Chapter Eight- Causal Reasoning

At the root of the whole theory of induction is the notion of physical cause. To certain phenomena, certain phenomena always do, and, as we believe, always will, succeed. The invariable antecedent is termed the 'cause,' the invariable consequent, the 'effect.' Upon the universality of this truth depends the possibility of reducing the inductive process to rules.
--John Stuart Mill

In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects.
--David Hume

The International Agency for Research on Cancer, an arm of the World Health Organization, labels a substance as a *probable human carcinogen* when at least two animal studies indicate a substance causes cancer.[1] This agency gives little or no weight, however, to studies indicating that substances do *not* cause cancer. Thus, crystalline silica, the main ingredient in beach sand and rocks, is declared as a substance "known to cause cancer," even though several studies have found no evidence of such a causal connection.

We read or hear in the media many stories about causes of cancer and other diseases. Often the stories seem to contradict one another. One story says that a study found that eating broccoli reduces cancer risk. Another story says that a new study casts doubt on the protective effects of broccoli. What are we to make of such studies and reports? How reliable are they?

While it would go beyond the limits of this book to try to resolve the issue of how to apply to humans the results of studies done on animals, we can shed some light on how to evaluate the animal tests themselves. And we can review the kinds of things to consider when evaluating causal reasoning. We'll begin by clarifying what we mean when we say that something is a **cause** or a **causal factor** of something else.

1. Causes as necessary and sufficient conditions

Whenever we say that *some x causes some y*, we are saying that *x* is a significant factor in bringing about *y*. Furthermore, when we think we know the *cause* of something, we think we have explained it. Though it is possible, of course, to know that, say, *smoking causes cancer* and not have a clue as to the actual mechanisms by which the chemicals in cigarette smoke bring about cell damage. Even if we cannot understand how chemicals in smoke turn normal cells into cancerous ones, the knowledge that smoking causes cancer can be applied by us. This gives us some control over the effects of smoking. In short, knowing *that* smoking causes cancer is not the same as knowing *how* smoking causes cancer. Our concern here is with claims *that* something is a cause of something else. The *how*, the actual causal mechanisms, cannot be our concern. Those studies must belong to the different disciplines studying the causes of diseases, the causes of bridges or buildings collapsing, the causes of hurricanes or earthquakes, and so on.

Sometimes, when we say that *x* is a *cause* or *causal factor* of *y*, we mean that *x* is a *necessary condition* for *y*. To say, for example that *smoking cigarettes is a significant causal factor in the development of lung cancer* is to say not only that (1) smoking is a significant factor in bringing about lung cancer but that (2) it is necessary

for one to smoke to develop the lung cancer caused by smoking and that (3) if one does not smoke one cannot get the lung cancer caused by smoking. So, even if we don't understand *how* smoking causes cancer, we can control the effects of smoking by not smoking.

Still, we all know that even if it is true that smoking causes cancer, there are people who have smoked for many years, lived to be old, and never developed cancer. So, how could smoking cause cancer if it is possible to smoke and not develop cancer? The answer is that smoking is *not* a **sufficient condition** for cancer. Not all causal events involve causes that are sufficient conditions. For example, some viruses are *necessary*, but *not sufficient*, conditions for getting the flu. Thus, for example, if a particular type of virus had not been present in Grimes' body, Grimes would not have contracted the flu. But had Grimes' immune system been working better or had Grimes been vaccinated (had his flu shot), he would not have gotten the flu even though the virus that causes the flu would still have been present in his body.

Had the virus been a **sufficient condition** for getting the flu, then Grimes would have gotten the flu no matter how well his immune system was working and no matter whether he did or did not have his flu shot. **If a causal factor is a sufficient condition for an effect, then that causal factor is sufficient to bring about its effect.** Smoking is not a sufficient condition for cancer. So, it is possible to smoke and not develop cancer, even though smoking is a significant causal factor in the development of lung cancer.

A third sense of 'cause' is that of a condition that is **both necessary and sufficient**. For example, the spirochete *Treponema pallidum* is both a necessary and a sufficient condition for syphilis. You cannot get syphilis if the spirochete is not present, and if the spirochete is present, you have syphilis. Thus, when we say *smoking causes cancer*, *a virus causes the flu*, and *a spirochete causes syphilis*, we do not mean exactly the same thing by 'causes' in the each case.

To say *x is a cause of y* is to say that *x* is a *necessary condition*, a *sufficient condition,* or *both a necessary and a sufficient condition* for bringing about *y*.

2. Causal claims: individuals and populations

We know that there are certain patterns to the behavior of both humans and natural phenomena. Some of these patterns involve variables (aspects that can change) which are complex, numerous, and difficult to control or measure. For example, cancer and heart disease seem to be related to many factors: genetic make-up, smoking, exercise, diet, weight, stress, etc. Finding the cause or causes of heart disease is not a matter of finding a single condition or lack of a single condition. Predicting who will get cancer or heart disease is usually difficult. We can be very sure that a person subjected to high doses of certain kinds of radiation will soon develop cancer. However, most cancer predictions will have to be couched in terms of *statistical probabilities*. Thus, instead of saying that a certain smoker is likely to die because of the effects of smoking, we say that *the death rate for smokers is 1.57 times the death rate for non-smokers* or that *the relative death rate from lung cancer is over 10 times greater in smokers than in non-smokers.*

It is difficult to translate a statistical probability into a specific prediction regarding any particular individual. Nevertheless, it seems clear from the data that smokers run a *greater risk* of harming their health and shortening their lives than non-smokers do. How much of a risk each individual smoker runs cannot be precisely determined. For some, the risk may be insignificant; for others, it may be enormous.

It is one thing to know that a study found that smokers died of lung cancer at a rate ten times higher than expected. In itself, this statistic does not prove that smoking is a significant causal factor in the development of lung cancer, but it supports the hypothesis that it is. But, even if smoking is a relevant and significant factor in the development of lung cancer *in general*, what does this relevance and significance mean for any given individual who now smokes but does not have lung cancer?

Determining that smoking causes lung cancer is not like establishing that if a baseball is thrown at a typical window in a typical house, the window will break. We can measure the fragility of the window and the force of the thrown object with sufficient accuracy to be able to predict which particular window will break when hit by the

thrown ball. But we cannot measure a given individual's susceptibility to lung cancer with anything near the same kind of accuracy, even though we know that if the individual smokes his or her susceptibility to lung cancer is greater than it would be if the individual did not smoke. Thus, even if we can accurately say that smoking causes lung cancer, we cannot accurately say that any given individual who smokes will contract lung cancer due to smoking. We can accurately say, however, that if one smokes, then one runs a considerably higher risk of contracting lung cancer than one would if one did not smoke.

3. Sequences, correlations, and causes

Causal relationships are regular patterns in nature that are characterized by **sequences of events** and **correlations**. Thus, if there is a causal relationship between smoking and lung cancer, then the smoking must precede the development of lung cancer and there must be a *correlation* between smoking and lung cancer. If there is a correlation between smoking and lung cancer, then there must at least be a significant difference between the percentage of smokers with lung cancer and the percentage of non-smokers with lung cancer. A significant difference is one that is not likely due to chance.[2] In a later section, we will discuss methods of testing causal hypotheses. There we will explain in more detail how one establishes whether a correlation is significant.

4. Causal fallacies

It is obvious that many events occur in sequences without any causal connection between them. A car passes by and a leaf falls from a nearby tree, for example. Moreover, not all correlated events are causally related. A correlation may be due to **chance** or **coincidence**. For example, correlations exist between batting averages of major league ball players and presidential election years (the averages are usually lower during election years). Correlations have been established between sex crimes and phases of the moon and between hair color and temperament. Finally, there is the story of the jungle natives who beat their drums every time there is an eclipse. It never fails—the sun always returns after the ceremony: a case of *perfect correlation but no causality*. *To conclude that one thing must be the cause of the other solely because the two things are correlated is to commit the fallacy of false cause or questionable cause.*

Another characteristic of causally related events is that the cause precedes the effect. However, *simply because one event precedes another does not mean that there is a causal relationship between the two events.* To reason that one thing must be the cause of the other solely because it preceded the other is to commit the **post hoc fallacy**, a type of *false cause* reasoning. The name is taken from the Latin *post hoc ergo propter hoc* (after this, therefore because of this). Just because the pain in your wrist went away after you started wearing your new magic copper or magnetic bracelet, does not mean that the bracelet caused the pain to be relieved. Just because the patient died right after the priest gave the patient a blessing, does not mean the blessing caused the death! Just because you thought of your mother right before she telephoned, does not mean your thoughts caused her to call!

