SEVEN

Elaborate Theories

> [We should] trust rather to the multitude and variety of . . . arguments than to the conclusiveness of any one. [Our] reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected.
>
> —CHARLES SANDERS PEIRCE[1]

> All human errors are impatience, the premature breaking off of what is methodical.
>
> —FRANZ KAFKA[2]

What Are "Elaborate Theories"?

Cochran's Discussion of Fisher's Advice

In an observational study, in a study of treatment effects without random assignment of treatments, an association between treatment received and observed outcome is ambiguous: it could reflect an effect caused by the treatment, or an unmeasured bias in the way treatments were assigned, or a combination of the two. In what ways can this ambiguity be reduced?

Sir Ronald Fisher invented randomized experimentation, and William G. Cochran first presented observational studies as a topic in statistics defined by its contrast with randomized experiments. How did they perceive the problem of recognizing effects actually caused by treatments when treatments are not randomly assigned? Cochran wrote,

First, as regards planning. About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold . . . This multi-phasic attack is one of the most potent weapons in observational studies. In particular, the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others . . . The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis . . . [The investigator] cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance.[3]

Suppose that a causal theory makes several predictions—that it is elaborate. Then there are several associations predicted by the theory, and each can be checked against data. Suppose that a skeptic challenges one of these associations, the skeptic's claim being that the association is produced by a particular bias in who is treated, not by any effect caused by the treatment. A second association predicted by the causal theory cannot be explained by the skeptic's first postulated bias, the skeptic must justify continued skepticism by postulating a second bias to explain the second association, and so on. The weighing of evidence for or against a causal theory may become subtler as there is more evidence to weigh.

Let us consider a quick, simple example of the use of an elaborate theory in an observational study, before returning to the general topic.

A Simple Example: Does a Parent's Occupation Put Children at Risk?

Adults are often exposed to occupational hazards at work, and a variety of laws regulate these exposures to ensure the safety of workers. Might a parent

be exposed to an occupational hazard in such a way that his or her child i also exposed, even if the child never enters the workplace?

David Morton and colleagues asked whether parents who work in an in dustry using lead might bring lead home in their clothes and hair, thereb exposing their children.[4] They measured the level of lead in the blood of 3 children with a parent who worked in a battery manufacturing plant i Oklahoma that used lead in the production of batteries. As it turned ou all these parents were fathers. Morton and colleagues found control childre whose parents did not work in the battery plant. They paired each treate child to a control child from a different household whose age differed by a most one year and who lived close by in the same neighborhood. If a treate child lived in a home facing a major road, the control child was selected fro the same side of the same road. If the treated child lived in an apartmer complex, the control child came from the same complex. This matching wa intended to ensure that treated and control children faced similar levels c environmental lead at home, say, from automobile exhaust or nearby indus trial pollution. For both groups, they measured levels of lead in the children blood, recorded in micrograms of lead per deciliter (μg/dl) of blood. As write in 2016, the U.S. Centers for Disease Control and Prevention says, "Ex perts now use a reference level of 5 micrograms per deciliter to identif children with blood lead levels that are much higher than most children levels."[5] At the time of Morton and colleagues' study, in 1982, a higher leve was often used, 30 μg/dl.

Additional information was collected. First, some workers held jobs i the battery plant constantly exposing them to lead, whereas others ha limited exposure. Using information from the plant manager, the father exposure to lead was graded high, medium, or low depending upon th specific job the father performed. There were 19 fathers with high exposure 7 with medium exposure, and 8 with low exposure. Second, an interviewe questioned homemakers about occupational hygiene, and on that basis th workers were graded as having good, moderately good, or poor hygiene. Fo instance, a lead worker had good hygiene if he showered, shampooed, an changed shoes and clothes at work before going home, whereas changin clothes without showering was considered moderately good. As one migh expect, fathers with little exposure to lead at work rarely showered an changed before going home, so we will look at the hygiene for 19 father with high exposure, where 13 had poor hygiene, 3 had moderately good hy

giene, and 3 had good hygiene. For plotting purposes, the $6 = 3 + 3$ fathers with moderately good or good hygiene are combined into a single group called OK hygiene.

What is the elaborate theory? If a father's exposure to lead has an effect on the lead level of his children, then we expect (i) higher levels of lead in the blood of treated children than in matched control children, (ii) higher levels of lead in the blood of children whose fathers have higher exposure to lead at the battery plant, and (iii) higher blood levels if a high-exposure father practices poor hygiene. Additionally, (iv) the lead exposure of the father of a treated child should not predict the blood lead level of the control child to which the treated child is paired. One way prediction (iv) could fail is if high-exposure fathers live in a poor neighborhood near the battery plant, and people who live near the battery plant are exposed to air pollution from the plant, and matched control children from the same neighborhood are also exposed to the same air pollution.

The results will be displayed using boxplots, a widely used graphical display invented by John W. Tukey.[6] Figure 7.1 illustrates a boxplot with fictional data about a variable *Y*. A boxplot has a central box, vertical lines that continue up and down from the box, and may contain individual points. The central horizontal line in a boxplot is placed at the median, the middle value in sorted data, so half the values of *Y* are above this median line and half are below. The upper horizontal line in the box is at the upper quartile, so one quarter of the values of *Y* are above the upper horizontal line, and three quarters are below. In parallel, the lower horizontal line in the box is at the lower quartile, so one quarter of the values of *Y* are below the lower horizontal line, and three quarters are above. Saying the same thing differently: a quarter of the *Y*'s are below the box, a quarter are in the lower part of the box, a quarter are in the upper part of the box, and a quarter are above the box. Points judged extreme by a certain conventional standard are labeled individually, and there are five such points in Figure 7.1: four at the top, and one at the bottom. The vertical lines extend upward and downward from the box to the last value of *Y* that is not judged extreme.

