} = M_{\mathrm{C}}$, so,

$$M_{\mathrm{C}} = \frac{nM_{\mathrm{U}} + mM_{T}}{n+m}$$

From the data we can ascertain M_{C}, M_{U}, n, and m, and thus estimate M_{T}. Solving for M_{T} in the above gives us M_{T}, the expected null value of the mean of T.

$$M_{T} = \frac{(n+m)\, M_{\mathrm{C}} - nM_{\mathrm{U}}}{m}$$

Comparing the obtained M_{T} with the expected M_{T} gives us the treatment impact.

While this is so obvious that it has no doubt been independently invented by applied statisticians faced with a particular version of the problem, we are not aware of any such case, nor does it appear in any textbook to date so far as we know. It seems a recurrent enough problem to merit such inclusion. An appropriate error-term for the M_{T} - M_{T} difference has been derived by Boruch (unpublished). The standard error of M_{T} can be estimated from the data. The standard error of M_{T} is much larger, as it is a standard error based upon the standard errors of M_{C} and M_{U}, both estimatable from the obtained data. The fact that one cannot assume the standard error of M_{T} and M_{T} to be equal adds considerable complexity to the problem. It turns out that this purification produces no gain in precision over the "analyze 'em as you randomize 'em" rule, comparing the total E (diluted by the untreated) and the control group. It does, however, produce a purified estimate of the effect, again, assuming the mere invitation to be without an effect.

What is particularly attractive about this design is the ability to do without a pretest. It does require the pretreatment identification of invitees, and the availability of a posttest on all three groups. It also requires that one assume that attrition from the posttest is purely random, or, if systematic, the assumption that it shows the same bias across all three groups. (Estimations of maximum plausible effects due to differential attrition from the posttest measure would assist in the interpretation of apparent effects.)

Where one had very large samples, and (though no comparable pretest) many background measures on all cases, one might be tempted to analyze the correlates of acceptance of treatment in the experimental group, and use these as measures in a post-hoc purification of the control group, justifying a comparison of the resulting C and T. However (in analogue to the shrinkage of a cross-validated multiple correlation) the covariates selected would have overfitted T, producing a regression-artifact problem. LISREL measurement modeling would do better. But best quasi-experimental practice would be to use the experience in predicting acceptance of treatment (and availability for posttest measurement) in selecting a purified pool for a subsequent randomized invitation study.

VI. Treatment-Effect Correlations

One of the problems in evaluating the efficacy of HIV or AIDS therapies, or of HIV prevention educational efforts in nonexperimental settings is that so many treatments are being received by any given segment of the population, and also that there are systematic selection biases in who gets exposed to which. The Treatment-Effect Correlation is an exploratory survey technique that potentially can control for both selection bias and unevenly distributed multiple treatments.

What it requires is two waves of measurements on "the same" "outcome" or "status" variables (let us call these pretest and posttest), plus an independent measure of what "treatments" each individual had been exposed to in between the two waves of measurement.[2] The strength (or frequency) of exposure can be indexed in degree, or a dichotomous dummy variable of presence or absence can be employed. The correlation of treatment measures with the pretest measures shows the selection bias,

[2] Measurement of exposure to “other” interventions was employed in the CDC AIDS Community Demonstration Project (1999), the forerunner of “Community Promise,” a CDC-supported evidence-based intervention. Not only did the project collect data on intervention, department of health, clinic, school and media exposures to HIV education messages, but, under the advice of Al McAlister validated self-report by including at each data collection point a “bogus” set of materials produced that day by computer which, in the New York site, only 1 out of 1,000 respondents said they had seen. While exposure to project materials increased over the 36 months of the project, this increase was accompanied by a decline in other sources of HIV information, assisting inferences about impact of the intervention.

that, with the posttest measures, shows the joint effect of selection bias plus the impact of treatment.

Even some experienced statisticians will ask, why not just use the pretest measure as a covariate? This is because error and unique variance in covariates, or independent variables can cause pseudo-effects (Cook & Campbell, 1979, Chapters 4 & 7; Campbell & Boruch, 1975). For example, if the treatment-pretest correlation were .50, the treatment-posttest correlation also .50, indicating no effect, and the pretest-posttest correlation .80, the partial correlation would be about .19, rather than zero, showing a pseudo-effect that could be highly significant with reasonable sample size. Other regression adjustments would show similar pseudo-effects. To avoid such biases, we must treat pretests and posttests similarly, not one as an independent variable, one as a dependent, with all of the effects of errors in both being thrown into the dependent variable.
However, LISREL measurement models would be acceptable if the pretest and posttest were constrained to the same model. The correlations among the treatments should make possible the detection of both main effects and interactions.