4.1 The regressive fallacy

Things like stock market prices, golf scores, and chronic back pain inevitably fluctuate. Periods of low prices, low scores, and little or no pain are followed by periods of higher prices and scores, and greater pain. Likewise, periods of high prices and scores, and little pain are followed by periods of lower prices and scores, and more severe pain. This tendency to move toward the average away from extremes was called "regression" by Sir Francis Galton in a

study of the average heights of sons of very tall and very short parents. (The study was published in 1885 and was called "Regression Toward Mediocrity in Hereditary Stature.") He found that sons of very tall or very short parents tended to be tall or short, respectively, but not as tall or as short as their parents. To ignore these natural fluctuations and tendencies often leads to self-deception regarding their causes and to a type of post hoc fallacy. The **regressive fallacy** is the failure to take into account natural and inevitable fluctuations of things when ascribing causes to them (Gilovich 1993: 26).

For example, a professional golfer with chronic back pain or arthritis might try a copper bracelet on his wrist or magnetic insoles in his shoes. He is likely to try such gizmos when he is not playing well or is not feeling well. He notices that after using the copper or the magnets that his scores improve and his pain diminishes or leaves. He concludes that the copper bracelet or the magnetic insole is the cause. It never dawns on him that the scores and the pain are probably improving due to natural and expected fluctuations. Nor does it occur to him that he could check a record of all his golf scores before he used the gizmo and see if the same kind of pattern has occurred frequently in the past. If he takes his average score as a base, most likely he would find that after a very low score he tended to shoot a higher score in the direction of his average. Likewise, he would find that after a very high score, he tended to shoot a lower score in the direction of his average.

Many people are led to believe in the causal effectiveness of worthless remedies because of the regressive fallacy. The intensity and duration of pain from arthritis, chronic backache, gout, and the like, naturally fluctuates. A remedy such as a chiropractic spinal manipulation, acupuncture, or a magnetic belt is likely to be sought when one is at an extreme in the fluctuation. Such an extreme is naturally going to be followed by a diminishing of pain. It is easy to deceive ourselves into thinking that the remedy we sought caused our reduction in pain. It is because of the ease with which we can deceive ourselves about causality in such matters that scientists do controlled experiments to test causal claims.

4.2 The clustering illusion

The clustering illusion is the intuition that random events occurring in clusters are not really random events. The illusion is due to selective thinking based on a false assumption. For example, it strikes most people as unexpected if heads comes up four times in a row during a series of coin flips. However, in a series of 20 flips, there is a 50% chance of getting four heads in a row (Gilovich). "In a random selection of twenty-three persons there is a 50 percent chance that at least two of them celebrate the same birthdate."[3] What are the odds of anyone dreaming of a person dying and then that person actually dying within 12 hours of the dream? A statistician has calculated that in Britain this should happen to someone every two weeks (Blackmore 2004: 301). It may seem unexpected, but the chances are better than even that a given neighborhood in California will have a statistically significant cluster of cancer cases.[4]

What would be rare, unexpected, and unlikely due to chance would be to flip a coin twenty times and have each result be the alternate of the previous flip. In any series of such random flips, it is more unlikely than likely that short runs of 2, 4, 6, 8, etc., will yield what we know logically is predicted by chance. In the long run, a coin flip will yield 50% heads and 50% tails (assuming a fair flip and fair coin). But in any short run, a wide variety of probabilities are expected, including some runs that seem highly improbable.

Finding a statistically unusual number of cancers in a given neighborhood--such as six or seven times greater than the average--is not rare or unexpected. Much depends on chance and much depends on where you draw the boundaries of the neighborhood. Clusters of cancers that are seven thousand times higher than expected, such as the incidence of mesothelioma (a rare form of cancer caused by inhaling asbestos) in Karian, Turkey, are very rare and unexpected. The incidence of thyroid cancer in children near Chernobyl in the Ukraine was one hundred times higher after the disaster in 1986.[5] Such extreme differences as in Turkey and the Ukraine are never expected by chance.

Sometimes a subject in an ESP experiment or a dowser might be correct at a higher than chance rate. However, such results do not indicate that an event is not a chance event. In fact, such results are predictable by the laws of

chance. Rather than being signs of non-randomness, they are actually signs of randomness. ESP researchers are especially prone to take streaks of "hits" by their subjects as evidence that psychic power varies from time to time. Their use of optional starting and stopping (counting only data that supports their belief in ESP) is based on the presumption of psychic variation and an apparent ignorance of the probabilities of random events. One would expect, by the laws of chance, that occasionally a subject would guess cards or pictures (the usual test for ESP) at a greater than chance rate for a certain run. Combining the **clustering illusion** with **confirmation bias** is a formula for self-deception and delusion.

A classic study on the clustering illusion was done regarding the belief in the "hot hand" in basketball.[6] It is commonly believed by basketball players, coaches, and fans that players have "hot streaks" and "cold streaks." A detailed analysis was done of the Philadelphia 76ers shooters during the 1980-81 season. It failed to show that players hit or miss shots in clusters at anything other than what would be expected by chance. The researchers also analyzed free throws by the Boston Celtics over two seasons and found that when a player made his first shot, he made the second shot 75% of the time and when he missed the first shot he made the second shot 75% of the time. Basketball players do shoot in streaks, but within the bounds of chance. It is an illusion that players are 'hot' or 'cold'. When presented with this evidence, believers in the "hot hand" are likely to reject it because they "know better" from experience.

In epidemiology, the clustering illusion is known as the **Texas-sharpshooter fallacy**. The term refers to the story of the Texas sharpshooter who shoots holes in the side of a barn and then draws a bull's-eye around the bullet holes. Individual cases of disease are noted and then the boundaries are drawn.[7] Kahneman and Tversky called it "belief in the Law of Small Numbers" because they identified the clustering illusion with the fallacy of assuming that the pattern of a large population will be replicated in all of its subsets.[8] In logic, this fallacy is known as the **fallacy of division**, **assuming that the parts must be exactly like the whole.**[9] For example, just because the leukemia rate for children is such that, say, six children in your city would be expected to contract the disease in a given year, there is no immediate cause for alarm if two or three times that number are diagnosed this year. Such fluctuations are expected due to chance. However, if your area consistently has several times more new cases than the national average, there could well be an environmental factor causally related to the cancers. Of course, it might also be due to the fact that there is good treatment in your area and people are moving there with their sick children because of it.

4.3 Correlation and causality

When two events (call them *x* and *y*) are significantly correlated, that does not necessarily mean that they are causally related. The correlation could be spurious and coincidental. Even if the correlation is not coincidental, it is possible that *x* causes *y* or *y* causes *x* or that *z* causes both *x* and *y*. We may find a significant correlation between the rise in sex education classes and the rise in teenage pregnancy. The classes may be stimulating the teens to experiment sexually, leading to the increase in teen pregnancies. Or, the classes may have been instituted in response to the rise in teen pregnancy. Or, it may just be a coincidence. There may be a significant correlation between hip size and grade point average among sorority sisters, but it is unlikely that either factor causes the other. Perhaps the larger sisters study more, while their thinner ones are socializing when they should be studying. Perhaps it's just coincidence. The moral is simple: Correlation does not prove causality. Yet many scientists seem to be on a quest to do nothing more than find statistically significant correlations and conclude that when they find them they have evidence of a causal connection.

Many defenders of psychic phenomena, for example, think very highly of Robert Jahn's experiments with people trying to use their minds to affect the random output of various electronic or mechanical devices. Jahn was an engineering professor until he got interested in the paranormal. The work took place at Princeton University in the Princeton Engineering Anomalies Research (PEAR) lab. In 1986, Jahn, Brenda Dunne, and Roger Nelson reported on millions of trials by 33 people over seven years of trying to use their minds to override random number

generators (RNG). (Think of these devices as randomly producing one of two variables, a 0 and 1, for example. Your task is to try to will more 0s than 1s, or vice versa.) In 1987, parapsychologist Dean Radin and Nelson did a meta-analysis of the RNG experiments and found that they produced odds against chance beyond a trillion to one (Radin 1997: 140). This sounds impressive, but as Radin says "in terms of a 50% hit rate, the overall experimental effect, calculated per study, was *about* 51%, where 50% would be expected by chance" [emphasis added] (141). Similar results were found with experiments where people tried to use their minds to affect the outcome of rolls of the dice, according to Radin.

However, according to psychologist Ray Hyman, "the percentage of hits in the intended direction was only 50.02%." And one 'operator' (the term used to describe the subjects in these studies) was responsible for 23% of the total data base. His hit rate was 50.05%. Take out this operator and the hit rate becomes 50.01% (Hyman 1989: 152). This reminds us that statistical significance does not imply importance. Furthermore, Stanley Jeffers, a physicist at York University, Ontario, has repeated the Jahn experiments but with chance results (Alcock 2003: 135-152).

