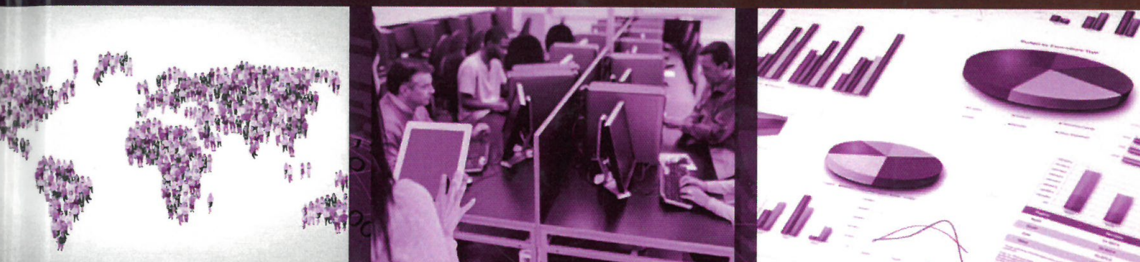
Laboratory Experiments in the Social Sciences

Second Edition



Edited by

Murray Webster, Jr. and Jane Sell

Chapter 3

Logical and Philosophical Foundations of Experimental Research in the Social Sciences

Shane R. Thye
University of South Carolina, Columbia, South Carolina

I INTRODUCTION

Social scientists of various academic stripes depict the experimental method in diverse ways—sometimes good, mostly bad. Lieberson (1985, p. 228) asserts that "the experimental simulation fails at present in social science research because we are continuously making counterfactual conditional statements that have outrageously weak grounds." Psychologists, who normally embrace experimental methods, have made even more radical claims: "The dissimilarity between the life situation and the laboratory situation is so marked that the laboratory experiment really tells us *nothing*" (Harré & Secord, 1972, p. 51, italics in original). Even writers of popular methods textbooks make critical observations. Babbie (1989, p. 232) argues that "the greatest weakness of laboratory experiments lies in their artificiality." There are numerous costs and benefits associated with any research method, but for whatever reason, the experimental method seems to inspire its fair share of critics.

If you are reading this, you are probably a student who has been assigned this chapter as part of a research methods course, or perhaps you are just interested in the logic and philosophical underpinnings of experimental methods. You may even, intuitively, believe that some or all of the preceding critiques are valid. After all, it has been pointed out for decades that experiments *are* artificial, and indeed they are (Webster & Kervin, 1971). The term *artificial* usually means "novel" in the sense that subjects are placed into situations (e.g., a lab) that are unusual or perhaps unnatural. But when you stop to think about it, don't all research methods involve somewhat novel, unusual, or artificial conditions? What is so natural about filling out a survey or being probed through an in-depth interview? We do not routinely do these things. The experimental situation is just one kind of situation that people respond to, no more or less novel than

Laboratory Experiments in the Social Sciences. http://dx.doi.org/10.1016/B978-0-12-404681-8.00003-0

your very first visit to New York City, going on a first date with someone you have met online, or giving birth to your first child. Novelty in and of itself is a part of everyday life and does not really distinguish experimentation from other popular research methods.

Henshel (1980) correctly pointed out more than three decades ago that *unnatural* experimentation is desirable and, in fact, the dominant mode in fields such as medicine, chemistry, and physics. I can do no better than he did, so I will use his example. Consider the modern airplane. We have all seen footage of early attempts at flight using awkward, wing-flapping devices pushed off the edge of cliffs only to plunge to the ground. They never flew. Those inventors attempted to mimic nature—that is, the way that birds obtain lift by flapping their wings—and they all failed. Some even went so far as to imitate the layered structure of a bird's secondary wing feathers on their flying machines. It did not help. It was not until the Wright brothers began to experiment with a fixed-wing aircraft that flight was finally obtained. Nowhere in nature does a fixed-wing aircraft exist, but because of unnatural experimentation, we enjoy the convenience of modern air travel.[1] Some of the most important conveniences in contemporary life are the result of unnatural experimentation, from the X-rays that diagnose disease to the silicone chips that power your smartphone. My hope is that by the end of this chapter, I can show you that the novelty and artificiality of the experimental method are enormous benefits, not liabilities (see also Webster & Kervin (1971) for a statement as to why).