As Tukey emphasized, boxplots convey quite a bit of information about a variable *Y*. The median line shows the typical value of *Y*. The box outlined by the upper and lower quartiles indicates a range that includes a central half of the *Y*'s, thereby indicating how much the *Y*'s typically vary. The rest of the plot tells us about atypical values of *Y*.

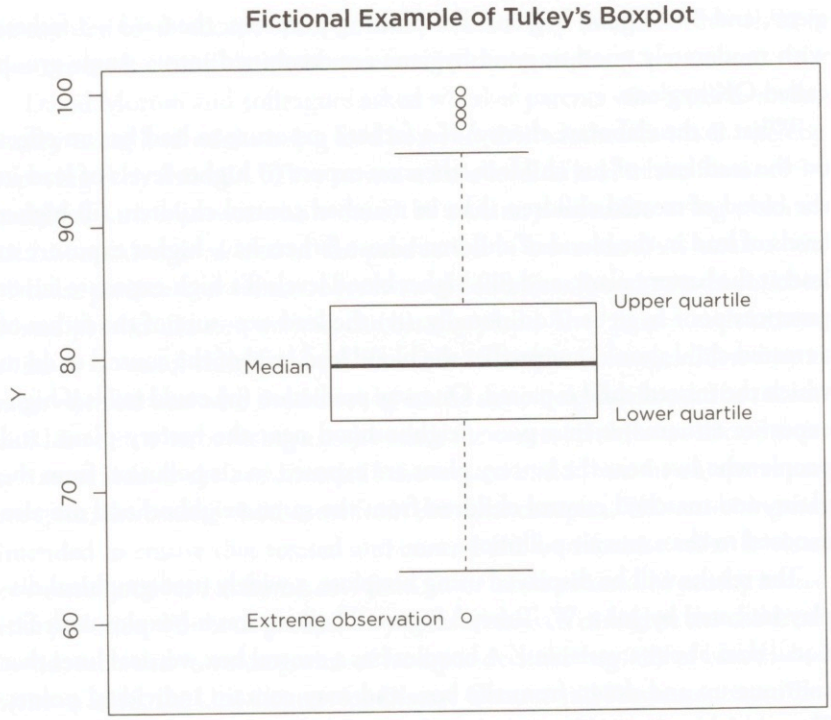

Figure 7.1. A boxplot for a fictional variable *Y*. Half of the values of *Y* are above the median (the middle line in the box), and half are below. A quarter of the values of *Y* are above the upper quartile (the upper line in the box), and three quarters are below. A quarter of the values of *Y* are below the lower quartile (the lower line in the box), and three quarters are above. Individual values of *Y* judged to be extreme by a certain standard are plotted as individual points.

Figure 7.2 checks the elaborate theory against the data. In Figure 7.2, a treated child is one whose father works in the battery plant, whereas a matched control child is of similar age and lives in the same neighborhood. Panel (a) on the upper left of Figure 7.2 compares the lead levels in the blood of treated children and matched control children. The blood lead levels are much lower for control children. Panel (c) on the lower left of Figure 7.2 focuses on the treated children, distinguishing among them on the basis of their fathers' levels of exposure to lead at the battery plant. If a father has high exposure, his child is more likely to have a higher level of lead in the blood. The level of exposure to lead of the treated child does not predict the