Based on the results of these experiments, Radin claims that "researchers have produced persuasive, consistent, replicated evidence that mental intention is associated with the behavior of …physical systems" (Radin 1997: 144). He also claims that "the experimental results are not likely due to chance, selective reporting, poor experimental design, only a few individuals, or only a few experimenters" (Radin 1997: 144). Radin is considered by many as a leading parapsychologist, yet it is difficult to see why anyone would find these correlations indicative of anything important, much less as indicative of psychic powers.

An even more problematic area of scientific research that seeks statistically significant correlations is the area of healing prayer. Several scientists have tried to prove that praying for somebody can heal them by the intercession of some spiritual force. Some have found significant correlations between prayer and healing. Some critics have accused these researchers of fraud. But my concern is that these scientists are assuming that finding a statistic that is not likely due to chance is evidence for supernatural intervention. They also don't seem to realize what it would mean if supernatural beings (SBs) could cause things to happen in the natural world. This might sound like a good thing. After all, who wouldn't like to be able to contradict the laws of nature whenever it was convenient to do so? However, if SBs could contravene the laws of nature at will, human experience and science would be impossible. We are able to experience the world only because we perceive it to be an orderly and lawful world. If SBs could intervene in nature at will, then the order and lawfulness of the world of experience and of the world that science attempts to understand would be impossible. If that order and lawfulness were impossible, then so would be the experience and understanding of it. Finally, if spirits could intervene in our world at will, none of the tests for causal claims, which we will turn to now, would even be possible.

5. Testing causal claims

There are several ways we can test causal claims. Three of the most important empirical methods are (1) **the controlled experiment,** (2) **the prospective study**, and (3) **the retrospective study.**

Each of the empirical methods of testing causal claims presupposes certain *logical methods of analysis* as well. The logical methods were systematically presented by John Stuart Mill in the nineteenth century, and thus are known as *Mill's Methods*.

5.1 Mill's methods

The first of Mill's methods is called the **method of agreement**. If six people at a dinner party get sick and the only food they all ate was the salmon, then the salmon probably caused the sickness, all else being equal. When only 'C' is shared in common by instances of 'E', then 'C' is probably the cause of 'E'. (Of course, the difficulty is in knowing that 'C' is the only thing besides 'E' that the instances have in common.)

The second method is called the **method of difference**. If two people have dinner, one having steak and the other salmon, and the one having salmon get sick, then the salmon probably caused the sickness, all else being equal. When 'E' occurs where 'C' is present but does not occur where 'C' is not present, then 'C' is probably the cause or a significant causal factor of 'E'. (Again, the difficulty is in knowing that 'C' is the only significant difference between the items.)

The third method is the **joint method of agreement and difference**. If a group of laboratory rats is randomly divided into two groups that are treated exactly alike except that one group is given a large dose of arsenic while the other is not, and all those given the arsenic die shortly thereafter, while none of those not given the arsenic die shortly thereafter, then the arsenic probably caused the deaths. When two groups differ only in 'C' and the group which has 'C' also develops 'E' but the group with no 'C' does not develop 'E', then we reason that 'C' is the cause or a significant causal factor of 'E'. (Everything here depends on knowing that everyone in one group got 'C' and no one in the other group got 'C' and that the only significant difference between the two groups is 'C'.)

A fourth method is the **method of concomitant variation**. If 'C' and 'E' are causally related there ought to be a correlation between them such that 'E' varies directly with the presence or absence of 'C'. For example, if lead causes lung cancer, then there should have been proportionately fewer cases of lung cancer as the amount of lead in the atmosphere decreased, as it did during the gasoline shortage during World War II and in the period following the removal of lead from gasoline in the U.S. Also, there should have been a proportionate increase in lung cancer as gasoline usage increased after the war.

Mill's methods are not tests of causal hypotheses but they are presupposed to some degree by the empirical tests we will now consider.

5.2 Controlled experiments

The preferred scientific way to test a causal claim is by doing a **controlled experiment**. If, for example, we want to test the causal claim that *crystalline silica causes lung cancer*, we might (1) randomly assign mice to two groups; (2) introduce crystalline silica into one group (the **experimental group**) but not the other (the **control group**), and otherwise treat the two groups identically for a specified length of time; and then (3) determine if there is a significant difference in the lung cancer rates of the two groups. If *crystalline silica causes lung cancer*, we should observe a significantly higher incidence of lung cancer in our experimental group. A significant difference would be one that is not likely due to chance. If we found that the lung cancer rate in the experimental group was 8.4 percent and in the control group was 7.9 percent, it is possible that the 0.5 percent difference is due to chance. That is, it is possible that had we studied two groups of mice, neither of which had been given the silica, we might have gotten similar results. However, if our study results in twice as many lung cancers in the experimental group as in the control group, we would be justified in concluding that silica is probably a cause of lung cancer. Such a huge difference in groups is not likely due to chance and is probably due to the silica.

If we did not use a **control group**, we would have no way of knowing whether an 8.4 percent rate of lung cancer among the mice we gave the silica to was significant or not. For all we know, that might be a typical rate among mice in general. By having a control group, we can compare the rate of the experimental group to a group that has *not* been affected by the substance we are testing. If silica is *not* a cause of lung cancer, then we should not see a rise in lung cancer rate among mice that have ingested silica. Thus, if our study resulted in the two

groups having very similar lung cancer rates, we would be justified in concluding that silica is probably not a cause of lung cancer.

To justify thinking that a difference in effect observed between an experimental and a control group is indicative of a causal event, it is necessary that the groups be of adequate size and that the study go on long enough for the effects, if any, to reveal themselves. If the control and experimental groups are too small, we cannot be sure that any difference we observe is not due to chance. One advantage to mice studies is that the mice can be bred to be nearly identical. If the mice were not nearly identical, much larger groups of mice would have to be studied. Studies that can now be done with *hundreds* of mice would require *thousands* instead. Nevertheless, it is important that the mice be *randomly* assigned to their groups, to reduce the chances of any unknown bias entering the process.

Why study mice, you might ask? If we are interested in whether silica causes lung cancer in humans, should we be studying humans instead of mice? Yes and no. We couldn't in good conscience do a controlled experiment with humans that requires us to give the experimental group members a substance that is potentially lethal. Animal studies do introduce, however, analogical issues that must be dealt with.

5.2.1 Animal Studies

Much scientific reasoning is based upon research done on animals such as rats and mice. The reasoning depends upon drawing comparisons between humans and rodents. Obviously, there are many dissimilarities between humans and rodents. There are also many physiological, anatomical, and structural similarities. As an example of the difficulty in evaluating such reasoning, we will look at some comments that were frequently heard regarding a study done on the effects of saccharin on laboratory animals.

A group of 183 laboratory rats was randomly divided into an experimental group of 94 and a control group of 89. The experimental group was fed the same diet as the control group with the exception of saccharin. Saccharin was given to 78 first-generation and 94 second-generation rats. (The rats were observed for two generations; the second-generation experimental rats were fed saccharin from the moment of conception.) The amount of saccharin given the rats varied. The maximum dose of saccharin amounted to 7 percent of the rats' diet. In the experimental group, cancers developed only in the rats given the 5 percent dose: in the first-generation, 7 male rats and 0 female rats developed bladder cancers; in the second-generation 12 males and 2 females developed bladder cancers. In the control group, only one rat—a first-generation male—developed bladder cancer. The differences are not likely due to chance and indicate a causal relationship between saccharin and bladder cancer in rats.[10]

Many well-informed scientists concluded from this study that saccharin probably causes cancer in humans. Some non-scientists criticized the study because they believed that since the quantity of saccharin given to the rats (about 7 percent of their daily diet) was the equivalent of the amount of saccharin in about 800 12-ounce cans of diet soda, the study had little relevance for humans. These critics felt that the *quantity* of saccharin given the rats was obviously *relevant* to whether or not it would cause cancer in humans. They thought that since the amount humans are likely to ingest was significantly different, it invalidated any conclusion regarding the carcinogenic effect of saccharin on humans. On the surface, the critics appear to be quite reasonable. A bit of critical thinking, however, reveals that the critics were not thinking very critically. It might turn out to be correct that giving high doses of substances to animals is not a very good way to determine the potential carcinogenic effects of those substances on humans. However, this cannot be known a priori or intuitively; it can only be discovered empirically.