Perhaps the most important feature of controlled experimental research is that it gives one the ability to claim, with some degree of confidence, that two factors are causally linked. Social scientists are in the business of establishing causal laws that explain real-world events. This objective is realized through the painstaking building and systematic testing of scientific theory. As a sidebar, to say that one has an explanation for some phenomenon is to say that there is a well-supported scientific theory of that phenomenon. For example, if you "explain" why skydivers descend to the earth by invoking the notion of gravity, what you really mean is that the observed descent conforms to the theory of gravitational forces. To date, controlled experimentation is the most widely embraced method for establishing scientific theory because it allows scientists to pinpoint cause–effect relations and eliminate alternative explanations. Laboratory experimentation is the gold standard for isolating causation because the logic of experimental research embodies the logic of scientific inquiry. *Thus, the advantage of the experimental method is that it allows one to see the world in terms of causal relations.*

Ironically, this also is a heavy burden carried by the experimental scientist. Consider a headline that read, "Bottled Water Linked to Healthier Babies." If it is true that expectant mothers who drink bottled water tend to have babies with

1. It is true that a soaring bird does not flap its wings, but the appropriate analogy for a soaring bird is an unpowered glider. Neither has the capacity to obtain lift like a powered fixed-wing aircraft.

fewer birth defects, higher birth weights, and other health benefits, this simple yet powerful intervention may carry major health consequences. However, my excitement over the power of water was short-lived as my newfound "parent" identity had to be reconciled with the "experimentalist" in me who also had an opinion. During that transformation, a number of questions emerged. What is the *theoretical connection* between bottled water and healthy babies? Is there a biological mechanism, or could the effect be due to other factors? The latter issue concerns whether or not the relationship is real. So-called spurious factors are unrecognized causes (e.g., socioeconomic status) that produce the illusion that two things (drinking bottled water and having healthy babies) are causally linked. Could it be that socioeconomic status causes both? Perhaps moms who can *afford* bottled water have healthier babies because they can *afford* to get better prenatal care, join gyms, eat healthier food, take vitamins, receive treatment at elite hospitals, and so on. For the same reason, I would lay heavy odds that moms who drive new BMWs also have healthier babies than moms who do not own a car, but that would be unlikely to grab the headlines.

In what follows, I examine the underlying logic of experimental design and analysis. I consider how experimental research bears on establishing of causation, explore recent critiques of Fisherian (Fisher, 1935, 1956) methods and show how these are flawed, and examine various forms of experimental design. The aim is not to provide a comprehensive discussion of all facets of experimentation. Instead, I hope to illuminate the logic and philosophical underpinnings of experimental research, dispel a number of myths and misconceptions, and generally excite the reader about the prospects of building scientific theory via experimentation. Along the way, I consider a number of issues germane to all research, such as how evidence bears on theory, notions of causation, and the logic of applying or generalizing findings to other settings.

II CLUES TO CAUSATION

The notion of "causality" has always been a challenge, in part, because causation is not directly observable. Rather, causation must be inferred from some manner of evidence. This section considers the various ways that scientists think about causality and the methods scientists use to infer that two phenomena are causally linked. The concept of causality has a twisting and convoluted history in the philosophy of science, and there are many ongoing discussions and debates. Most notions of causation are traced to Aristotle (340 BC/1947), who offered four different conceptions of causation. Of these, Aristotle's *efficient cause* captures the notion that one event (X) sets into motion, forces, creates, or makes another event (Y) occur. Although this kind of causation corresponds well with the everyday meaning of the term, it has become the focal point of controversy and debate.

Galileo developed an alternative causal ontology that equated causation with necessary and sufficient conditions. Galileo argued that to say event *a* caused event *b* was to say that event *a* is a necessary and sufficient condition for event *b* (see Bunge, 1979, p. 33). A *necessary condition* exists when event *b never* occurs in the absence of event *a*. That is, event *b* follows event *a* with 100% regularity. A *sufficient condition* exists when event *b always* follows event *a* with perfect regularity. To illustrate, a lawnmower will only start if it has the correct kind of fuel in its fuel tank. Thus, the proper kind of fuel is a necessary condition for starting the mower because it will *never* start without it. At the same time, fuel alone is not enough to start the mower; it requires a working engine and all requisite parts. This means that proper fuel is not a sufficient condition to make a mower start. The combination of the proper fuel and a working engine are necessary and sufficient conditions; when both are in place, the mower *always* starts, and if either is missing the mower *never* starts.

The philosopher David Hume (1748/1955) took the more radical position associated with British empiricism. Prior to Hume, popular notions of causation traced to Aristotle involved one event forcing, or setting into motion, another event. Hume offers a softer notion that robs causation of its force. He argued that we can never directly observe a causal force in operation, but instead, all we can observe is the conjunction or correlation of two events that we presume are causally linked. Hume used the term "constant conjunction" to describe the situation in which X always occurs in the presence of Y. Thus, for Hume and the empiricists, causation is an elusive thing—we can never be sure that one thing causes another or that events correlated today will be correlated tomorrow. More radical Humeans, such as Bertrand Russell, reject the idea of causation altogether. In his now infamous 1913 paper, Russell wrote, "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (Russell, 1913).

