



---

---

# Causal Inference and the Language of Experimentation

---

---

## INTRODUCTION

The major purpose of this book is to outline the experimental approach to causal research in field settings. We hope that our work will prove useful to persons interested in both theoretical and applied social research. The book is addressed to two contrasting audiences. One group consists of persons trained in laboratory research who want to conduct their current work in the real world where conditions are more difficult to control, and who want to make this transition with a minimal loss in the quality of causal inference and with a greater awareness of the particular ambiguities they will usually have to accept in interpreting the results of their field research projects. The second audience is of social scientists who are acquainted primarily with descriptive research where the investigator does not manipulate or intrude upon the processes being observed, who are aware of the dangers of inferring causation from passive observational data, and who are nonetheless interested in inferences about effects, benefits, influences and the like. Our hope is that such social scientists will learn from this book how to collect the kinds of data and perform the kinds of statistical analyses that will render causal inferences as sound as possible.

Since the book is largely about drawing causal inferences in field research, we obviously need to define cause. Part of this chapter is devoted to this topic. We shall deliberately adopt an outmoded position derived from Mill's inductivist canons, a modified version of Popper's falsificationism, and a functionalist analysis of why cause is important in human affairs. We will contrast our "critical-realist" position with some of the many other philosophical beliefs about causation.

In chapter 2 we shall introduce a set of terms for understanding some of the modifications to classical experimental designs that work in field settings necessitates. We have adopted many of the terms developed by Campbell (1957) and Campbell and Stanley (1963) in their systematic attempt to explicate the threats to valid inference that arise when the major features of laboratory research with humans are not present (e.g., random assignment, physical isolation of respon-

dents, short duration of the experimental treatment). Their efforts have been substantially extended in our present work.

In chapters 3 through 7 we describe a variety of forms that designs with non-comparable (i.e., nonrandomized) groups can assume, and present an outline of how the data collected within such design frameworks might be statistically analyzed. These chapters stress the assumptions that have to be accepted before responsible inferences about the causal impact of treatments can be drawn from particular designs and data analyses.

Even though most of the controls associated with the laboratory cannot—and should not—be created in field settings, classical designs based on random assignment can nevertheless sometimes be implemented in such settings. In chapter 8, we discuss a variety of factors which prevent the use of random assignment in field settings or which cause the randomization to break down before the study is completed. Strategies that may overcome these obstacles are stressed. Chapter 8 is concerned, therefore, with the implementation of randomized experimental designs in the field rather than with the form of such designs.

The present book is not intended to be a definitive treatise on field research. Many topics central to such a purpose will only be lightly touched upon, including: how to check on the importance of one's guiding research questions; how to physically sample respondents; how to construct and validate measures; how to collect qualitative data; how to present research findings. Nor is the book a comprehensive treatise on evaluation research. Although it may be of some relevance, we do not discuss such crucial evaluation issues as: how to incorporate into the evaluation the concerns of various constituencies with an interest in the eventual findings; how to discover whether the program is being implemented in anything like the promised form; how to use qualitative or quantitative knowledge about a program to assess whether much of the research resources should be devoted to answering questions about causal impact; how to measure and use in the analysis details about the heterogeneity of a program or project.

Instead, the book is intended to help persons who conduct both basic and applied research, who have already decided that they want a causal question answered, and who are prepared to look elsewhere for details about how to carry out field operations that, while significant for a study, may be less important in answering the causal questions posed in that study. In applied fields, such as evaluation, causal questions usually have an explicit context that at least specifies the population of persons who might be affected by a putative cause. For example, the sponsors and practitioners of research often ask causal questions about a specific kind of person who lives or works in a particular type of setting. Although we discuss sampling and measurement issues implicated by the context surrounding causal questions, we are more concerned with the experimental designs and statistical analyses that facilitate causal inference.

## THE LANGUAGE OF EXPERIMENTATION

The word *experiment* denotes a test, as when one experiments with getting up two hours earlier to see if this makes one's working day more productive. The test is usually of a causal proposition: for example, does garlic or curry add a better

flavor  
where  
instan  
faster  
T  
ment  
evalu  
garlic  
and  
of cc  
tions  
or cu  
are  
actu  
I  
prot  
proc  
ligh  
spa  
spa  
exa  
(19  
fea

flavor to certain rice dishes? There are some uses of the concept of experiment where the link with cause is not immediately obvious, yet still paramount. For instance, an airplane is "experimental" only if one wants to test whether it flies faster, more efficiently, or more safely than some alternative.

The notion of a "trial" or deliberate manipulation is also linked to experimenting. Actually getting up earlier on some mornings is the most direct way of evaluating how one's productivity changes; using curry on some occasions and garlic at others will enable one to evaluate which seasoning improves the rice dish; and without flying the experimental airplane, it will be difficult to test. There are, of course, both deliberate and unplanned trials based on simulating real manipulations and experiences, as when one tries to imagine the different tastes that garlic or curry might cause, or when one "pilots" the airplane in a simulator. But these are proto-trials, as it were, and more credibility is attributed to tests based on actual eating or flying than to simulated tests.