Consider the application to therapies for an HIV+ panel sample. The pretest and posttest could consist of CD-4 T-Cell counts or of a more accurate indicator of disease progression one year apart. The pharmaceutical treatments each panel member had tried out in between for the first time (scored as frequency and/or total strength of each) would be correlated with pretest and posttest, and changes in correlation interpreted as signs of causal impact (more clearly when an increased correlation on the posttest for a measure of positive health than when a decrease in a negative correlation).

For HIV/AIDS-Education programs, surveys one year apart on safer-sex practices, knowledge, or attitudes could provide pretest and posttest, to be correlated with exposure levels to various educational efforts. However, were information about these treatments collected in the same interview or questionnaire as the posttest, this provides an artifactual basis for higher correlation (see Campbell & Clayton, 1961; Campbell & Stanley, 1966, p. 67).

An independent survey of treatment exposure, midway between pretest and posttest might be ideal. Or the treatment exposure questions might be included on both pretest and posttest, in the "cross-lagged panel correlation" format (Cook & Campbell, 1979, pp. 309-321) but solved for "cross-lagged causal paths."'

The method could also be used for community attributes. Our federal monitoring system should monitor not only HIV+ levels, but also educational efforts, and (most expensively) educational levels as to HIV risks. If the distribution of the collection of educational efforts is both uneven, and datable, we might speculate on estimates of collective effects of all programs. Let's say we have a net educational-intensity score for a hundred regions between 1990 and 1992. The cross-sectional educational dosage profile will no doubt correlate with both pre-1990 and post-1992 HIV incidence rates, perhaps in a progressive manner, the most needy getting the higher dosage, or perhaps regressively. But if the educational dosage is effective, the correlation should change. (If the distribution is progressive, that is, if the 1990-1992 educational dosage correlated positively with the pre-1990 HIV incidence rates, then effectiveness would reduce the correlation of dosage and post-1992 HIV rates, perhaps moving it into the negative range. But many other factors can reduce such a correlation [Campbell, 1971, as revised for private distribution]). Most interpretable would be if the treatment pre-test correlation were near zero, and the treatment post-test correlation were significantly negative.

VII. Using Interview and Questionnaire Measures of Beliefs and Behaviors

Undoubtedly, if educational programs reduce rates of new HIV cases, it is because beliefs have been changed, and these have affected behaviors. While we are unlikely to measure preventive behaviors, we can get verbal reports on them. Our causal model is thus, in part, as shown in Figure 15.

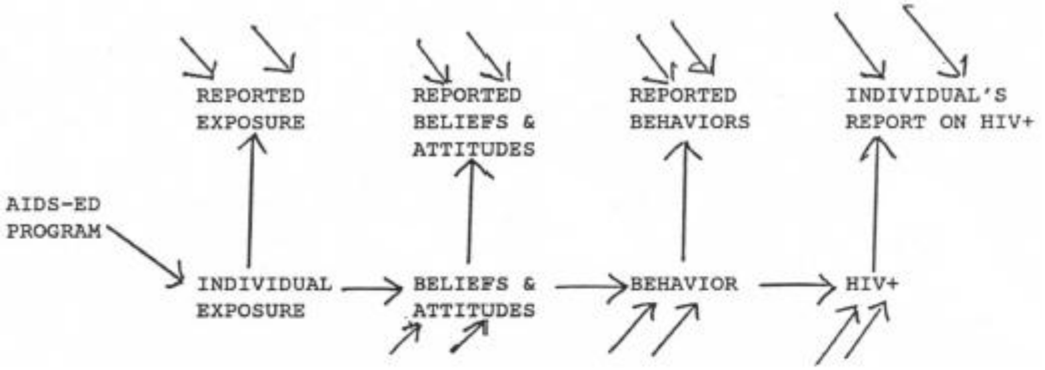


Figure 15. Causal Paths presumed in interview outcome measures.

The advertising community has found it easier to measure impact at the reported exposure level, avoiding using sales. In our case, the trace of the HIV/AIDS Education program might be most noticeable at Reported Beliefs, since people tend to forget sources of information.