Despite the historical disagreements surrounding notions of causation, the majority of social scientists generally agree on certain basic requirements that must be satisfied to support causal inference (Davis, 1985; Hage & Meeker, 1988). One can think of these requisites as "clues" to assess if a relation is truly causal or not. Next, I consider six conditions that scientists use to assess causation. This discussion focuses not on points of disagreement but, rather, on the general principles on which there is consensus.

A Covariation

When a cause occurs, then so should its effect; when a cause does not occur, then neither should its effect (this is one of the basic maxims underlying Mill's canons of inference discussed later). In short, causes and their effects should covary or be correlated. At the same time, it is very important to remember that things which covary may not be causally related. For instance, the price of bourbon is correlated with the price of new cars in any given month: when bourbon is expensive, cars are expensive. In this case, the prices of bourbon and

new cars are not causally linked, but both are caused by the prevailing economic conditions. As Hume would agree, it is easy to focus on the correlation between car prices and bourbon while blindly missing the underlying causal force. As such, this leads to a very important principle: *correlation alone does not imply causation*. Sifting causation from correlation is a focal issue in virtually all social science research.[2]

B Contiguity

There is always some time lag between a cause and its effect. When the time lag between a cause (a paper cut) and its effect (bleeding) is short, we say the two events are contiguous. Some cause–effect relations are contiguous, whereas others are not (e.g., conception and childbirth). In general, social scientists presume that noncontiguous causes set into motion other processes that, in turn, have effects at a later point in time. For example, greater parental education can lead to a variety of lifestyle benefits (e.g., greater income, more social capital, and advanced reading and verbal skills) that have effects on their children's educational attainment. Importantly, however, a causal claim between parent and child education levels is not warranted unless there is a theory that specifies the intermediary factors occurring between the two points in time. For example, more educated parents are more likely to read to and with their children than are uneducated parents, and the greater reading by those children helps them succeed in school.

C Time and Asymmetry

One of the most basic principles of causation is that if X causes Y, then X must occur before Y in time. It is important to note that simply because X precedes Y does not mean X causes Y. Just because it rained before I failed my exam does not mean the rain caused me to fail (if I do make the connection between rain and failure, I am committing the *post hoc fallacy*). A related idea is that causation is assumed to run in one direction, such that causes have asymmetric effects. That is, we cannot simultaneously assert that X causes Y and Y causes X. Also, although the notion of *reciprocal causation* $X_1 \rightarrow Y \rightarrow X_2$ seems to violate the assumptions of time and asymmetry, it does not because the X at time 2 is not the same X as the X at time 1. Instead, reciprocal causation implies that cause–effect sequences dynamically unfold. For instance, cybernetic feedback systems involve reciprocal causation in which the cause–effect relationships are reversed through time. The "cruise control" feature in your car operates on this

2. The basic experimental design provides an elegant solution to the problem of sorting causal relations from correlations. Because this problem has an even tighter stranglehold on those who deploy survey, historical, qualitative, or ethnographic methods, researchers in these domains often use the experiment as a template to design their own studies (Lieberson, 1991).

principle. That is, an initial cause (the deceleration of your car) triggers an effect (an increase in engine RPM) that in turn feeds back on that initial cause (the acceleration of your car).

D Nonspuriousness

When two things occur together but are truly caused by some third force, the original relationship is said to be *spurious*. There are many unusual spurious relations that illustrate the point. For instance, few people know that there is a strong positive correlation between ice cream sales and rape. That is, in months when ice cream sales are high, many rapes are reported. Does this mean that ice cream sales are somehow causally linked to the occurrence of rapes? The answer is of course not! The relationship is spurious because both factors are caused by a third factor—temperature. In warmer months, more ice cream is sold and there are more sexual predators frequenting outdoor venues and social gatherings. In cold weather, both factors are attenuated. It would be a mistake to believe that rapes and ice cream sales are casually related without seeking alternative explanations. To establish causation, one must be able to, with some degree of confidence, rule out alternative explanations. Ruling out alternatives is a key activity in science, and as demonstrated later, experimental research provides the best known method for doing so.