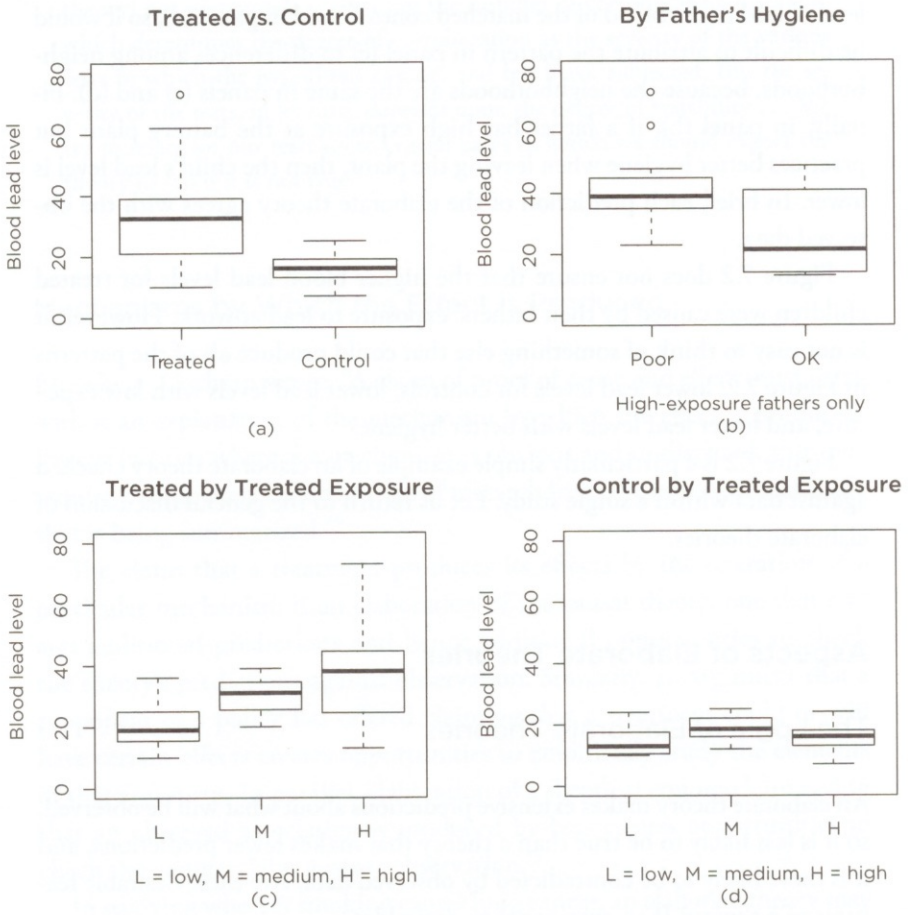

Figure 7.2. Checking an elaborate theory. Lead levels in the blood of children, μg/dl. Panel (a) compares children whose fathers worked in the battery plant to matched control children of similar age from the same neighborhood. Panel (c) separates treated children based on their fathers' levels of exposure to lead at the battery plant. Panel (d) separates control children based on the level of exposure of the father of their pair-matched treated child. Panel (b) looks only at children of fathers with high exposure to lead, separating them on the basis of the father's hygiene before leaving the factory, either poor or OK, where the OK group merges three fathers with good hygiene and three fathers with moderately good hygiene.

level of lead in the blood of the matched control child in panel (d), so it would be difficult to attribute the pattern in panel (c) to differences among neighborhoods, because the neighborhoods are the same in panels (c) and (d). Finally, in panel (b), if a father has high exposure at the battery plant but practices better hygiene when leaving the plant, then the child's lead level is lower. In brief, each prediction of the elaborate theory agrees with the observed data.

Figure 7.2 does not ensure that the higher blood lead levels for treated children were caused by their fathers' exposure to lead at work. However, it is not easy to think of something else that could produce all of the patterns in Figure 7.2: lower lead levels for controls, lower lead levels with low exposure, and lower lead levels with better hygiene.

Figure 7.2 is a particularly simple example of an elaborate theory checked against data within a single study. Let us return to the general discussion of elaborate theories.

Aspects of Elaborate Theories

The Logic of Elaborate Theories

An elaborate theory makes extensive predictions about what will be observed so it is less likely to be true than a theory that makes fewer predictions, and it is more likely to be contradicted by observed data. Are these desirable features of a theory? If so, why are they desirable?

Philosophers of science have often argued that they are desirable features of a scientific theory. In an essay, "On Selecting Hypotheses," Charles Sanders Peirce wrote, "But if I had the choice between two hypotheses . . . I should prefer . . . [the one which] would predict more, and could be put more thoroughly to the test . . . It is a very grave mistake to attach much importance to the antecedent likelihood of hypotheses . . . Every hypothesis should be put to the test by forcing it to make verifiable predictions."[7] Similarly, Karl Popper wrote,

> Theories may be more, or less, severely testable; that is to say, more, or less, easily falsifiable. The degree of their testability is of significance for the selection of theories . . . We should try to assess what tests, what trials, [the

> theory] has withstood . . . It is not the number of corroborating instances which determines the degree of corroboration as the severity of the various tests to which the hypothesis can be, and has been, subjected. But the severity of the tests, in its turn, depends upon the degree of testability . . . We try to select for our tests those crucial cases in which we should expect the theory to fail if it is not true.[8]

Mechanisms by Which the Effect Is Produced

Elsewhere, Cochran wrote, "A claim of proof of cause and effect must carry with it an explanation of the mechanism by which the effect is produced. Except in cases where the mechanism is obvious and undisputed, this may require a completely different type of research from the observational study that is being summarized."[9]

The claim that a treatment produces its effects by the operation of a particular mechanism is an elaboration of the causal theory, one that creates additional predictions and hence additional opportunities to check the theory's predictions against observation. Similarly, an argument that a proponent of a policy has offered claiming that a treatment could or will have certain effects creates opportunities to empirically study the elements of that argument. In parallel, elaboration of a skeptical counterclaim saying that an observed association is produced by bias creates opportunities to check the counterclaim against observation.[10]

In studying whether smoking causes lung cancer, an elaborate theory may predict that (i) smokers will develop lung cancer more often than nonsmokers in observational studies of people,[11] (ii) laboratory animals experimentally exposed to tars in cigarettes will develop cancer,[12] and (iii) the autopsied lungs of smokers who died of something other than lung cancer will exhibit cellular damage similar to that of individuals who died of lung cancer and unlike the lungs of nonsmokers.[13] These very different types of research each have weaknesses—smoking is not randomized, and mice are not people—so each comparison may be unconvincing on its own, but agreement between studies with very different weakness may be compelling.

Quantitative observational studies of many people may complement narrative, ethnographic, or qualitative studies of a few people.[14] A covariate that is not measured in a large quantitative study may be available for the

asking in an ethnographic study. Here, again, the weaknesses of one type of investigation may differ from those of another. Agreement among studies with very different weaknesses gradually makes it more difficult to attribute a shared conclusion to bias produced by their weaknesses.

How to Elaborate a Causal Theory

> What goes on in science is not that we try to have theories that accommodate our experiences; it's closer that we try to have experiences that adjudicate among our theories.
>
> —JERRY FODOR[15]

In thinking about how best to elaborate a causal theory, what considerations are important? The quote from Cochran at the beginning of the chapter makes a key suggestion: with an elaborate theory, "the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others." Suppose that a causal theory has met with some challenge, or is evidently open to some challenge—that is, some reasonably specific counterclaim explaining how the observed association between treatment and outcome is produced by a biased comparison, not an effect caused by the treatment. We are especially interested in an elaboration of the causal theory such that the causal theory and this counterclaim make different predictions about something we can observe. In this way, the elaboration of the causal theory helps to adjudicate between the causal theory and a specific counterclaim.

Let us consider an example.