Particularly for prompt measurement in pilot studies of alternative media and message content, the HIV Education community needs to use these proximal indicators obtainable by verbal reports on small samples of persons.

Should we go over our cafeteria of quasi-experimental design speculating about verbal reports? We could institute, perhaps, an annual optional anonymous questionnaire of high school seniors, and use it in the time series design, using different high schools to test different program content, etc. (The advertising industry uses studio audiences. Regularly assembling groups could be asked to accept HIV/AIDS Educational Programs).

We would rather speculate about random assignment designs making it possible to pinpoint question-answerer and exposure, and then speculate about likely imperfections of implementation and the quasi-experimental designs that result or suggest themselves.

For tryouts of alternative messages, we believe we need to maximize clarity of causal inference (and minimize costs) rather than try to represent the dilute dosage that would be typical once the program was in place. We also favor selecting samples for maximum clarity of causal inference, rather than representativeness per se. Here is a scenario, influenced by the already available research showing the impact of visual presentations (Turner, Miller, & Moses, 1989, Chapter 5). By telephone, we secure thousands of persons having VCR's and willing to accept by mail and critique trial HIV/AIDS Education programs. From this list, we select equivalent, blocked random samples, to use with Videotape A, Videotape B, printed alternatives, and controls who get nothing for two months. They are asked to send in critiques, and to return the cassette by the franked envelope provided. All these samples are then later interviewed (perhaps by mail, with anonymity except for treatment type code. The questionnaire also asks about HIV/AIDS Education exposure.) The treatment name-lists are then retained for later search in the pregnancy and other HIV test catchments, anonymity again being preserved.

Systematic collection of causal testimony

Quasi experiments range widely in the extent to which they approximate "true" experiments. This paper has oscillated on this dimension. Here is a suggestion from the remote end: Rather than inferring cause from the comparison of measures (treated vs. untreated, before treatment - after treatment) drawing inferences of effect without regard to whether or not the participants were aware of the treatment, its impact, or any change on their own part (or if aware of change would have explained it differently), this suggestion trusts people as competent observers and reporters of their own exposure to treatments and the resulting changes in their own beliefs, attitudes, and behavior.

There are hundreds of community workers involved in HIV/AIDS-Education and Drug-Abuse-Education, who are in effective conversation with persons in their own community. They listen as well as preach (or perhaps, instead of preaching). They hear some first-hand (and some convincing second-hand) reports on people who testify to having changed their behavior in HIV avoidant ways. These testimonies often include reports on why they have changed. These community workers are themselves engaged in trying to change behavior. They get told (sometimes at least) when their own efforts are foolish, ineffective, or counter-productive. They get told about some of the foolish, ineffective, or counter-productive HIV/AIDS-Education efforts put out by other agencies and media. They also, we hope, get testimony about the efficacy of some of their own educational efforts.

All of these informative, anecdotes are made use of by these community workers in what the Program Evaluation specialists call "Formative Evaluation," that is, in the tailoring of their own multi-faceted outreach program, guiding their own trial-and-error selection of components and their choice of new components to add. "Formative Evaluation" (in this aspect, at least) is informal "Impact Evaluation." Were these front-line street workers to prepare reports for others, their knowledge could be classified as "Participant Observer Research,"[3] "Ethnographic Program Evaluation," or "Qualitative Evaluation." Such reports can have the goal of impact analysis fully as much as do "experimental" studies.

Currently, such sources of impact estimation go unused for our overall evaluation of HIV/AIDS-Education programs, or for their design. It seems reasonable to explore ways in which this might be done. But before we do this, some warnings: These community educators and therapists are already overburdened with required paper work, quarterly reports, questionnaires, etc. Most of this paperwork is alienating. It is done (if and when done) with confidence that it will never be read or used, unless a bureaucratic excuse is needed to cut funding. Further funding is their most important goal in filling out these forms, both to keep their own livelihood and also (for most) to keep going local programs they feel are both needed and useful. Descriptive accuracy is a secondary motive. Success reports are certain to be exaggerated. Favorable anecdotes are sure to win out over unfavorable ones, if the latter reflect on the community worker's own program. Moreover, these community workers would much rather talk than write. This is a conjectural situational analysis. But certainly any effort to use these resources for HIV/AIDS-Educational program impact

[3] Indeed, it was participant observation, routinely collected during the AIDS Community Demonstration Project that led to Krauss's first NIMH project, the Parent/Preadolescent Training for HIV Prevention (PATH) mentioned above. While the Community AIDS Demonstration project in New York City was aimed at adult women who had been sexual partners of men they knew or suspected injected drugs, “reaction to intervention” interviews with community women consistently elicited statements such as: “It is too late for me. I have already taken risks. Help me save my children.”

estimation should review present reporting requirements and the reports so produced with such conjectures in mind.