Deliberate trials have long been used to test causal propositions. Our ancestors probably did not start out with a plan to strike flints together until sparks were produced. At some point they probably observed an accidental fire caused by lightning or sparks, and they also noted that rubbing stone against stone caused sparks. From this, they could have developed and tested the hypothesis that such sparks combined with flammable material can produce a flame. A much later example of a deliberate trial is illustrated in the following story from Boring (1954), an example from seventeenth-century field research that is redolent with features of modern science.

In 1648 the Torricellian vacuum was known to physics in general and to Pascal in particular. This is the vacuum formed at the upper closed end of a tube which has first been filled with mercury and then inverted with its lower open end in a dish of mercury. The column of mercury falls in the tube until it is about 30 inches high and remains there, leaving a vacuum above it. Pascal was of the opinion that the column is supported by the weight of the air that presses upon the mercury in the dish (he was right; the Torricellian tube is a barometer) and that the column should be shorter at higher altitudes where the weight of the atmosphere would be less. So he asked his brother-in-law, Perier, who was at Clermont, to perform for him the obvious experiment at the Puy-de-Dôme, a mountain in the neighborhood about 3,000 feet ("500 fathoms") high as measured from the Convent at the bottom to the mountain's top. On Saturday, September 19, 1648, Perier, with three friends of the Clermont clergy and three laymen, two Torricellian tubes, two dishes and plenty of mercury, set out for the Puy-de-Dôme. At the foot they stopped at the Convent, set up both tubes, found the height of the column in each to be 26 old French inches plus 3 1/2 Paris lines (28.04 modern inches), left one tube set up at the Convent with Father Chastin to watch it so as to see whether it changed during the day, disassembled the other tube and carried it to the top of the mountain, 3,000 feet above the Convent and 4,800 feet above sea-level. There they set it up again and found to their excited pleasure that the height of the mercury column was only 23 French inches and 2 Paris lines (24.71 inches), much less than it was down below, just as Pascal had hoped it would be. To make sure, they took measurements in five places at the top, on one side and the other of the mountain top, inside a shelter and outside, but the column heights were all the same. Then they came down, stopping on the way

to take a measurement at an intermediate altitude, where the mercury column proved to be of intermediate height (26.65 inches). Back at the Convent, Father Chastin said that the other tube had not varied during the day, and then, setting up their second tube, the climbers found that it too again measured 26 inches 3 1/2 lines. These are reasonable determinations for these altitudes, showing about the usual one inch of change in the mercury column for every 1,000 feet of change in altitude.

In this experiment there was no elaborate design, and it took place 195 years too soon for the experimenters to have read John Stuart Mill's *Logic*, but the principle of control and of the Method of Difference is there. How important it was for them to have left a barometer at the base of the Puy-de-Dôme to make sure that changes in the tube that they carried up the mountain were due to elevation and not to general atmospheric changes or to other unknown circumstances! How wise of the party at the top to have made the measurement under as many different conditions as they could think of with altitude constant! How intelligent of them to take a reading on the way down and thus to turn the Method of Difference into the Method of Concomitant Variation!

Despite the creative use of experimental design features from the seventeenth century onward, it was not until the past century or so that experimental design notions became systematized. This systematization at first emphasized physical control of conditions—isolation, insulation, sterilization, strong steel chamber walls, soundproofing, lead shielding against Hertzian waves, and so forth. Much more recently, as biological research moved from the laboratory to the open field, the modern theory of experimental control through randomized assignment to treatment emerged. In agricultural work the emphasis is usually on whether a new practice or technique will increase the yield per acre. Note that, unlike Pascal's work in physics, this problem implies a particular single cause, the effects of which the researcher would like to evaluate. To do this, he or she creates different agricultural plots and deliberately assigns to each a different type of seed, fertilizer, method of raking, or whatever is under investigation. We shall refer to these possible causes as *treatments*, though the term *independent variables* could also have been used. We shall refer to possible effects of the treatment as *outcomes*, though the term *dependent variables* could also have been used and will occasionally be mentioned. Outcomes can, of course, be measured at many time intervals before, during, and after an experiment. As we shall see later, the scheduling of outcome measurement is one of the more important tools an experimenter has for detecting effects and for attributing them to the treatment.

To infer treatment effects, one needs a comparison. If the researcher applied fertilizer and then measured the yield, a number—produce per acre—would result. But we would not know if a larger or smaller number would have been obtained without the fertilizer. Many sources for comparison exist in experimentation—most have different purposes and not all are equally efficacious for any one purpose. The researcher, for example, could compare this year's yield in the experimental plot with last year's yield from the same plot; or he or she could compare this year's yield with that of some neighboring plot. The first of the comparisons would not be very useful, since crop yields depend on many factors (rainfall, sun, and so forth) which change from year to year. The second compari-

son is more useful; however, a neighboring plot may have a slightly different soil composition or be slightly more shaded. Either of these might account for observed differences in crop yield.