Effects on Crime Rates of Restricting Handgun Purchases

An interesting example is from a study by Garren Wintemute, Mona Wright, Christiana Drake, and James Beaumont of the possible effects that restrictions on handgun purchases might have on crime rates.[16] Table 7.1 is adapted from their table 2. U.S. federal law restricts the purchase of a handgun by people who have been convicted of a felony. In 1991, the state of California went further, prohibiting for 10 years the purchase of a handgun by a person

Table 7.1. Crime rates in California, 1989–1991, among people who tried to purchase a handgun following a conviction for a violent misdemeanor

		Events per 100 person-years of observation		
Handgun purchase	*Number of people*	*Any crime*	*Gun or violent crime*	*Nonviolent crime with no gun*
Denied	927	14.1	8.0	9.3
Approved	727	15.5	9.9	8.6

convicted of certain violent misdemeanors, including assault, resisting arrest, and brandishing a firearm. Wintemute and colleagues compared two groups of people who would have qualified in 1991 for this restriction on handgun purchases. One group had attempted to make a handgun purchase in 1991, and because of the new restriction the purchase was denied. The second group attempted to purchase a handgun in 1989 or 1990, and because the new restriction was not in effect, the purchase was approved. In other words, the two groups, denied and approved, were the same in the sense that the change in law changed their ability to legally purchase a handgun, but they differed in when they attempted to make the purchase. In a table not unlike Table 1.1 for the ProCESS Trial, Wintemute and colleagues compared these two groups in terms of measured covariates, arguing that the groups were similar in terms of gender, age, race, and ethnicity, number of prior convictions, and number of convictions for gun or violent crimes. They then compared outcomes—namely, arrests for crimes committed after the purchase attempt.

There is a technical issue in Table 7.1 that I will mention but not discuss in detail. People were observed for different periods of time. If Harry was observed for a longer period of time than Sally, then Harry is more likely than Sally to be observed to commit a crime, just because we kept an eye on him for a longer time. We do not want this trivial issue to be confused with any possible effects of the change in law. For this reason, Wintemute and colleagues used several technical tools to address unequal periods of observation, the most elementary of these being evident in Table 7.1. Specifically, arrest rates are not per person but per person-year of observation. If Harry is observed for two years, he contributes two person-years to the study. If Sally is observed for one year, she contributes one-person year to the study.

If Harry is arrested twice in his two years and Sally is arrested once in her one year, then the arrest rates for Harry and Sally are the same, one arrest per person-year. The rates in Table 7.1 are high: a rate of 15 arrests per 100 person-years for Harry means, roughly speaking, a 15% chance Harry is arrested each year. Remember that the denied and approved groups both consist of people who had been convicted previously of certain violent misdemeanors.

The comparison of denied and approved groups suffers from one evident defect: the denials all occurred in a later year than the approvals. If criminal activity shifted greatly over the period from 1989 to 1991, then that shift might be confused with an effect of the change in law. For instance, changes in the unemployment rate or the activities of the police from 1989 to 1991 might affect whether a person is inclined to commit a crime. If Harry and Sally are out of work and short of cash in a particular year because of the sour economic situation in that year, then they might be more inclined to commit a crime that yields cash. Perhaps the ups and downs in criminal activity reflect the changing economic situation, not the change in law restricting handguns. We want an elaboration of the causal theory so that a sour economy predicts one thing will happen, but an effect of handgun restrictions predicts something else will happen.

The causal theory that handgun restrictions affect crime rates permits an obvious elaboration: restrictions on handguns should reduce specifically crimes for which possession of a handgun is relevant. If Harry is inclined to demand someone's wallet at gunpoint, then not being in possession of a gun at that delicate moment might prove inconvenient. This is less of an inconvenience for Sally, who is inclined to pick someone's pocket. A rise in the unemployment rate in a particular year might increase the rate of crimes committed for monetary gain, but that tendency seems unlikely to be restricted to crimes for which a gun is relevant. In the elaborate causal theory, an effect of the change in law restricting access to guns predicts a different pattern of associations than does a sour economic situation.

Table 7.1 and other analyses by Wintemute and colleagues show a lower rate of arrests for gun and/or violent crime in the group whose gun purchase was denied, but not a lower rate of nonviolent crime without a gun. The visible pattern in Table 7.1 is easier to explain as an effect of restrictions on gun purchases, harder to explain in terms of a sour economy. The orig-

inal study should be consulted for further analyses that involve, as I have mentioned, some additional technical detail.

Could Table 7.1 still be produced by a biased comparison? Yes, but not by a bias that affects crimes of all kinds in a similar way. If the police had decided in 1991 to crack down on violent gun crime and ignore pickpockets, then that shift might depress violent gun crime in the denied group without depressing nonviolent crime in the denied group, consistent with Table 7.1, even if there were no effect of the restrictions on handgun purchases. Of course, it would be easy to check in other ways whether such a shift in police behavior took place.

An elaborate theory is most valuable if it helps to adjudicate between a treatment effect and some plausible counterclaim denying such an effect.

Elaborate Theories and Tests of Ignorable Treatment Assignment

Recall the situation discussed in Chapter 5. In Chapter 5, you have a treated and a control condition, $Z_i = 1$ or $Z_i = 0$, an observed outcome, R_i, that equals $R_i = r_{\mathrm{T}i}$ if $Z_i = 1$ or $R_i = r_{\mathrm{C}i}$ if $Z_i = 0$, and an observed covariate, x_i. In this situation, you would have what you need for causal inference if treatment assignment were ignorable, but you have no way to check whether indeed treatment assignment is ignorable. How can you check whether treatment assignment is ignorable?