First Author Notes

For the NIMH AIDS Survey Research Methodology Conference, July 11-12, 1991, Holiday Inn Crowne Plaza, Rockville, MD. (Revised from a paper with the same title, delivered at the NRC Conference on Nonexperimental Approaches to Evaluating AIDS Prevention Programs, Washington, DC, January 12-13, 1990, and cited in Coyle, Boruch, and Turner, 1991.)

This paper underwent multiple subsequent revisions, the first a revision of a rough draft discussed at the January, 1990, meeting. I am indebted to Heidi Brown, Beatrice Krauss, Ping Wu, and William Trochim for help in that version's preparation.

References

Association of the Bar of the City of New York & the Drug Abuse Council, Inc. (1973). *New York drug law evaluation project*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research (producer and distributor], 1980. doi:10.3886/ICPSR07656.v1)

Baldus, D. C. (1973). Welfare as a loan: An empirical study of the recovery of public assistance payments in the United States. *Stanford Law Review, 25*, 123-250.

Bentler, P. M. (1990, January). *Structural equation modeling and AIDS prevention research*. Presented at the NRC Conference on Nonexperimental Approaches to Evaluating AIDS Prevention Programs, Washington, DC.

Box, G. E. P. & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association. 70*, 70-92.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*, 409-429. Reprinted in E. S. Overman, (Ed.). (1988). *Methodology and epistemoloqy for social science* (pp. 261-289). Chicago, IL: University of Chicago Press.

Campbell, D. T. (1971). Temporal changes in treatment-effect correlations: A quasi-experimental model for institutional records and longitudinal studies. In G. V. Glass (Ed.), *Proceedings of the 1970 invitational conference on testing problems* (pp. 93-110). Princeton, NJ: Educational Testing Service.

Campbell, D. T. (1976). Focal local indicators for social program evaluation. *Social Indicators Research, 3*, 237-256.

Campbell, D. T. (1984). Hospital and landsting as continuously monitoring social polygrams: Advocacy and warning. In B. Cronholm & L. von Knorring (Eds.), *Evaluation of mental health services programs* (pp. 13-39). Stockholm: Forskningsraadet Medicinska.

Campbell, D. T. (1985). Quasi-experimental approaches in therapeutic research. *Muscle & Nerve, 8*, 483-485.

Campbell, D. T. (1987). Guidelines for monitoring the scientific competence of preventive intervention research centers: An exercise in the sociology of scientific validity. *Knowledge: Creation, Diffusion, Utilization, 8*, 389-430.

Campbell, D. T. (1990, January). *Quasi-experimental design in AIDS prevention research.* Presented at the NRC Conference on Nonexperimental Approaches to Evaluating AIDS Prevention Programs, Washington, DC.

Campbell, D. T. (1994). Systems theory and social experimentation. In S. A. Umpleby & V. N. Sadovsky (Eds.), *Reconstructing knowledge and action; Systems theory in the United States and the Soviet Union*. New York: Hemisphere Publishing Corporation.

Campbell, D. T. & Boruch, R. F. (1971). *Measurement and experimentation in social settings*. Unpublished project description for NSF Grant GS-30273X.

Campbell, D. T. & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A Bennett & A. A. Lumsdaine (Eds.). *Evaluation and experience; Some critical issues in assessing social programs* (pp. 195-296). New York: Academic Press.

Campbell, D. T., Boruch, R. F., Schwartz, R. D., & Steinberg, J. (1977). Confidentiality-preserving modes of access to files and to interfile exchange for useful statistical analysis. *Evaluation Quarterly. 1*, 269-300.

Campbell, D. T. & Clayton, K. N. (1961). Avoiding regression effects in panel studies of communication impact. *Studies in Public Communication, 3*, 99-118.

Campbell, D. T. & Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3*, 33-54.

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* (pp. 171-246). Chicago, IL: Rand McNally. Reprinted in 1966 as *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton-Mifflin.