E Consistency

Philosophers and scientists have long debated whether one should think of causality as *deterministic* (i.e., that a given cause X will always lead to effect Y) or *probabilistic* (i.e., that the presence of cause X will increase the likelihood of effect Y). Early writers such as Galileo (1636/1954) and Mill (1872/1973) leaned toward the deterministic end of the spectrum. However, with the advent of modern statistical tools, the majority of contemporary philosophers and social scientists evoke probabilistic notions of causality. To illustrate, there is abundant evidence that smoking causes a higher rate of lung cancer. Still, not *every* smoker develops lung cancer, and not *every* lung cancer victim is a smoker. It seems more accurate to say that smoking increases the probability of lung cancer. Hage and Meeker (1988) argue that probabilistic notions of cause are preferable for three reasons. First, there can be unrecognized countervailing causal forces (e.g., antibodies or gene combinations that make people resilient to cancer) that play a role. Second, most phenomena are affected by a multiplicity of causes (e.g., body chemistry or exposure to carcinogens) that can interact to obfuscate true causal relations. Third, chaos and complexity theories have shown that both natural and social phenomena can behave in unpredictable and nonlinear ways (Waldrop, 1992). To illustrate, those with advanced cancer can for some inexplicable reason go into remission. The upshot is that the deterministic views of causation may be overly simplistic.

F Theoretical Plausibility

Finally, scientists always view causal claims about the world with one eye trained on established scientific laws. When claims about the world violate or are inconsistent with those laws, then without unequivocal evidence to the contrary, the confirmation status of those claims is questionable. Simply stated, extraordinary claims require extraordinary evidence. For instance, Newton's law of inertia states that unless acted upon, a body at rest stays at rest and a body in motion stays in motion. Based on this thinking, many have attempted to build "perpetual motion" machines ignoring the broader context of other physical laws. A true perpetual motion machine (if one could be built) would, in fact, violate several existing scientific laws. For example, such a machine would need to consume no energy and run with perfect efficiency (thus violating the second law of thermodynamics) or produce energy without consuming energy as it runs (thus violating the first law of thermodynamics). Thus, although Newton's law of inertia *suggests* that perpetual motion machines are possible, in the context of the laws of thermodynamics, it seems that such machines are very unlikely to ever be produced. It should not be surprising that, to date, dreams for perpetual motion machines have remained just that.

III MILL'S CANONS AND INFERRING CAUSALITY

What is the logical connection between data and causation? Asked differently, how does one infer causation based on the outcome of some empirical test? Many, if not all, social scientists operate by approximating a model of evidence that can be traced to John Stuart Mill (1872/1973), who was influenced by Sir Francis Bacon and subsequently influenced Sir Ronald A. Fisher (1935, 1956). Mill presumed that nature is uniform, and as such, if a cause–effect relationship occurred once, he presumed it would occur again in similar circumstances. Mill developed five methods (or canons) to assess causation; these canons lie at the heart of contemporary inference and experimental design today. Here, I briefly discuss two of the more important canons: the method of difference and the method of agreement. Whereas the method of difference aims to find a lone *difference* across two circumstances, the method of agreement seeks to find a lone similarity.

> *The method of difference*: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indisputable part of the cause, of the phenomenon" (1872/1973, p. 391).
>
> *The method of agreement*: "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon" (1872/1973, p. 390).

Perhaps the best way to illustrate the methods is by example. Imagine that you and a friend attend a wild game cookout, sample a variety dishes, and later that night you feel ill but your friend does not. How would you determine the cause of the illness? Imagine that you and your friend both sampled shrimp, pheasant, and duck, while you also enjoyed oysters but your friend did not. The top panel of Figure 3.1 illustrates this scenario. Notice that there are two differences between you and your friend: you dined on oysters and then became ill