A critical thinker knows that in areas outside of his or her own area of expertise, authorities or experts in those fields must be consulted and trusted beyond non-experts, provided that the area of the field of expertise at issue is not controversial among the experts themselves. The field of cancer research is well established, and there are things known by these researchers that may not be understood by the public. In any case, before assuming that the quantity of saccharin given the rats invalidates the study, a critical thinker would investigate the issue. A little research would reveal that at the time the saccharin studies were done most researchers agreed that anything that had been found to cause cancer seemed to cause it without regard to the quantity of the carcinogen. Quantity seemed only to affect the *speed* of the development of cancer. Giving high doses allowed researchers to discover potential carcinogens much more quickly and with much smaller samples than would otherwise be needed. Instead

of a few hundred mice studied for a short period, the saccharin study would have gone on for years and would have required about 85,000 mice at dosages close to normal human exposure.[11] Scientists give large doses of suspected carcinogens to speed up and streamline their studies, not because they are ignorant as to what a relevant dosage would be. Most cancer researchers, in other words, did not believe that the massive doses given rodents invalidated their studies.

On the other hand, a little research on why scientists use large doses would also have revealed that there is some controversy in this area. Not all researchers agree with the practice of giving massive doses of chemicals to the animals. For example, Bruce Ames, a professor of molecular and cell biology at the University of California at Berkeley, and Lois Gold, a cancer researcher at Lawrence Berkeley Laboratory, think that many chemicals described as hazardous may actually be harmless at normal human exposures.[12] However, even though Ames and Gold think that the method of giving massive doses of substances to mice is bankrupt, they do *not* claim either that the analogy between mice and humans is a bad one or that common sense tells them that the quantity of dosage makes the analogy irrelevant. Rather, they base their conclusion on recent evidence that they believe shows that "tumors are induced by the high doses themselves, because they tend to stimulate chronic cell division."

A 1992 report by the National Institute of Environmental Health Sciences, the branch of the National Institute of Health that directs animal studies, states that many of the assumptions driving rat and mouse research "do not appear to be valid."[13] The report claims that the practice of giving rodents the maximum dosages they can tolerate (M.T.D.) produces about two-thirds "false positives." "In other words, two-thirds of the substances that proved to be cancerous in the animal tests would present no cancer danger to humans at normal doses."[14] In addition, there is also evidence that some substances that are *not* carcinogenic in animal studies nevertheless cause cancer in humans, e.g., arsenic.

A new research method, based on an *analogy* between animal cells and human cells, looks promising. However, it is "costly and time consuming," according to Dr. Robert Maronpot, who has used the method in a study of oxazepam, a direct chemical relative of Valium, one of the most-often prescribed drugs in America. The method involves examining frozen DNA sections from animals given varying dosages of substances. Dr. Maronpot found that the rats and mice in M.T.D. studies develop cancer because of the high doses of oxazepam. "Oxazepam would not be a problem even for a mouse at normal human dosage levels," he said.

What this means is not that animal studies are irrelevant for drawing conclusions about humans. Nor does it mean that massive doses of substances should *never* be given in animal studies. It does seem to indicate, however, that reliance on statistics in animal studies to establish probabilities of cancer-causing substances will be diminished.[15]

5.2.3 Control Groups

Comparing experimental to control groups is useful for discovering causal connections only if the two groups are alike in all relevant respects except for the characteristic being tested. For example, it would be irrelevant to compare a group of smokers to a group of non-smokers in order to study the effects of smoking on a person's health if the smokers are sedentary, overweight, under extreme stress, eat vast quantities of fried foods, are alcoholics and have family histories of heart disease, while the non-smokers are active, of average or below average weight, lead calm lives, eat healthful foods, drink alcoholic beverages in moderation and do not have family histories of heart disease. The differences between the two groups would make the comparison irrelevant; for, any of the differences in weight, activity, health history, and the like, could account for the differences in health between the two groups. *In short, all potentially relevant causal factors should be identical between the experimental and control groups except for the factor being tested.*

Likewise, our controlled experiment to test the claim that silica causes lung cancer would have been invalidated if the two groups of mice were significantly different to begin with. For example, the experiment would be invalid if one group consisted of mice bred to be clones and to be used for scientific experimentation, while the

other group consisted of field mice. Any significant difference in lung cancer rates we might observe between the two groups might be due to significant differences they had before the experiment began.

How can one tell whether one has controlled for all potential causal factors except the one we are testing? You cannot, at least not with absolute certainty. There is always the possibility that you have overlooked something, or are ignorant of the real cause of the effect being studied. Again, it requires *background knowledge* in order to be able to judge competently what is or is not potentially relevant in a given situation. Moreover, there is always some possibility of error. Hence, as with all inductive reasoning, conclusions regarding causes must be stated as being to some degree *probable*.

On the other hand, if one were not to use a control group, one could not know that any effect observed was due to the suspected cause. For example, you might think a particular acne medication is a miracle worker at getting rid of pimples because you put it on for a week and your blemishes went away. However, if you had done nothing at all, your blemishes might have gone away. On the other hand, perhaps you also did something else that week (e.g., gave your face a nightly ice bath), which was actually the cause of the blemish reduction. However, if you put the acne medication on one side of your face but not the other and treat both sides of your face equally during the week, then if there is a difference in effect, it is probably due to the medication. Without a control, you cannot be sure what caused the effect.

5.2.4 Controlled studies with humans

One of the benefits of working with animals is that one doesn't have to worry about their beliefs or demeanors affecting the outcome of the study. The beliefs and demeanors of humans, both researchers and subjects in experiments, can affect the outcome of a controlled study.

Robert Rosenthal has found that even slight differences in instructions given to control and experimental groups can affect the outcome of an experiment. Different vocal intonations, subtle gestures, even slight changes in posture, might influence the subjects.

The **experimenter effect** is the term used to describe any of a number of subtle cues or signals from an experimenter that affect the performance or response of subjects in an experiment. The cues may be unconscious nonverbal cues, such as muscular tension or gestures. They may be vocal cues, such as tone of voice. Research has demonstrated that the expectations and biases of an experimenter can be communicated to experimental subjects in subtle, unintentional ways, and that these cues can significantly affect the outcome of the experiment (Rosenthal 1998).

Many researchers have found evidence that people in experiments who are given inert substances (**placebos**) often respond as if they have been given an active substance (the **placebo effect**). The placebo effect may explain, in part, why therapies such as homeopathy, which repeatedly fail controlled testing, have many satisfied customers nonetheless (Carroll 2003: 293).

Double-blind experiments can reduce experimenter and placebo effects. In a **double-blind** experiment, neither the subjects of the experiment nor the persons administering the experiment know certain important aspects of the experiment. For example, an experimenter who works for a drug company that is trying to produce a new drug for depression would be wise to have another experimenter randomize the participants into the group getting the new drug and the group getting a placebo. The subjects should not be told what group they are in; their expectations might affect the outcome. The experimenter who does the randomization should keep to himself any information regarding who is in which group and should not be the one to distribute the pills/placebos to the subjects. The one who hands out the pills/placebos and keeps records of who gets which should not be the one who records the effects on the participants. In this way, any bias on the part of the experimenters is minimized. Only after the experiment is concluded should the members of the groups be unblinded. There should be an exception, of course, if something bizarre began happening, such as several patients dying of heart attacks. In such cases, the experiment might be halted and the participants unblinded to examine the possibility that the new drug might be killing people (Carroll: skepdic.com/experimentereffect.html).

The experimenter effect may explain why many experiments can be conducted successfully only by one person or one group of persons, while others repeatedly fail in their attempts to replicate the results. Of course, there are other reasons why studies cannot be replicated. The original experimenter may have committed errors in design, controls, or calculations. He may be deceived himself about his ability to avoid being deceived. Or he may have committed fraud. While the experimenter effect does not discriminate among sciences, parapsychologists maintain that their science is especially characterized by the ability of believers in psi (ESP and psychokinesis) to get positive results, while skeptics get negative results. Many parapsychologists also believe that subjects in their card-guessing experiments who believe in psi produce above chance results, while those who are skeptics produce below chance results (the **sheep-goat effect**). Skeptics note that in any card-guessing experiment there should be some who score above chance and some who score below. What is not expected is that one person should consistently and repeatedly score significantly above chance. Some parapsychologists who have subjects that seem to perform psychic tasks such as bending metal with their minds or guessing correctly 20 out of 25 cards whose faces they can't see have found that when they institute tighter controls on their subjects, the subjects lose their ability to perform. Rather than admit that the apparent success of psychic ability was due to poor controls, these scientists claim that psi diminishes with increased testing (the **decline effect**) or that psi only works when the experimenter shows absolute trust in the subject. Having rigorous controls to avoid cheating or sensory leakage (unconscious and inadvertent communication between the subject and others) shows distrust, they argue, and destroys psychic ability. Skeptics think these are **ad hoc hypotheses**; that is, claims created to explain away facts that seem to refute their belief in psi.