The CDC AIDS Community Demonstration Projects Research Group. (1999). Community-level HIV intervention in five cities: Final outcome data from the CDC AIDS Community Demonstration Projects. *American Journal of Public Health, 89*, 336-345. PMCID: 1508588

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation; Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Coyle, S. L., Boruch, R. F., & Turner, C. F. (Eds.). (1991). *Evaluating AIDS prevention programs; Expanded edition*. Washington, DC: National Academy Press.

Fleiss, J. L. & Tanur, J. M. (1973). The analysis of covariance in psychopathology. In M. Hammer, K. Salzinger, & S. Sutton (Eds.). *Psychopathology; Contributions from the social, behavioral, and biological sciences* (pp. 509-527). New York: Wiley.

Folker, G. (2009). NIAID honors AIDS activist Martin Delaney. Accessed at *http://www.eurekalert.org/pub_releases/2009-01/nioa-nha012209.php* September 2012.

Goedert, J. J, & Blattner., W. A. (1988). The epidemiology and natural history of Human Immunodeficiency Virus. In V. T. DeVita, S. Hellman, & S. A. Rosenberg (Eds.), *AIDS: Etiology, diagnosis, treatment and prevention* (pp. 33-61). Philadelphia: Lippincott.

Kessler, R. C. (1993). Quasiexperimental design in AIDS psychosocial research. In D. G. Ostrow & R. C. Kessler (Eds.), *Methodological issues in AIDS behavioral research* (pp. 76–92). New York: Plenum.

Krauss, B. J., Goldsamt, L., Bula, E., Godfrey, C., Yee, D. S., & Palij, M. (2000). Pre-test assessment as a component of safer sex intervention: A pilot study of brief one-session interventions for women partners of male injection drug users in New York City. *Journal of Urban Health, 77,* 383-395. PMID: 10976612. Included as an effective program in the meta-analysis by Scott-Sheldon, L. A. J., & Johnson, B. T. (2006). Eroticizing creates safer sex: A research synthesis. *The Journal of Primary Prevention, 27,* 619-640.

Maini, M. K, Gilson, R. J., Chavda, N., Gill, S., Fakoya, A., Ross, E. J., Phillips, A. N., & Weller, I. V. (1996). Reference ranges and sources of variability of CD4 counts in HIV-seronegative women and men. *Genitourinary Medicine, 72,* 27-31.

McCleary, R. & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. Beverly Hills, CA: Sage Publications.

Moffitt, R. (1991). The use of selection modeling to evaluate AIDS interventions with observational data. In S. L. Coyle, R. F. Boruch, and C. F. Turner (Eds.), *Evaluating AIDS prevention programs; Expanded edition* (pp. 342-364). Washington, DC: National Academy Press.

Riecken, H. W. & Boruch, R. F. (Eds.) (1974). *Social experimentation*. New York: Academic Press.

Roos, N. P. (1973). Evaluating the Impact of Health Programs: Moving from Here to There. Unpublished paper. Department of Social and Preventive Medicine, University of Manitoba, Winnipeg, Canada, July.

Ross, H.L. (1973). Law, science and accidents: The British road safety act of 1967. *Journal of Legal Studies, 2* (1), 1-78.

Sax, P. E., Boswell, S. L., White-Guthro, M., & Hirsch, M. S. (1995). Potential clinical implications of interlaboratory variability in CD4 + T-lymphocyte counts of patients infected with Human Immunodeficiency Virus. *Clinical Infectious Diseases, 21*, 1121-1125.

Trochim, W. M. K. (1984). *Research design for program evaluation*. Beverly Hills, CA: Sage Publications.

Trochim, W. M. K. & Campbell, D. T. (1996). The regression point displacement design for evaluating community-based pilot programs and demonstration projects. Unpublished paper, analysis described at http://www/socialresearchmethods.net/kb/statrpd.php, Accessed September 2012.

Turner, C. F., Miller, H. G., & Moses, L. E. (Ed.). (1989). *AIDS; Sexual behavior and intravenous drug use*. Washington, DC: National Academy Press.

Volberding, P. (1990). Rationale for variations in clinical trial design in different HIV disease stages. *Journal of Acquired Immunodeficiency Syndromes, 3*, S40-S44.

Whittaker, G. F. (1974). *An economic and statistical evaluation of the performance of hospitals in a merged system: An application of a quasi-experimental design,* Northwestern University, Ph.D. Dissertation, Department of Accounting and Information Systems.

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

Zimring, F.E. (1975). Firearms and federal law: The Gun Control Act of 1968. *Journal of Legal Studies, 4*, 133-198.