Although it would help to apply the old and new fertilizer to several plots, the number of plots in each treatment group may not by itself help causal inference. More important is the manner in which treatments are assigned. If all the plots treated with the new fertilizer are located on the southern side of an agricultural station (perhaps because these plots are nearer to where the new fertilizer is delivered), and if all the plots with the old fertilizer are on the northern side, then clearly one set of plots will get more sun than the other. Thus, any differences in crop yield between plots with one fertilizer rather than the other may be attributed to differences in sunlight and not to differences in the fertilizer used.

One of the great breakthroughs in experimental design was the realization that random assignment provided a means of comparing the yields of different treatments in a manner that ruled out most alternative interpretations. Random assignment requires experimental units, which can be plots of land in agriculture, individual persons in social psychology experiments, intact classrooms in education studies, and even neighborhoods in some criminal justice research. Treatments are then assigned to these units by some equivalent of a coin toss, a process of random selection which determines the treatment that each receives. Given a sufficient number of units relative to the variability between units, the random selection procedure will make the average unit in any one treatment group comparable to the average unit in any other treatment group before the treatments are applied.

In our hypothetical agricultural example, the plots of land to which the new fertilizer is to be applied would be dotted haphazardly around the part of the agricultural station set aside for this experiment and would be interspersed with the haphazardly arranged plots to which the old fertilizer is to be applied. When there is random assignment, any differences in yield observed at the close of the experimental period cannot be due to differences in the number of sunlight hours from plot to plot since the plots receiving one treatment are, *on the average*, identical to those receiving the other treatment. In addition, differences in yield cannot be due to differences in the composition of soil from plot to plot since the soil is comparable, *on the average*, in the new and old fertilizer plots. Of course, each individual plot remains different from any other, just as each human is different from every other in a social experiment. However, the average plot in each agricultural treatment group is initially comparable, just as in social experiments the average human in each treatment group is initially comparable. Random assignment is the great *ceteris paribus*—that is, other things being equal—of causal inference. Its dependence on having many units per group has the beneficial side effect of permitting multiple tests or replications since the basic experiment is, in a sense, recreated in all treatment plots. Indeed, the error terms that are used for testing treatment effects indirectly assess whether the findings can be replicated.

All experiments involve at least a treatment, an outcome measure, units of assignment, and some comparison from which change can be inferred and hopefully attributed to the treatment. *Randomized experiments* are characterized by the

use of initial random assignment for inferring treatment-caused change. It is more difficult to assign individuals or larger social groups to treatments at random than it is to assign agricultural plots. It is also more difficult to assign individuals to treatments at random in field settings than in laboratory settings. The field researcher is often a guest at the sites where he or she works while the laboratory researcher has almost total control over the setting and acts as the respondent's host. Such considerations imply that random assignment will be less frequent with humans than with objects and less frequent with humans in the field than in the laboratory.

Although the term was not coined until later, Stouffer (1950) and Campbell (1957) placed a special emphasis on quasi-experiments—experiments that have treatments, outcome measures, and experimental units, but do not use random assignment to create the comparisons from which treatment-caused change is inferred. Instead, the comparisons depend on nonequivalent groups that differ from each other in many ways other than the presence of a treatment whose effects are being tested. The task confronting persons who try to interpret the results from quasi-experiments is basically one of separating the effects of a treatment from those due to the initial noncomparability between the average units in each treatment group; only the effects of the treatment are of research interest. To achieve this separation of effects, the researcher has to explicate the specific threats to valid causal inference that random assignment rules out and then in some way deal with these threats. In a sense, quasi-experiments require making explicit the irrelevant causal forces hidden within the *ceteris paribus* of random assignment.

Several distinctions are traditionally made among types of quasi-experiments. Nonequivalent group designs are typically those in which responses of a treatment group and a comparison group are measured before and after a treatment. This would be the case where two school classes are compared to each other and measures, perhaps of achievement, are collected at the beginning and end of the school year. Interrupted time-series designs are those in which the effects of a treatment are inferred from comparing measures of performance taken at many time intervals before a treatment with measures taken at many intervals afterwards. For example, attendance at school might be observed every day for a year and then every day for the next year following a new school policy about attendance. As we shall see in chapter 5, many interrupted time-series designs are improved (i.e., frequently occurring alternative interpretations are ruled out) by combining the longitudinal component of time series with the cross-sectional comparability of nonequivalent group designs.

The term correlational-design occurs in older methodological literature, most often to refer to efforts at causal inference based on measures taken all at one time, with differential levels of both effects and exposures to presumed causes, being measured as they occur naturally, without any experimental intervention. We find the term *correlational* misleading since the mode of statistical analysis is not the crucial issue. We discuss such methods in chapter 7 using the term, "Passive Observational Methods" to replace Correlational Methods.

The reduced possibilities for control available in field settings led to the development of the theory of quasi-experiments and to a refined specification of the controls needed if random assignment has not been achieved. Lack of control has