5.2.5 Replication

Since there are so many variables that can affect the outcome of a controlled study, it is important that we not put too much faith in a single study. If we or others are unable to replicate the results of a controlled study, it is likely that the results of our study were spurious. If, on the other hand, our results can be replicated by ourselves and by others, that is an indication that the results were not spurious, but genuine. However, if the original study design was flawed in some way or our control protocols were not adequate, then replication is meaningless. For example, when Fleischmann and Pons announced that they had successfully produced cold fusion, they were greeted with both skepticism and with attempts by others to duplicate their experiments. (This was difficult, since they didn't publish all the details of their experiment.) Even when others seemed to duplicate the results of their experiments, many remained skeptical. Further inquiry determined that the results that seemed so promising were due to faulty equipment and measurements, not cold fusion.

One of the traits of a cogent argument is that the evidence be sufficient to warrant accepting the conclusion. In causal arguments, this generally requires—among other things—that a finding of a significant correlation between two variables, such as magnets and pain, be reproducible. Replication of a significant correlation usually indicates that the finding was not a fluke or due to methodological error. Yet, the news media and research centers often create the illusion of importance of single studies. For example, the *University of Virginia News* published a story about a study done on magnet therapy to reduce fibromyalgia pain. The study, conducted by University of Virginia (UV) researchers, was published in the *Journal of Alternative and Complementary Medicine*, which asserts that it "includes observational and analytical reports on treatments outside the realm of allopathic medicine...."

The only people who refer to conventional medicine as *allopathic* are opponents of conventional medicine. So, they may not be the most objective folks in the world when it comes to evaluating anything "alternative." Be that as it may, the study must stand or fall on its own merits, not on the biases of those who publish it. Furthermore, the study must be distinguished from the press release put out by UV. The headline of the UV article states that Magnet Therapy Shows Limited Potential for Pain Relief. The first paragraph states that "the results of the study were inconclusive." Even so, the researchers claimed that magnet therapy reduced fibromyalgia pain intensity enough in one group of study participants to be "clinically meaningful." (Perhaps UV considers 'limited potential' as the middle ground between 'inconclusive' and 'clinically meaningful.')

The UV study involved 94 fibromyalgia patients who were randomly assigned to one of four groups. One control group "received sham pads containing magnets that had been demagnetized through heat processing" and the other got nothing special. One treatment group got "whole-body exposure to a low, uniformly static magnetic field of negative polarity. The other...[got]...a low static magnetic field that varied spatially and in polarity. The subjects were treated and tracked for six months."

"Three measures of pain were used: functional status reported by study participants on a standardized fibromyalgia questionnaire used nationwide, number of tender points on the body, and pain intensity ratings." One of the investigators, Ann Gill Taylor, R.N., Ed.D. stated: "When we compared the groups, we did not find significant statistical differences in most of the outcome measures." Taylor is a professor of nursing and director of the Center for Study of Complementary and Alternative Therapies at UV. "However, we did find a statistically significant difference in pain intensity reduction for one of the active magnet pad groups," said Taylor. The article doesn't mention how many outcome measures were used.

The study's principal investigator was Dr. Alan P. Alfano, assistant professor of physical medicine and rehabilitation and medical director of the UV HealthSouth Rehabilitation Hospital. Alfano claimed that "Finding any positive results in the groups using the magnets was surprising, given how little we know about how magnets work to reduce pain." Frankly, I find it surprising that Alfano finds that surprising, since it is unlikely he would have conducted the study if he didn't think there might be some pain relief benefit to using magnets. His statement assumes they work to reduce pain and the task is to figure out how. Alfano is also quoted as saying that "The results tell us maybe this therapy works, and that maybe more research is justified. You can't draw final conclusions from only one study." This last claim is absolutely correct. His double use of the weasel word 'maybe' indicates that he realizes that one shouldn't even make the claim that more research ought to be done based on the results of one study, especially if the results aren't that impressive.

Not knowing how many outcome measures the researchers used makes it difficult to assess the significance of finding one or two outcomes that look promising. Given all the variables that go into pain and measuring pain, and the variations in the individuals suffering pain (even those diagnosed as having the same disorder), it should be expected that if you measure enough outcomes you are going to find something statistically significant. Whether that's meaningful or not is another issue. A competent researcher would not want to make any strong causal claims about magnets and pain on the basis of finding one or two statistically significant outcomes in a study that found that most outcomes showed nothing significant.

But even if most of the outcomes had been statistically significant in this study of 94 patients, that still would not amount to strong scientific evidence in support of magnet therapy. The experiment would need to be replicated. Given the variables mentioned above, it would not be surprising if this study were replicated but found different outcomes statistically significant. Several studies might find several different outcomes statistically significant and some researcher might then do a **meta-study** (a study that combines the data of several studies and treats them as if they were part of one study) and claim that when one takes all the small studies together one gets one large study with very significant results. Actually, what you would get is one misleading study.

If other researchers repeat the UV study, looking only at the outcome that was statistically significant in the original study, and they duplicate the results of the UV study, then we should conclude that this looks promising. But one replication shouldn't seal the deal on the causal connection between magnets and pain relief. One lab might duplicate another lab's results but both might using faulty equipment manufactured by the same company. Or both might be using the same faulty subjective measures to evaluate their data. Several studies that showed nothing significant for magnets and pain might be followed by several that find significant results, even if all the studies are methodologically sound. Why? Because you are dealing with human beings, very complex organisms who won't necessarily react the same way to the same treatment. Even the same person won't necessarily react the same way to the same treatment at different times.

So, a single study on something like magnets and pain relief should rarely be taken by anybody as significant scientific evidence of a causal connection between the two. Likewise, a single study of this issue that finds nothing significant should not be taken as proof that magnets are useless. However, when dozens of studies find little support that magnets are effective in warding off pain, then it seems reasonable to conclude that there is no good reason to believe in magnet therapy (Carroll 2003: 209).

5.3 Prospective studies

Another way to test the causal claim that *crystalline silica causes lung cancer* would be to study a large random sample of humans. Divide the sample into two groups: those who have been exposed to significant amounts of silica and those who have not. Compare the lung cancer rates of the two groups over time. If *crystalline silica causes lung cancer*, we should find that the lung cancer rate is significantly higher in the exposed group.

This type of study is called a **prospective study,** and it can only be done to test a substance that is already present in the population. Thus, we could not do a prospective study on a new drug or chemical. Only substances that have been present in an environment long enough to have an effect can be tested by the prospective method.

One drawback to the prospective study is that there is no control over potentially relevant causal factors other than the one being tested. For example, we already know that smoking, coal dust, and environmental pollution are significant causal factors in lung cancer. In a controlled experiment, one has the advantage of being able to systematically control for such factors, called **X-factors**, which might be causing the effect being studied. We make sure that our control and experimental groups are alike in lifestyle, age, family health history, weight, etc. However, in a prospective study, our sample is chosen at random. Thus, it is possible that we could get a disproportionate number of smokers in the group that has ingested silica. If so, any difference in lung cancer rates between that group and the one which has not ingested silica might be due to smoking, not silica. However, this drawback can be mitigated by using a very large random sample. In a very large random sample, potential causal factors other than silica (**X-factors**) are likely to be evenly distributed amongst those exposed to silica and those not exposed to silica. If there were only 100 people in the sample, by chance they might all be coal miners or all live in heavily polluted industrial areas or be smokers. If I study 100,000 randomly selected people, the odds are that not all those in the silica group live in heavily polluted areas or smoke, while all those in the non-silica group live healthy lives in pristine countryside villages. It is more likely that potential X-factors will be evenly distributed among both the silica and non-silica populations in a very large random sample.