An elaborate theory changes this situation. With such a theory, there are things you could observe that would force you to abandon either the elaborate theory or the claim that treatment assignment is ignorable. With firm commitment to an elaborate theory, you can test ignorable treatment assignment. Moreover, framed as a statistical test of hypotheses, it is possible to ask about the properties of the test, for instance, whether the test is likely to detect failures of ignorable assignment when particular types of bias are present.[17]

The Crossword Analogy

Cochran's quoted discussion of elaborate theories spoke of weighing evidence from "a heterogeneous collection of results of varying quality, each bearing

on some consequence of the causal hypothesis . . . some [of which] are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance." Here, there are many strands of evidence of uneven quality, none strong enough to be compelling on its own, with some strands intersecting in ways that provide mutual support while others clash in ways that bar the emergence of a coherent picture. So a process is needed that appraises mutual support among studies in the presence of conflict among studies, a process that sets aside bits of evidence to see whether a strongly supported, coherent picture emerges from what remains.

In a related context, the philosopher Susan Haack suggested an analogy between the development of scientific knowledge and the solving of a crossword puzzle:

> The model is not . . . how one determines the soundness or otherwise of a mathematical proof; it is, rather, how one determines the reasonableness or otherwise of entries in a crossword puzzle. This model is more hospitable to a graduational account . . . The crossword model permits pervasive mutual support, rather than, like the model of a mathematical proof, encouraging an essentially one-directional conception . . . How reasonable one's confidence is that a certain entry in a crossword is correct depends on: how much support is given to this entry by the clue and any intersecting entries that have already been filled in; how reasonable, independently of the entry in question, one's confidence is that those other already filled-in entries are correct; and how many of the intersecting entries have been filled in.[18]

Haack is making two points. First, much of the conviction that a penciled-in crossword puzzle is correct stems from appropriate intersections of entries, rather than the strength of the individual clues that support individual entries. In parallel, in science, conviction often results from the appropriate intersection or agreement of several or many inconclusive studies with different vulnerabilities. The number of studies, their sample sizes, and their levels of statistical significance are not the critical issues. The critical issue is whether the vulnerability that makes one study doubtful is absent from another study. Observational studies of people, experimental studies of laboratory animals, and experimental studies of interactions among biomolecules are each vulnerable to error in determining the effects of treatments on people, but their vulnerabilities are very different.

Second and more subtly, Haack wishes to exhibit the possibility of mutual support without vicious circularity. Suppose I can deduce A from B, and I can also deduce B from A; then, to believe A and B are both true on that basis alone would be to err by vicious circularity, because logically A and B could both be false. Two entries in a crossword may meet appropriately yet both be incorrect. In the quote above, Haack asks, "How reasonable, independently of the entry in question, one's confidence is that those other already filled-in entries are correct"? Haack suggests that mutual support among entries in a crossword is possible without vicious circularity by a process of demarcation. Demarcation means setting aside a specific part of the evidence when appraising another part.[19] When using the evidence supporting A and its appropriate intersection with B in appraising the evidence available for B, it is appropriate to leave aside the evidence that B provides for A. In parallel, when using the evidence supporting B and its appropriate intersection with A in appraising the evidence available for A, it is appropriate to leave aside the evidence that A provides for B. In a crossword puzzle in which 1-down intersects 3-across, we may ask what evidence we have for our tentative solution to 1-down apart from its appropriate intersection with 3-across, and we may ask what evidence we have for our tentative solution to 3-across apart from its appropriate intersection with 1-down. We might conclude that the only compelling evidence for our tentative solution to 1-down comes from its appropriate intersection with 3-across, and the only compelling evidence for our tentative solution to 3-across comes from its appropriate intersection with 1-down; then, we would regard these two solutions as very tentative. Alternatively, we might conclude that our tentative solution to 1-down is supported by strong evidence apart from its appropriate intersection with 3-across, and that 3-across is supported by strong evidence apart from its appropriate intersection with 1-down; then, the appropriate intersection of these two entries provides additional support that they are each correct.

In appraising evidence from several observational studies, or from one study with several comparisons, we might ask: What evidence is available in support of a particular conclusion apart from those studies that suffer from a particular vulnerability? We might ask this question repeatedly, for different vulnerabilities.

The studies of fetal alcohol syndrome provide an example.

Fetal Alcohol Syndrome

> Make ready a feast of the princes. There it is your pleasure to eat the roast flesh, to drink as much as you please the cups of the wine that is sweet as honey.
>
> —HOMER, *The Iliad*[20]

The *Iliad* and intoxication are older than the written word, but the substantial effects of alcohol on the developing human fetus received systematic investigation in the latter half of the twentieth century.[21] At high, steady doses, the effects of alcohol on human fetal development are difficult to miss. The initial reports were case series of children born to severely alcoholic mothers with no comparison group.[22] As one case-series began, "Eight unrelated children of three different ethnic groups, all born to mothers who were chronic alcoholics, have a similar pattern of craniofacial, limb, and cardiovascular defects associated with prenatal-onset growth deficiency and developmental delay." The greatest harms were done to developing brains, but we are exquisitely aware of faces, and photographs of the children's faces revealed something terribly wrong.