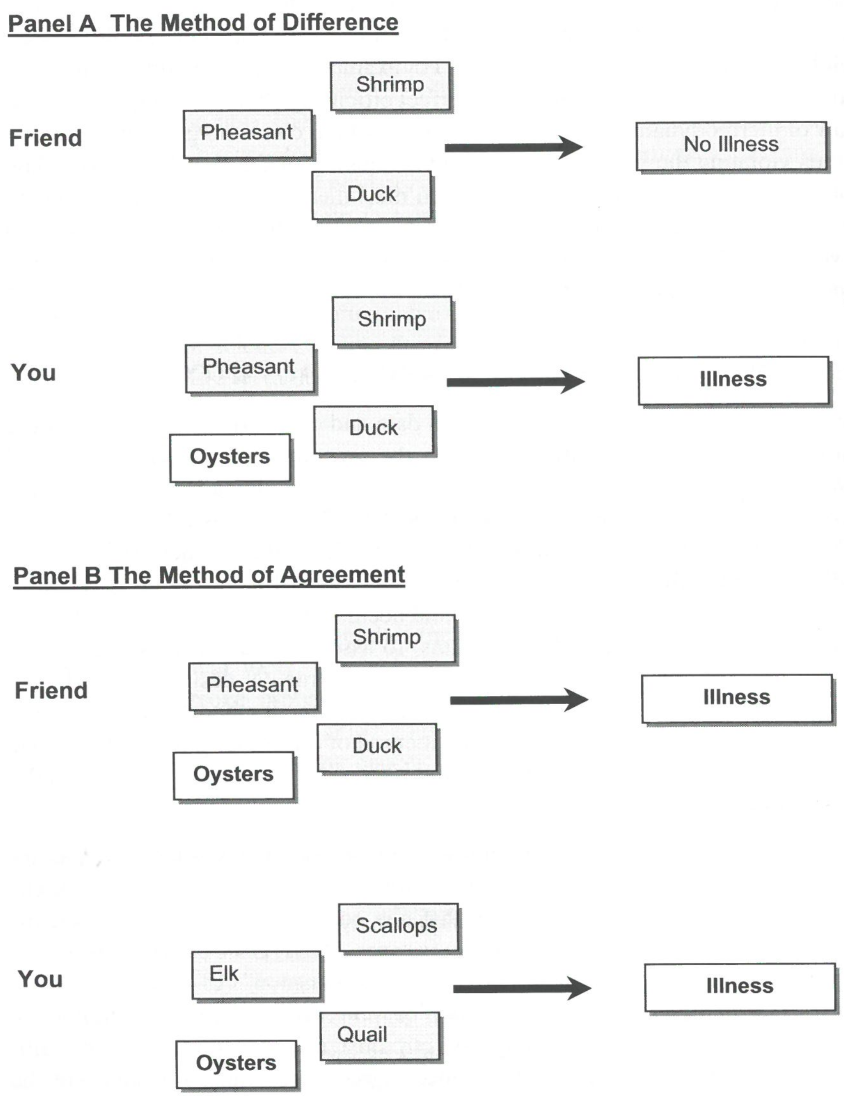


FIGURE 3.1 Mill's method of difference and method of agreement.

while your friend did neither. Mill's method of difference suggests that oysters are the cause of the illness because it is the single condition that distinguishes illness from health. The same method can also be used as a method of elimination. Notice that both you and your friend ate shrimp, pheasant, and duck, but only you got ill. Thus, shrimp, pheasant, and duck can be eliminated from the list of possible causes. In this way, the method can be used to eliminate alternative causes because *a deterministic cause that remains constant can never produce effects that are different.*

Now consider the method of agreement, which indicates that if a single factor (eating oysters) occurs in conjunction with a common effect, then that single factor is the likely cause. Assume that your friend enjoyed shrimp, pheasant, duck, and oysters, while you ate scallops, elk, quail, and oysters. Later that evening, you both became ill. Because eating oysters is the common denominator preceding illness, you might infer that oysters are the cause. This illustrates another important principle: *a causal factor that differs across two circumstances can never generate precisely the same effect.* In this case, if eating shrimp were the true cause of illness, because your friend ate shrimp and you did not, we would expect your friend (but not you) to be ill. The method of agreement once again suggests that the oysters are suspect.

A Limitations of Mill's Canons

Mill's (1872/1973) method of difference and method of agreement provide useful guidelines for thinking about causation. Even so, it is now recognized that the canons are limited in a number of regards. A good critique of the methods can be found in Cohen and Nagel (1934; see also B. Cohen, 1989), and it may be useful to summarize their analyses here. Cohen and Nagel (1934, p. 249) point out that Mill's methods are neither "methods of proof" nor "methods of discovery" (see also Hempel, 1965). In terms of proof, the canons generally presume that *all other possible causes* are contained in the factors examined. In the previous example, we presume that all possible causes of illness reside in the food that was eaten. However, the methods cannot *prove* this to be definitively so because there *could* be unrecognized causes in things besides the food. Imagine that you and your friend have an undetected mild shellfish allergy that only reacts when two or more varieties of shellfish are eaten simultaneously. Referring back to Figure 3.1, in every case of illness, both oysters and one other shellfish (shrimp or scallops) were consumed. As such, it might not be oysters per se causing the illness but, rather, the unique combination of food and the allergy—what statisticians call an *interaction effect*. The method of difference and method of agreement are blind to this possibility.

Second, Mill's (1872/1973) methods cannot be used as "methods of discovery" because they presuppose that one can identify, *a priori*, potential causes of the illness. In Figure 3.1, differences and similarities across food items are analyzed as potential causes for the illness. However, there are other potential causes that are unmeasured or unknown. This includes personal factors (food consumed before the party), social factors (contact with other sick people), historical factors (flu or allergy season), or even genetic or biological factors (a weak immune system) that might fuel illness in a difficult to detect way. In principle, there are an *infinite* number of additional causes that could contribute to illness, and the lion's share of these will be unmeasured and unknown. Again, the methods are not methods of discovery because they are oblivious to possible alternatives.