Having to study very large samples has some obvious disadvantages besides not being able to control for X-factors. Such studies may be time-consuming and costly, as it may take years to collect and analyze the data. However, there are some advantages to a prospective study over the controlled experiment. One advantage is that it allows us to study the effects of substances on human beings without doing experiments on humans or other mammals. In addition, for substances suspected of being health hazards, which have been in use for many years, prospective studies can be done in a relatively short time. The effects of smoking may take 20 or more years to develop. A controlled experiment on humans to study the effects of smoking would be very difficult and lengthy, as well as immoral. Whereas, the American Cancer Society's famous prospective study (1952) on the effects of cigarette smoking took only a few years to complete because large numbers of men had been long-term smokers (i.e., had smoked for 20 years or more). In that study, the sample included about 200,000 men between the ages of 50 and 69. Women were not included in the sample because it was believed that there would not be very many women who had been long-time smokers. The study found that the mortality rate from lung cancer for persons who smoke one to two packs of cigarettes a day is 11.5 times greater than for those who do not smoke. Although women were not included in the sample, it would have been reasonable to conclude that similar results would be obtained from a large sample of long-term women smokers. Why? Reasoning analogically, we would argue that since women are like men in many relevant respects regarding the issue at hand, namely, in having the same kind of respiratory system, we would expect that it is probable that what is true of men for smoking and lung cancer will also be true of women. Nevertheless, it was less certain, from this study, that women who smoke run the same risks as men who smoke—simply because women were not included in the study. For all we know, there may be some relevant difference in lifestyle or physiology of women that would make a difference. In fact, however, later studies of the effects of smoking on women have demonstrated that women are susceptible to the same kinds of health hazards from smoking as men are.

A later and larger prospective study (685,748 women and 521,555 men, begun in 1982) by the American Cancer Society found mortality risks among current smokers are higher than those among nonsmokers. In all, it is estimated that cigarette smoking causes approximately 23 percent of all cancer deaths in women and smoking is responsible for 42 percent of all male cancer deaths (Shopland et al. 1991). The risk of developing any of the smoking-related cancers is dose-related; that is, the more cigarettes consumed daily, the younger the age at which one initiates smoking, and the more years one smokes, the greater the risk. This kind of data strengthens the case against smoking. Among male cigarette smokers, the risk of lung cancer is more than 2,000 percent higher than among male nonsmokers; for women, the risks were approximately 1,200 percent greater.

Prospective studies are useful in testing causal hypotheses regarding substances that have been widely used for many years. However, for substances that have only recently been introduced into human populations, many years will have to go by before the effects on humans can be known by prospective studies. Since an experiment on laboratory animals can yield results in a relatively short time, it seems inevitable that such experiments will continue to be used to test the potential harmful effects of new substances introduced into the human ecological system.

Mill's Method of **concomitant variation** is often used in designing a prospective study. For example, suppose you wanted to test the claim that *lack of calcium is a cause of leg cramps*. One could do a prospective study, which would involve getting a very large random sample from the general population. We would need dietary information and the incidence and severity of leg cramps of those surveyed. From the dietary information, we could determine the amount of calcium a person ingests. We would predict that if lack of calcium causes leg cramps there would be a correlation between the amount of calcium people ingest and the incidence of leg cramping. We would expect that as the amount of calcium in the diet increases the amount of leg cramps would decrease and as the amount of calcium in a diet decreases, the amount of leg cramping should increase. On a graph, our predicted data would look like this:

2.5.3 Retrospective studies

If the data matches our prediction, we can say we have *confirmed* our hypothesis, but we should not assert that we have proved that lack of calcium causes leg cramps. A single study should not be taken as proof of anything. We need replication under critically examined conditions.

5.4 Retrospective studies

Retrospective studies *are those done on populations already demonstrating the effect whose cause is being sought.* In a retrospective study, we compare a group demonstrating the effect to a group which does not demonstrate the effect. We must have some idea as to what is the cause of the effect, and we test our idea by looking to see if there is a significant difference in the presence of the suspected cause between the two groups. The significance of the study depends heavily upon the two groups being similar in relevant respects.

For example, imagine that it is known that the lung cancer rate among workers at a granite-quarry is very high. We might suspect that silica is the cause of many of the lung cancers. After all, such workers are exposed to silica dust and some animal studies have indicated that silica causes lung cancer. To do a retrospective study, we must first find a comparable group of people who do not have lung cancer (that is, they do not show the effect). This comparable group must be like the group showing the effect in all relevant ways except for having lung cancer. They should be in the same age group, have the same kind of smoking rates, live in similar environments, have similar family health histories, and the like. If *silica causes lung cancer*, we should find that the group with no lung cancer has been exposed to significantly less silica than the lung cancer group. If it turns out, that the no lung cancer group has been exposed to similar quantities of silica as the lung cancer group, then it is probably the case that silica does not cause lung cancer. Also, if we were to compare workers at the quarry who have lung cancer with those who don't and were to find that all of those who have lung cancer also smoke, while none of those who don't have lung cancer smoke, this would indicate that smoking, not silica, is likely the cause of the lung cancers.

One advantage of the retrospective study is that often it can be done much more quickly than an experimental or a prospective study. One drawback with the retrospective method is that samples are not randomly selected, leaving open the possibility of a variety of X-factors as potential causes.

For example, imagine that 35 people out of 100 at a company picnic get food poisoning. To discover whether it was the potato salad, for example, that caused the illness, we couldn't take a group of people and feed them the salad and compare them with those who don't eat any potato salad. To do so would be impractical as well as immoral. We could, however, poll each of the 100 people and ask them what they ate. If we discovered that there were a total of six different food items at the picnic and that only people who ate the potato salad got sick, we might justifiably conclude that the potato salad was the causal factor of the illness. By Mill's joint method of agreement and difference we reason that if something is a cause of something else, we expect the cause to be present where the effect is present—the agreement—and we expect to find that the effect is not present in a different group where the cause isn't present—the difference. We could be wrong, of course, in concluding that the potato salad was the causal factor of the illness. Something unknown to us could have caused the sickness.[16] Nevertheless, I think under such circumstances we would be wise to dispose of the potato salad.

Another example of a retrospective study is discussed in Giere's *Understanding Scientific Reasoning* (1996: 297-303). A few years after the use of birth control pills became widespread in the United States and England, it was noticed that blood clots among young women seemed to be on the rise. To determine whether or not there was a significant or chance correlation between taking birth control pills and blood clotting would have taken years using either the experimental or the prospective design method. Therefore, a retrospective study was made of young married women who had recently been hospitalized for blood clots. It was found that 45 percent of them had been taking the pill. This group was compared to a control group of married women matched for age, number of children and a few other things thought to be possibly relevant to developing blood clots. Each of the control group women had been recently admitted to a hospital for a serious medical problem other than blood clotting. It was found that only 9 percent of the control group were on the pill. The difference between the two groups seems significant but it isn't. The researchers failed to consider the smoking habits of each group. The difference between the two groups turned out to be due more to smoking than to the birth control pill.

Scientists use specific methods to test causal hypotheses to avoid self-deception and false cause reasoning. Yet, even experts sometimes forget that correlations by themselves do not prove causal connections and that *experimental and control groups must be alike in all relevant respects except for the tested factor*. In the 1970's the U.S. Government issued a warning to women using birth control pills. The warning asserted that studies had shown that for women between the ages of 40-44 the death rate from heart attacks was 3.4 times higher for women on birth control pills than for those not on the pill. Since smoking was known at the time to be a relevant causal factor in heart attacks, it should have been controlled for. In fact, when the women were divided into smokers and non-smokers, something quite startling was discovered: Smokers on the pill die from heart attacks at a rate 3.4 times greater than women smokers not on the pill and smokers on the pill die from heart attacks at a rate 8.4 times greater than women non-smokers either on or off the pill. Taking the pill does not seem to be a causal factor in death by heart attack *unless a woman also smokes*, in which case the risk of dying of a heart attack is 4.2 times as

great as for sister smokers not on the pill. (Note: later studies have not found that women who smoke and use birth control pills are at any higher risk of heart attack than women who only smoke.)

In another study, it was found that 84 granite-quarry workers in Vermont had died of lung cancer. Those who did the study concluded that exposure to silica caused the lung cancers. However, other researchers obtained smoking histories of those in the study and found that all 84 of those who died from lung cancer were smokers. Thus, perhaps none of those who got lung cancer got it from exposure to silica.[17]

6. Importance of testing under controlled conditions

Anybody can claim their product works wonders. But not everybody puts their product to the test. In fact, many new products that promise relief from pain or instant happiness are never tested at all. The money behind the product has all gone into marketing. Little, if any, is spent on research. That is why it is important before investing your money in a new product to find out if the product has been tested, how it was tested, and what were the results. For example, the DKL LifeGuard Model 2, from DielectroKinetic Laboratories, was advertised as being able to detect a living human being by receiving a signal from the heartbeat at distances of up to 20 meters through any material. This would be a great tool, if it really works, for law enforcement agencies. Sandia Labs tested the device using a double-blind, random method. (Sandia is a national security laboratory operated for the U.S. Department of Energy by the Sandia Corporation, a Lockheed Martin Co.) The causal hypothesis they tested could be worded as follows: *the human heartbeat causes a directional signal to activate in the Lifeguard, thereby allowing the user of the LifeGuard to find a hidden human being (the target) up to 20 meters away, regardless of what objects might be between the LifeGuard and the target.*

The testing procedure was quite simple: five large plastic packing crates were set up in a line at 30-foot intervals and the test operator, using the DKL LifeGuard Model 2, tried to detect in which of the five crates a human being was hiding. Whether a crate would be empty or contain a person for each trial was determined by random assignment. This is to avoid using a pattern that might be detected by the subject. The tests showed that the device performed no better than expected from random chance. The test operator was a DKL representative. The only time the test operator did well in detecting his targets was when he had prior knowledge of the target's location. The LifeGuard was successful ten out of ten times when the operator knew where the target was. It may seem ludicrous to test the device by telling the operator where the objects are, but it establishes a baseline and affirms that the device is working in the eyes of the manufacturer. Only when the operator agrees that his device is working should the test proceed to the second stage, the double-blind test. For, the operator will not be as likely to come up with an ad hoc hypothesis to explain away his failure in a double-blind test if he has agreed beforehand that the device is working properly.