The case series were not convincing on their own. As Carrie Randall wrote,

> Despite a shower of case reports, the implication of alcohol as a teratogen in humans was met with skepticism by the medical community. Alcohol was being used at the time to prevent premature labor . . . , so it was difficult to accept the proposal that it could cause harm to the developing fetus. Furthermore, because alcohol was so widely used, it was reasoned that if a causal relationship between prenatal alcohol exposure and birth defects existed, it would have been recognized and reported long before 1973.[23]

These case series were followed by studies with controls.[24] For instance, Beatrice Larroque and colleagues interviewed mothers at a French maternity hospital concerning their alcohol use during pregnancy, then examined their children when they were about 4.5 years old.[25] They compared moderately low and moderately high levels of alcohol consumption, finding that children whose mothers consumed the equivalent of four or more glasses of wine per day had substantially lower performance in a variety of cognitive

assessments. There was also evidence that mothers who drank more alcohol were different from those who drank less: the heavier drinkers had less education, were older, and were more often cigarette smokers; that is, there was evidence of bias in the comparisons, perhaps quite consequential bias. Larroque and colleagues made efforts to remove these biases analytically, in the spirit of the stratification in Table 5.1, but employed other methods. Many subsequent studies also found substantial cognitive deficits among children with prenatal exposures to alcohol, but not all studies concurred.[26] It is, however, no small concern that mothers who drink heavily during pregnancy may differ from those who abstain, differing perhaps in ways that have not been recorded and thus differing in ways that cannot be controlled by analytical adjustments.

Experiments were conducted with laboratory animals. Pregnant mice were given two doses of alcohol by injection into the blood, and later fetuses were found to exhibit facial malformations eerily similar to those found in human children exposed to high levels of alcohol.[27] When seven-day-old rats were given two doses of either a saline solution or alcohol, the brains of rats treated with alcohol "revealed a very dense and widely distributed pattern of neurodegeneration" that "deletes large numbers of neurons from several major regions of the developing brain."[28] Rhesus monkeys given prenatal exposures to alcohol exhibited cognitive deficits when compared with controls.[29] There are, of course, hazards in extrapolating from experiments on animals to effects on humans, particularly in the area of cognition.[30]

Studies based on autopsies and neuroimaging documented structural abnormalities in the brains of humans with substantial prenatal exposures to alcohol.[31] Social or economic differences might induce a spurious association between drinking while pregnant and a child's performance on cognitive tests; for instance, a parent's education and income predict a child's performance on cognitive tests.[32] Are social and economic differences plausible explanations of abnormalities in the structure of a child's brain?

In 2000, the U.S. National Institute on Alcohol Abuse and Alcoholism's *Tenth Special Report to the US Congress on Alcohol and Health* concluded:

> Fetal Alcohol Syndrome (FAS) is considered the most common nonhereditary cause of mental retardation. In addition to deficits in general intellectual functioning, individuals with FAS often demonstrate difficulties with

> learning, memory, attention, and problem solving as well as problems with mental health and social interactions . . . Estimates of FAS prevalence vary from 0.5 to 3 per 1,000 live births in most populations, with much higher rates in some communities.[33]

Many, perhaps all, of the individual strands of evidence that support the quoted conclusion are vulnerable to alternative interpretations. The woven tapestry of evidence—or the extensively filled in crossword puzzle—is considerably less vulnerable to alternative explanations. Wittgenstein taught us to ask, "What would a mistake here be like?"[34]

Some questions remain about the effects of light alcohol consumption, but there is no compelling evidence that low doses are safe.[35] In 2016, the U.S. Centers for Disease Control and Prevention wrote, "Why take the risk? . . . About half of all US pregnancies are unplanned and, even if planned, most women do not know they are pregnant until they are 4–6 weeks into the pregnancy . . . Sexually active women who stop using birth control should stop drinking alcohol."[36]

Many strands of evidence meet appropriately in support of the claim that sustained and heavy prenatal exposures to alcohol do enormous harm to the developing fetus. Could it be that each strand of evidence is vulnerable to an alternative explanation, as most if not all are vulnerable, and in each case the alternative explanation is correct? It is a logical possibility. Could this be true no matter how many different types of evidence are assembled so long as each one is open to some alternative interpretation? To claim this entails no error in logic: you make no error in logic—you do not contradict yourself—by making this claim. Logical possibility remains. Proof, in the mathematical sense of proof, is lacking. So long as all of the possible alternatives are jointly possible, there is no logical proof. But how important here is the absence of logical proof? In a more general context, Thomas Nagel wrote,

> There is no alternative to considering the alternatives and judging their relative merits . . . To dislodge a belief requires argument, and the argument has to show that some incompatible alternative is at least as plausible . . . Someone who said at every point that the apparently law-confirming experimental results were just coincidence would be crazy, but he would not be contradicting himself.[37]

In discussing fetal alcohol syndrome, I have been contrasting the weakness of individual studies that compose a body of evidence no longer weak.

Consensus and Repetition Do Not Make Weak Evidence Strong

There is a scientific consensus concerning fetal alcohol syndrome. How important is it that there is consensus? In the case of fetal alcohol syndrome, the current consensus reflects a change from a previous consensus. The consensus had been that alcohol was not a major focus of concern during pregnancy; then, the consensus changed.

The mere existence of consensus is not a useful guide. We should ask, Does a consensus have its origins and its ground in a rational and comprehensive appraisal of substantial evidence? Has the available evidence been open to vigorous challenge, and has it met each challenge? If a consensus has these origins, then the existence of consensus is of little consequence beyond its important origins. Conversely, a consensus that lacks these origins is of little consequence precisely because it lacks these origins.[38] Knowing the current consensus is helpful in forecasting a vote; having substantial evidence is helpful in judging what is true. That something is standardly believed or assumed is not, by itself, a reason to believe or assume it. Error and confusion are standard conditions of the human mind.

If a design for an observational study produces evidence that is open to a specific alternative interpretation besides a treatment effect, then repeating the same design in many studies will do little or nothing to adjudicate between a treatment effect and the alternative interpretation.[39] Mervyn Susser argued that evidence is strengthened when "diverse approaches produce similar results," particularly when these diverse approaches suffer from diverse weaknesses that are unlikely to align to produce similar results in the absence of an actual treatment effect.[40] For instance, in the case of fetal alcohol syndrome virtually all observational studies of pregnant women face the constant concern that women who drink heavily during pregnancy may differ as mothers from pregnant women who abstain from alcohol. This genuine concern is not an issue in the many controlled experiments on animals, which face legitimate but different concerns. Studies of the neurotoxicity of alcohol in rats suggest a biological

mechanism through which alcohol may produce effects, but leave open the consequences for cognition; however, this legitimate concern does not invalidate studies of cognitive deficits in rhesus monkeys given prenatal exposures to alcohol.