In summary, Mill's canons suffer two limitations: *proving* that a single factor is the unique cause and *discovering* causes beyond the immediate situation. Although it may not be immediately obvious, both limitations stem from a single issue: *any phenomenon may have an infinite number of intertwined causes that researchers cannot measure or identify.* Given this seemingly insurmountable problem, one might guess that Mill's (1872/1973) methods would have withered on the vine. Indeed, that might have occurred had it not been for the statistician R. A. Fisher,[3] who rescued the methods. Fisher (1935, 1956) did that by providing a logical and methodological basis for reducing or altogether eliminating the troublesome set of alternative explanations. Next, I illustrate Fisher's solution and other central features of experimental research through a detailed example. Following this, I reconsider Mill's (1872/1973) methods in view of Fisher's solution and Cohen and Nagel's (1934) critique.

IV FISHER'S SOLUTION AND HALLMARKS OF EXPERIMENTATION

The confluence of three features makes experimental research unique in scientific inquiry: random assignment, manipulation, and controlled measurement. To illustrate, presume that a researcher is intrigued by the relationship between viewing television violence and childhood aggression, a topic that captures much scrutiny. A typical hypothesis is that viewing television violence increases subsequent aggression by imitation. Correspondingly, let us assume that a researcher plans to study 100 fourth-grade children, their viewing habits, and their aggressive behaviors. The overall experimental strategy involves three distinct phases. First, the researcher must *randomly assign* each child to one of two groups. The first group will be exposed to violent television (called the treatment group) and the second group will be exposed to nonviolent television (called the control group). Second, the researcher *manipulates* the content of the violent television such that one program contains violence while the other does not. Third, the researcher *measures* aggressive behavior across the groups in a controlled environment. Let us consider each feature in more detail.

3. Despite the tremendous vision Fisher possessed as a statistical prodigy and brilliant geneticist, he was, ironically, severely myopic. He claimed that this aided, rather than hindered, his creative insight because he was forced to rely more heavily on mental instead of physical representations.

A Random Assignment

Fisher (1935, 1956) offered the concept of random assignment as a way to hold constant spurious causes. *Random assignment* is defined as the placement of objects (people or things) into the conditions of an experiment such that each object has exactly the same probability of being exposed to each condition. Thus, in the two condition experiment described previously, each child has exactly a 50% chance of being assigned to the experimental or control group. This procedure ensures that each group contains approximately 50 children, half male and half female, to within the limits of random chance. The benefit of random assignment is that it *equates at the group level.* In other words, random assignment ensures that the two groups are equal in terms of all *historical factors* (e.g., divorced parents, childhood poverty, and being abused), *genetic factors* (eye color, chromosome distribution, blood type, etc.), *physical factors* (gender, height, weight, strength, etc.), *personality factors* (preferences, values, phobias, etc.), and *social factors* (role identities, dog ownership, etc.). The reason is that each factor has exactly the same probability of appearing in each group. The brilliance of random assignment is that it mathematically equates the groups on factors that are known, unknown, measured, or unmeasured.[4]

A common criticism is that experimental results are biased if the traits under the control of random assignment interact with the dependent variable. This idea, however, is slightly off the mark. For instance, assume that: (1) both males and females respond aggressively, but differently, to watching violent television; and (2) our measure of aggression is sensitive to male, but not female, aggression. The result of the experiment would correctly show that males in the treatment group respond more aggressively than males in the control group *on this measure of aggression*. It would also correctly show no difference between the treatment and control females *on this measure of aggression*. Strictly speaking, the problem here is not one of random assignment or even one of incorrect inference. Instead, the problem is that one may not have a robust and valid measure of hostility that captures the kind of aggression expected to occur in both males and females. The relationship between exposure to violent television and aggressive behavior is properly guided by the theory that links the two phenomena. Such an interaction suggests a problem with the theory or a problem with the measurement procedures, but not the experiment per se.

B Manipulation

Following random assignment, the independent variable is manipulated such that the treatment group is exposed to violent images while the control group is not.

4. There are deep-rooted statistical reasons to employ random assignment. For instance, violating random assignment can cause observations to be correlated, which can bias any ensuing statistical test, inflate the standard error of that test, or both. Thus, random assignment rests on logical and statistical foundations.

Ideally, the manipulation would be exactly the same in both conditions except for the factor hypothesized as causal. In our example, the researcher could ask the experimental group to watch a television program of a couple engaged in a financial dispute that ends violently. The control group could watch the exact same couple end the exact same dispute in a nonviolent manner. Notice that in the context of random assignment, the basic experiment is comparable to Mill's method of difference in that the two groups are equated on virtually all factors except the independent variable. Apodictically, the Fisherian principle of random assignment is the cornerstone of experimental research, and in conjunction with controlled manipulation, these procedures render the method of difference workable.