If the device could perform as claimed, the operator should have received signals from each of the crates with a person within but no signals from the empty crates. In the main test of the LifeGuard, when neither the test operator nor the investigator keeping track of the operator's results knew which of five possible locations contained the target, the operator performed poorly (six out of 25) and took about four times longer than when the operator knew the target's location. If human heartbeats cause the device to activate, one would expect a significantly better performance than 6 of 25, which is what would be expected by chance.

The different performances—10 correct out of 10 tries versus 6 correct out of 25 tries—vividly illustrates the need for keeping the subject blind to the controls: it is needed to eliminate self-deception. The evaluator is kept blind to the controls to prevent him or her from subtly tipping off the subject, either knowingly or unknowingly. If the evaluator knew which crates were empty and which had persons, he or she might give a visual signal to the subject by looking only at the crates with persons. To eliminate the possibility of cheating or evaluator bias, the evaluator is kept in the dark regarding the controls.

The lack of testing under controlled conditions explains why many psychics, graphologists, astrologers, dowsers, paranormal therapists, and the like, believe in their abilities. To test a dowser it is not enough to have the dowser and his friends tell you that it works by pointing out all the wells that have been dug on the dowser's advice.

One should perform a random, double-blind test, such as the one done by Ray Hyman with an experienced dowser on the PBS program *Frontiers of Science* (Nov. 19, 1997). The dowser claimed he could find buried metal objects, as well as water. He agreed to a test that involved randomly selecting numbers which corresponded to buckets placed upside down in a field. The numbers determined which buckets a metal object would be placed under. The one doing the placing of the objects was not the same person who went around with the dowser as he tried to find the objects. The exact odds of finding a metal object by chance could be calculated. For example, if there are 100 buckets and 10 of them have a metal object, then getting 10% correct would be predicted by chance. That is, over a large number of attempts, getting about 10% correct would be expected of anyone, with or without a dowsing rod. On the other hand, if someone consistently got 80% or 90% correct, and we were sure he or she was not cheating, that would confirm the dowser's powers.

The dowser walked up and down the lines of buckets with his rod but said he couldn't get any strong readings. When he selected a bucket, he qualified his selection with words to the effect that he didn't think he'd be right. He was right about never being right! He didn't find a single metal object despite several attempts. His performance is typical of dowsers tested under controlled conditions. His response was also typical: He was genuinely surprised. Like most of us, the dowser is not aware of the many factors that can hinder us from doing a proper evaluation of events: self-deception, wishful thinking, suggestion, unconscious bias, selective thinking, and communal reinforcement (Carroll 2003: 82).

7. A word about statistical significance and meta-analyses

I noted above in the discussion of the Jahn PEAR studies that statistical significance is not necessarily important. Let me explain. The most common meaning of 'statistical significance' when applied to the results of a scientific study is "according to an arbitrary statistical formula, the results obtained are not likely due to chance." Most of the time, a formula is used that translates into layman's terms as: *One out of twenty such tests will yield spurious results.* However, scientists never put it that way. They will tell us that **p<0.05** (p = the probability of the results being spurious) or that "We observed a decrease of 5.1 percent (95 percent **confidence interval**, 1.5 to 8.7 percent; p = 0.02)." What they really mean is that 5 percent of the time, doing exactly the same study in exactly the same way, one can expect to get results that look significant when they're not. In short, one out of twenty scientific studies is wrong (Brignell 2000: 52).

How did scientists land upon this magical 95 percent confidence interval? The idea came from R. A. Fisher, one of the pioneers of statistics as a science. How did he land upon the significance of a probability of 0.05? There isn't any profound reason for it. He just found the number convenient (Brignell 2000: 52). Many scientists consider it to be the gold standard. Finding a statistical correlation that is significant at the 95 percent confidence level has become the goal of some researchers. And why wouldn't it, since anything less is unlikely to get published in many journals. Although, it is becoming more frequent to see studies claiming significance at the 90 percent confidence level, which means that one in *ten* such studies is likely to yield spurious results. The Environmental Protection Agency, for example, was willing to declare that secondhand smoke causes 3,000 lung cancer deaths a year on the basis of a study that used the 90 percent confidence interval.

If you think about this for a short time, you will see that there are some significant problems with this standard. A scientist who does ten or twenty experiments can expect at least one of them to yield significant results according to some arbitrary statistical formula. Guess which study the scientist writes up and submits to a scientific journal for publication? Not the nine or nineteen that found no significant correlation. Those get filed in the drawer. This is known as the *file-drawer effect* or *publication bias*. Also, the media tend to publish stories about studies that show positive results. This explains in part why we rarely see a follow-up to all the stories about the latest wonder drugs or cures that are hyped when someone publishes a paper after having found a statistically significant correlation. It also explains why we should not put too much faith in the results of any single study. Publication bias also makes it possible for some drugs to get marketed that have scientific support that they work, when in fact they may be little more than placebos.

Finally, when one combines the trend among scientists to do *meta-studies* (combining the results of several studies done by different scientists and treating them as if they comprised one large study) with the file-drawer effect, one has a recipe for exaggerated significance as well as exaggerated importance. Considering the fact that there may be many studies on a particular subject that have never been published because they got negative results, it is likely that a meta-study of that subject will have a disproportionate and unrepresentative collection of studies to analyze. And, even though researchers try not to mix heterogeneous studies, it is very difficult to find dozens of studies that all used the same subject selection criteria, employed the same treatments, and used the same statistical methods.

As scientist and author John Brignell says, "The safest thing to do with meta-analyses is take them with a large pinch of salt."

Exercise 8-1 Self-test: true or false? (Check your answers in Answers to Selected Exercises.)

1. A controlled experiment has the advantage of allowing us to systematically control for factors that could be causing the effect being studied.
2. One drawback to the prospective design is that it takes many years to perfect.
3. A necessary condition is one that necessarily produces an effect.
4. A sufficient condition is one that is sufficient to produce an effect
5. Publication bias refers to the tendency of scientific journals to publish articles that fit with the editors' biases and prejudices.
6. In a retrospective study, we compare a group demonstrating the effect to a group that does not demonstrate the effect.
7. To divide a large random sample into a group that has been exposed to a substance that might be carcinogenic and a group that has not been exposed to the substance, would be done if one were using the prospective method.
8. An experimental group is one that is introduced to the suspected causal factor, to be compared with another identical group not introduced to it.
9. If a sample from a very large human population is selected properly, the large random differences of opinions and behaviors one would expect to find in that population should tend to cancel out one another.
10. Causal relationships are regular patterns in nature that are characterized by sequences of events and significant correlations.
11. If a single study finds a statistically significant correlation between two variables, we should take that as proof of a causal connection between the variables.
12. It is impossible for non-causally related phenomena to exhibit patterns of regularity of behavior.
13. Each of the empirical methods of testing causal claims presupposes certain logical methods of analysis as well.
14. To say that something is a relevant causal factor in the development or production of something is to say that it is either part of or wholly a necessary condition, a sufficient condition, or both a necessary and sufficient condition for the effect to occur.
15. The *post hoc* fallacy is a type of sampling error due to carelessness after the fact.
16. All perfectly correlated events are causally related.
17. All causally related events are significantly correlated.
18. One disadvantage to the retrospective study is that samples are not randomly selected.
19. In causal reasoning, if done scientifically, there is little possibility of error and conclusions regarding causes are almost absolutely certain.
20. Replication of a scientific study proves the original results were correct.
21. Correlation proves causality.
22. Double-blind studies are done to minimize experimenter and test subject bias.
23. A single study demonstrating a significant correlation is rarely sufficient to justify belief that a causal connection has been established.
24. The results of studies on animals such as mice are completely irrelevant for humans.
25. An ad hoc hypothesis is a claim created to explain away facts that seem to refute one's belief.