George Polya argues that similar considerations apply in heuristic reasoning in mathematics, developing various qualitative consequences of probabilistic reasoning. Discussing the heuristic reasoning that precedes the proof of a purported theorem, he wrote,

> On the one hand, the examination of a new consequence [of the purported theorem] supplies strong inductive evidence when this consequence has not been made plausible by the consequences examined previously. In practice, this will be the case when this consequence has no immediate relation with the old ones, when it is removed from the preceding, when this new consequence is not only new, but of a new kind. On the other hand, the examination of a new consequence introduces strong inductive evidence when it has a good chance of compromising the theorem. In practice, this will be the case when the examination touches upon a new aspect of the theorem, an aspect of the theorem which had not been previously considered.[41]

* Evidence Factors

* What Are Evidence Factors?

Typically, if you analyze one data set several times in slightly different ways, you simply repeat yourself. Moreover, if you understand your previous analyses as confirmations of your current analysis, then you are suffering from a statistical version of schizophrenia; the concurring voices you hear are your own, a consensus of one. Indeed, to analyze one data set several times can be dishonest if you report only some of these analyses, especially if you select the reported analyses because only these analyses favor a particular conclusion you wish to reach.

There is an exception, however. The exception is a single planned analysis of an elaborate theory in which several statistically independent tests are performed using the same data, where the different tests are vulnerable to different biases. It takes quite a bit of care in research design to produce this

situation, and in this book I can only sketch the idea, so one must turn to the references for specifics.[42]

To say that two tests are statistically independent under the null hypothesis H_0 of no treatment effect is to say that if H_0 were true then the result of the first test would provide precisely no information about the result of the second. That is something rare, something requiring care in research design, because typically two analyses of the same data strongly predict one another. Because of this statistical independence, the two analyses may be combined as though they came from different studies by unrelated investigators, using methods similar to those used in meta-analyses that combine findings from unrelated studies.

In the simplest situation, to say that the two analyses are vulnerable to different biases is to say that the first analysis would be correct if treatment assignment were ignorable in one sense, and that the second analysis would be correct if treatment assignment were ignorable in a second sense, but massive failures of the first sense would not bias the second analysis, and massive failures in the second sense would not bias the first analysis. This is a much stronger property than statistical independence. One can always pointlessly produce two statistically independent analyses of one data set by splitting the data in half at random and analyzing the two parts separately, but parallel analyses of the two halves will be affected in the same way by the same biases.

* An Example of Evidence Factors

To illustrate evidence factors, return to the study by Morton and colleagues of the effects on children of a father's exposure to lead at work. Figure 7.3 reorganizes the data in Figure 7.2. Panel (a) of Figure 7.2 described the lead level in the blood of exposed and control children separately, and panel (a) of Figure 7.3 described the matched pair difference in lead levels, exposed-minus-control or equivalently treated-minus-control. These differences tend to be positive in Figure 7.3, so most exposed children had higher levels of lead than their matched controls. Panels (c) and (d) of Figure 7.2 describe the level of lead for individual children grouped by the level of exposure to lead of the exposed father, and panel (b) of Figure 7.3 groups the treated-minus-control pair differences by the level of exposure of the exposed father.

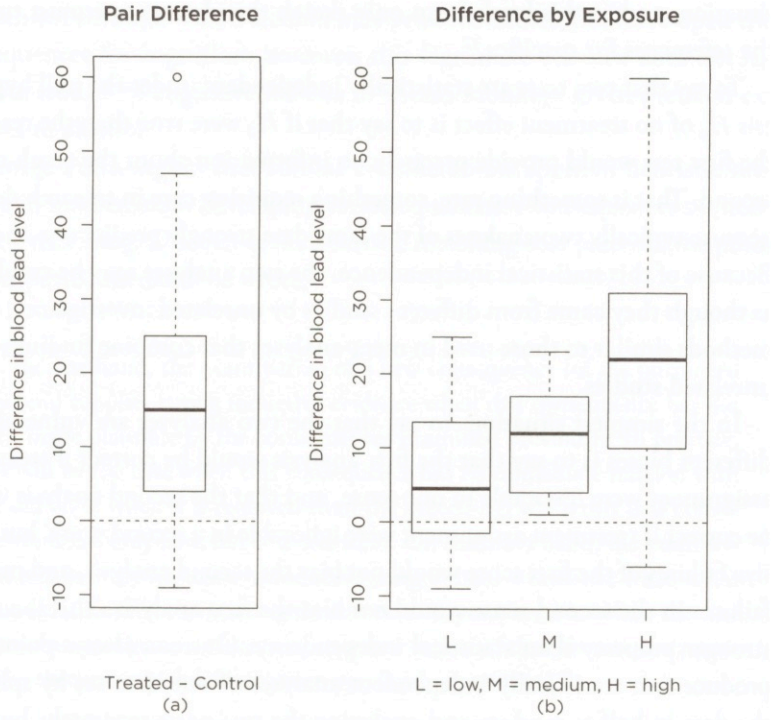

Figure 7.3. Matched pair differences, treated-minus-control, in levels of lead in children's blood, μg/dl. In each figure there is a horizontal line at zero. Panel (a) shows the differences, while panel (b) separates the differences into three groups based on the level of exposure to lead of the exposed father.

Due to one missing blood reading for one control child, Figure 7.3 refers to 33 pairs of two children.