C Controlled Measurement

The final step is to measure the dependent variable (aggressive behavior) in a controlled environment. Ideally, individuals from the treatment and control group would have the opportunity to aggress in exactly the same manner toward the same target. In actual research, aggression has been measured in a variety of ways, including the delivery of electric shock, hitting a doll with a mallet, or the slamming down of a telephone. Importantly, any difference between the treatment and control groups can only be attributed to the independent variable because, in principle, this is the only factor on which the two groups differ (remember that a cause that does not vary can have no effect). Overall, the experimental method is the method of causal inference because it equates two or more groups on all factors, manipulates a single presumed cause, and systematically records the unique effect.

How does the method measure up with respect to providing information on causation? Recall that causal inferences require information on three empirical cues (covariation, contiguity, and temporal ordering) and three other criteria (nonspuriousness, consistency, and theoretical plausibility). Overall, the basic experiment does an excellent job of attending to these matters. Evidence for covariance comes in the form of the effect appearing in the experimental group but not the control group. Also, the researcher has information on contiguity because he or she controls the time lag between the factors under investigation. Furthermore, because the experimenter manipulates the cause before the effect, the temporal ordering is correctly instated. Of course, random assignment controls for potentially spurious factors. Finally, the experiment is a repeatable event, and its data are always considered in the context of established theory.

V FISHER'S PREMATURE BURIAL AND POSTHUMOUS RESURRECTION

The basic experiment approximates Mill's (1872/1973) method of difference in structure and design. However, recall that Mill's method was deemed intractable

because there could be an infinite number of unknown and unknowable causes adding to (or interacting with) the presumed cause. A related problem is that Mill invokes a deterministic view of causation wherein empirical outcomes occur with perfect regularity, which of course never occurs (Willer & Walker, 2007). Fisher (1935) developed the principle of random assignment to remedy the ailing method. When subjects are randomly assigned to conditions, any unknown and unknowable factors are distributed equally and thus: (1) *the causal effects of those unknown and unknowable factors will be equated across experimental and control groups*; and (2) *factors that are the same can never cause a difference.* Thus, any differences can be reasonably attributed to variation from the independent variable. Although Fisher's principle of random assignment is widely recognized as the panacea, others still believe the method is grievously ill.

Bernard Cohen (1989, 1997) asserts that random assignment does not remedy the problem. He argues that as the set of unknown or unknowable causes grows large, the probability that the experimental and control groups will differ on at least one of these causal factors approaches 1.0. This relationship is shown in Figure 3.2. Furthermore, Cohen notes that random assignment only equates experimental and control groups with infinite sample sizes, and of course, real experiments always employ finite samples. He therefore concludes that there is always some unknown causal factor operating in any experiment, and as such, the original problems that plagued Mill's (1872/1973) methods remain unresolved. In the end, Cohen (1997) does the only befitting thing and offers "a decent burial" for J. S. Mill and R. A. Fisher.

At first blush, Cohen (1989, 1997) appears to have brought the illness out of remission. However, he may have been premature in laying our progenitors

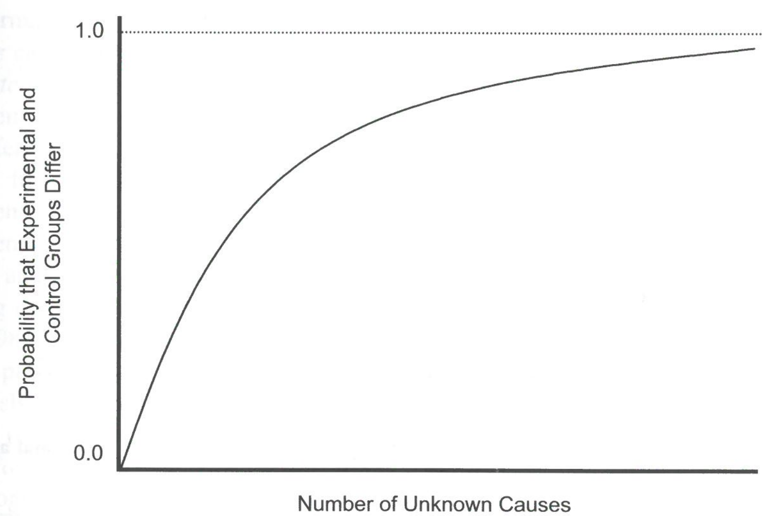


FIGURE 3.2 The relationship between unknown causes and probability that experimental and control groups differ.

and their methods to rest. His analysis centers on: (1) the number of potential alternative causes; and (2) the number of subjects in the experimental and control groups—claiming that the former is too large and the latter is too small for random assignment to properly operate. However, in a mathematical sense, Cohen (1997) overestimates both the number of alternative causes that *do* exist and the sample size that *is* required for random assignment to work properly. In terms of sample size, Cohen (1997) ignores the straightforward statistical relation between sample size and the nature of alternative causes. That relationship is described by the *law of large numbers*. This law dictates that as the size of the experimental and control groups increases, the average value of any factor differentiating those groups approaches a common value (i.e., the population value for that factor). This law also has implications for the *number* of alternative causes. The law implies that as group size increases and alternative causal factors become equated, the number of *differentiating possible causes will diminish at an exponential rate*. As this occurs, the probability that the experimental group differs from the control group, on any single factor, approaches zero.[5] The relationship between sample size and probability of differentiating factors is shown in Figure 3.3. Thus, probability theory and the law of large numbers mitigates (but does not totally eliminate) the issue of unknown factors.

There are additional logical and empirical grounds that further salvage the method. Logically, it is important to distinguish factors that *could make*

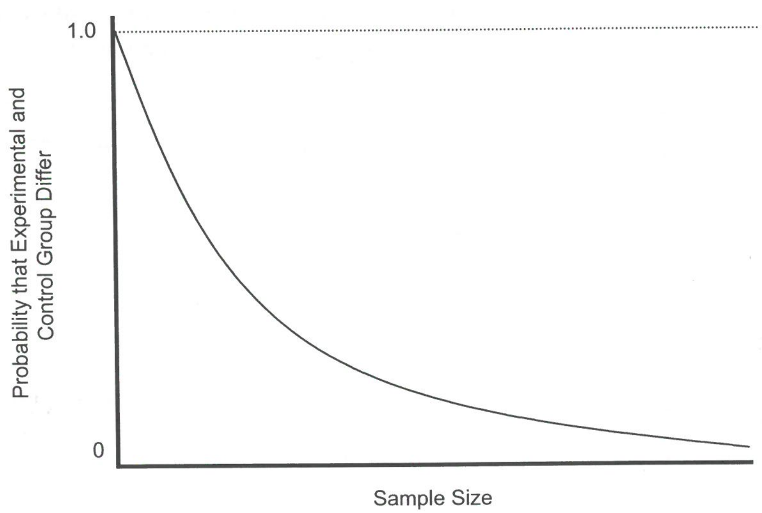


FIGURE 3.3 The relationship between sample size and probability that experimental and control groups differ.

5. For instance, with just 23 randomly selected people, there is a 50% chance 2 of them will share a birthday (Paulos, 1988)!

a difference from factors that *do make a difference*. Not all of the factors that make experimental groups different from control groups are relevant to the dependent variable, and therefore, not all factors must necessarily be equated. Many differences simply do not matter. For instance, physical theories do not consider the color of falling objects for the same reason that bargaining theories do not consider the height of the negotiator; neither factor is relevant to the theoretical processes and phenomena of concern (Willer & Walker, 2007). Scientists specifically control for *theoretically relevant* factors in a given study, and in true experiments they also control for *possibly relevant* factors using random assignment. Cohen heavily emphasizes those factors that are possibly relevant. However, given that laboratory research selectively focuses on specific theoretical problems, most of the infinite set of possibly relevant differences probably will be irrelevant.

Now, let's play devil's advocate and presume Cohen (1997) is correct in his assertion. That is, assume that a very large number of factors differentiate experimental and control groups and that these factors are relevant to the phenomena under consideration. From a statistical standpoint, ironically, even this situation does not pose a problem for the internal validity of the experiment. The *central limit theorem* explains how those factors would be distributed. This theorem suggests that for a very large number of independent causal factors differentiating experimental and control groups, the aggregate effect of those causal factors would quickly converge and become normally distributed. The shape of the distribution is important. Some causal factors would have positive effects on the dependent variable; others would have negative effects. When subjects are randomly assigned to conditions, the overall impact of these factors is to add *random variance* (or noise) to the dependent variable, and that noise would be normally distributed with an expected value of 0. Thus, *random assignment and the central limit theorem ensure the errors are normally distributed and guarantee there is no overall impact of these variables.* This means that even if the identified problems are real, they do not affect the basic logic of experimental inference or impact the validity of the method.

In summary, the method of difference and the principle of random assignment provide the logical and statistical foundation for contemporary experimental design and analysis. Mill and Fisher left in their wake a powerful set of analytic and statistical tools that have become the method of understanding causality in science. I have shown that the problems identified by Cohen (1989, 1997) and others are not problematic in a statistical or pragmatic sense. In practice, if the experimental method were flawed in the ways detailed previously, we would never expect to find consistent results produced by experimental research, nor would we expect to find cumulative theory growth in areas so informed. However, even Cohen (1989) acknowledges that cumulative research programs informed by experimental work abound. Despite the fact that both Mill and Fisher have long since passed away, the legacy of experimental testing they left behind continues to flourish.