One might hope that panel (a) of Figure 7.3 is analogous to a simple randomized experiment in which one child in each of 33 matched pairs was picked at random for exposure. One might hope that panel (b) of Figure 7.3 is analogous to a different simple randomized experiment in which levels of exposure were assigned to pairs at random. One might hope that panels (a) and (b) are jointly analogous to a randomized experiment in which both randomizations were done, within and among pairs. All three of these hopes may fail to be realized: there might be bias in treatment assignment within pairs or bias in assignment of levels of exposure to pairs. It turns out, however, that these several hopes are very different hopes; perhaps you will get

part, not all, of what you want. The hopes may be demarcated from each other, so we may ask about the strength of the evidence against the null hypothesis H_0 of no effect if one or another of the various hopes fails to be realized.

Let us consider a simple analysis of a conventional sort, then discuss some properties of the analysis. The analysis tests the null hypothesis H_0 of no treatment effect in two ways, one focused on panel (a) of Figure 7.3, the other on panel (b).

A simple analysis of panel (a) of Figure 7.3 asks whether the pair differences are too often and too substantially tilted toward positive values to be explained by unlucky but fair coin flips in assigning treatments within pairs. The test uses a familiar statistic proposed in 1945 by Frank Wilcoxon, called Wilcoxon's signed rank statistic.[43] It computes the absolute value of each pair difference, ranks those absolute values from one to the number of pairs—here, 33—then sums the ranks for pairs with a positive difference. The statistic turns out to be 499 in panel (a), whereas the largest it could have been is $1 + 2 + \cdots + 33 = 561$; so, as suggested by the appearance of Figure 7.3(a), the differences are heavily tilted toward positive values. If H_0 were true and if treatments were randomized within pairs, then an argument similar to that in Chapter 3 yields the distribution of Wilcoxon's statistic, from which we obtain the two-sided P-value of 1.15×10^{-5}. A boxplot such as panel (a), a boxplot so tilted toward positive values, would have been very improbable in a paired randomized experiment with no treatment effect.

A simple analysis of panel (b) of Figure 7.3 asks whether large pair differences are too often found with higher exposures to be explained by an unlucky but random assignment of exposure levels to pairs. The analysis uses another familiar statistic, Kendall's rank correlation test, which asks whether high values of two variables co-occur too often to be attributed to chance.[44] If H_0 were true and if levels of exposure were randomized among pairs, an argument similar to that in Chapter 3 yields the distribution of Kendall's statistic, from which we obtain a two sided P-value of 0.0104. If levels of exposure had been randomly allocated among pairs in the absence of a treatment effect, it is very unlikely that the steady increase in lead differences with increased levels of exposure in panel (b) of Figure 7.3 would have occurred.

So far, these are two simple statistical analyses of a conventional sort. However, the two analyses stand in an important relationship to one another

when the null hypothesis H_0 of no effect is true. If the assignment of exposure levels were seriously biased and far from randomized, this would invalidate the second test based on Kendall's statistic, but it would do no harm to the first test based on Wilcoxon's statistic, provided that treatments are assigned at random within pairs. In parallel, if the assignment of treatments within pairs were seriously biased and far from randomized, this would invalidate the Wilcoxon test, but it would do no harm to the second test based on Kendall's statistic, provided that exposure levels were assigned at random among pairs. In this sense, the evidence provided by each test is demarcated from the evidence provided by the other test. Moreover, if both assignments were independently randomized, then the two tests would be statistically independent under H_0; that is, if there were no treatment effect, each of the two P-values would provide precisely no information about the other.[45] Because of this, the two P-values can be combined using methods for combining independent P-values; for instance, using Fisher's method yields a combined P-value of 2.04×10^{-6}, considerably smaller than either of its components.[46]

In other words, the two comparisons in Figure 7.3, panels (a) and (b), are each fallible, but in very different ways; discard either one fearing that it is biased, and there is still evidence from the other. Moreover, the two comparisons together provide stronger evidence than either one on its own.

Wilcoxon's statistic and Kendall's statistic are exactly independent in Figure 7.3 under H_0 if both treatment assignments are randomized. This exact independence requires the use of these or similar rank statistics. Many other statistics yield a form of approximate independence in situations like Figure 7.3.[47]

So far, we have considered two extreme possibilities: a comparison, say, the treatment—control comparison in Figure 7.3, panel (a), is either randomized or so severely biased that it is useless. We have seen that either of the two comparisons in Figure 7.3, panels (a) and (b) may be useless without invalidating the other. Must we focus on extreme possibilities? Using the tools in Chapter 9, we may consider less extreme, intermediate cases, in which one comparison is biased to a limited degree but not entirely useless. Considerations of this kind might show that enormous biases affecting either comparison in Figure 7.3 and moderately large biases affecting the other would be insufficient to explain Figure 7.3 as something other than an effect caused by lead from the factory.[48]

There are forms of compulsive checking that do not check anything. Closely parallel analyses of the same data rarely check anything important. In contrast, in seeking evidence factors we are seeking analyses that are not in the slightest bit redundant; indeed, the findings of one analysis provide precisely no information about the findings in the other, so independent confirmation is possible. Wittgenstein remarked about the man who bought "several copies of the morning paper to assure himself that what it said was true."[49] We wish to avoid that man's misunderstanding.

Taking Stock

An elaborate causal theory is one that "predict[s] more and [can] be put more thoroughly to the test."[50] In the limited structure of Chapter 5, ignorable treatment assignment is needed for causal inference but it is not testable; however, an elaborate theory may make it testable. An elaboration of a causal theory is most useful if its predictions help to discriminate between ignorable treatment assignment and a particularly plausible pattern of unobserved biases.