Exercise 8-2

Each of the following consists of a base argument. Look at the base argument and get a general notion of the sufficiency of the premises for the stated conclusion. Following each base argument is a list of statements. The statements either change the strength with which the conclusion is asserted to follow from the evidence or they offer alternative premises. Treating each additional statement separately, determine whether it strengthens, weakens, or has no effect on the argument. (Note: while each of the following passages is based on an actual experiment, the data have been changed for the purposes of this exercise.)

*1. Over a period of several years, 100 rats were fed a diet which consisted of 7 percent saccharin (the experimental group). Another 100 rats were fed the same diet except for the saccharin (the control group) . The offspring of each group were put on the same diet as their parents. The rats used in the experiment were all carefully bred laboratory rats (i.e., nearly identical "twins"). Both male and female rats we used in equal numbers in the first generation. Bladder cancer was discovered in 3 percent of the first-generation experimental group. No first-generation control group rat developed bladder cancer. In the second generation, 14 percent of the experimental group and 2 percent of the control group developed bladder cancer. Therefore, saccharin very likely causes cancer in humans.

1.1 CONCLUSION: Saccharin causes cancer in rats.
1.2 PREMISE: The rats were collected from sewers and fields.
1.3 PREMISE: The experimental group members were from the sewers, but the control group members were carefully bred laboratory rats.
1.4 PREMISE: All the experimental rats developed bladder cancer.
1.5 PREMISE: One thousand rats were used in each group.
1.6 CONCLUSION: More studies should be done to determine if there is any causal link between saccharin ingestion and bladder cancer in humans.
1.7 PREMISE: Different lab technicians fed the rats different diets on different days, except for the saccharin, which was always kept at 7 percent of the experimental group's diet and was absent from the control group's diet.
1.8 PREMISE: The experimental group members were frequently x-rayed to see if any tumors were developing.
1.9 PREMISE: The 30 or so substances known to cause cancer in humans are all known to cause cancer in laboratory animals when given in high doses.
1.10 CONCLUSION: Pregnant women should be advised that using saccharin during pregnancy may increase the risk that their child will develop bladder cancer.

2. One hundred weanling rats on a diet of 20 percent Brazilian peanut meal showed liver damage after 9 weeks. After 6 months, 9 of 11 rats developed multiple liver tumors, and two of these rats had lung metastases (i.e., the malignancy spread). Therefore, peanuts are carcinogenic (i.e., cancer-causing).

2.1 CONCLUSION: Peanut meal may cause cancer in humans.
2.2 PREMISE: A control group fed the same diet as the weanling rats, except for the peanut meal, developed no tumors.
2.3 PREMISE: Several experiments, using peanuts from several different countries, produced the same results.
2.4 PREMISE: 800 of 1000 young turkeys died of liver lesions within two weeks of being fed a substance which contained the Brazilian peanut meal.
2.5 PREMISE: Peanut meal often contains aflatoxin and 20 millionths of a gram of aflatoxin will kill a day old duckling in 24 hours; and CONCLUSION: Studies should be done to determine the frequency of aflatoxin in peanuts and peanut-based foods. Note:This issue is discussed in detail in *Man Against Cancer*, Bernard Glemser (New York: Funk & Wagnalls, 1969).

3. A group of 50 sterile laboratory rats were being kept as pets by a lab technician. No consistent diet was given to the rats, but what each individual rat ate each day was recorded. 25 of the rats developed liver cancer. An analysis was made of each rat's diet, and it was discovered that each of the cancerous rats had consumed .5 milligrams of sodium nitrite each day for a year. Therefore, sodium nitrite probably causes liver cancer in humans.

3.1 PREMISE: The 25 non-cancerous rats consumed no sodium nitrite during the year.

3.2 PREMISE: Excluding water, ground rice and spinach, the only substance each of the cancerous rats ingested in common was the sodium nitrite.

3.3 PREMISE: Other studies have linked sodium nitrite to liver cancer in laboratory animals.

3.4 PREMISE: In another study, an experimental and a control group are fed the same diet, except for .5 milligrams of sodium nitrite given daily to each of the members of the experimental group. 40 percent of the experimental group develop cancer, while only 38 percent of the control group develop cancer.

Exercise 8-3

Evaluate the causal reasoning in the following passages. If reasonable, suggest at least one experiment to test the conclusion made in each argument (i.e., treat the conclusion as a hypothesis).

*1. I know Vanish-X Cream eliminates acne because last night I used some and today my skin is much clearer.

2. Leg cramps are caused by lack of calcium. I know because last year when I had leg cramps I started taking calcium supplements and within two weeks my cramps were gone.

3. Friends, do you want to get rich like Brother Billy Bob? Just send a card with your contribution to the Reverend I. Am Greedy like our listener Mrs. Goodfaith did. She writes, "Dear Brother Billy Bob: On the very day I sent you my check for $100, I found a $1000 check that I had misplaced months ago." Praise the Lord!

4. In our neighborhood, a recent survey revealed that 75 percent of the men who had lost their jobs in the last year had quit going to church before losing their jobs. That should prove to even the greatest skeptic that it does not pay to leave the church.

5. I never shave before I pitch a ball game because the last time I shaved before a game, I gave up 14 runs in the first inning!

*6. It's all those food additives that are causing the crazy criminal acts to increase. You check it out. There is a direct correlation between the increase in the use of food additives and the increase in senseless crimes of violence.

7. The inflation rate went down after President Reagan took office, so his economic policies must have caused it.

8. Since Senator Rodda has been in office the crime rate has gone up 10 percent a year. It's time to elect someone who will do something positive about crime!

9. Andy, Beth, Carol, Dave, Edie and Frank had dinner together. All six developed violent stomach cramps within 30 minutes of finishing their meal. The meal consisted of salad with vinegar and oil dressing, baked potatoes with sour cream and chives, broccoli, broiled sirloin steak, and red wine. Edie is on a diet and refused the potato, but she put sour cream on her steak. Andy and Beth don't drink wine. Carol is a vegetarian. Dave and Frank hate broccoli. Therefore, the sour cream was a causal factor in their stomach cramps.

*10. In the 18th century, millions of people died from smallpox. Edward Jenner, a country doctor of Gloucestershire, spent 20 years investigating the cause of small pox and trying to establish that inoculation with cowpox immunizes a person from smallpox. Cowpox is a much milder disease than smallpox. Jenner heard from local folk that those infected with cowpox never got smallpox. He documented about 20 cases of people infected with cowpox who, when inoculated with smallpox were unaffected. It was well-known that others who were infected with smallpox were always affected unless they had already had the disease. He concluded that cow-pox immunizes human beings against smallpox. See *A History of the Sciences*, Stephen F. Mason (New York: Collier Books, 1962), p. 519; see also Langley, *op. cit.*, pp. 93-94.

11. "Joyce Kenyon, a San Jose, Calif., computer-systems manager, thanks her [Zuni] hummingbird fetish for saving her neck on more than one occasion. Recently, in an `intuitive hit', it cautioned that she might want to assure that a backup to her company's computer system was in place. `It's a good thing I did the backup work because the computer crashed the next day.' she said." (L. A. Winokur, "Pushing Their Luck: Zuni Indians Peddle `Magical' Charms," *The Wall Street Journal*, April 28, 1993.)

12. A recent study found that men under 5-foot-7 had about 70 percent more heart attacks than those over 6-foot-1. The study was based on a survey of 22,071 male doctors from across the United States. It was found that for every inch of height a person's heart attack risk goes down 3 percent. This means that someone 5-foot-10 is 9 percent less likely than someone 5-foot-7 to suffer a heart attack. The researchers found that shorter men were more likely to be overweight and to have high cholesterol and blood pressure. but even when these factors were taken into consideration, their risk of heart attacks was still higher than taller men's. Another study found a similar link between height and heart attacks in women. Short people might be at risk because their blood vessels are skinnier, so they are more prone to becoming clogged.

13. "Criminals convicted of murder, rape and other violent crimes have significantly higher levels of the metal manganese in their hair, say brain researchers with the University of California, Irvine. The finding might indicate that such prisoners suffer from a metabolic disorder that affects the brain, said Dr. Monte Buchsbaum, a professor of psychiatry and one the study researchers. People who suffer manganese poisoning from industrial exposure develop a syndrome similar to Parkinson's

disease, he said. That disease causes shaking of the hands and loss of fine motor control consistent with damage to the brain's basal ganglia, a switchboard-like area that coordinates movement with sensory information.

"Either the prisoners are being exposed to higher level of manganese or they could have a disorder that results in higher levels of manganese," Buchsbaum said.... "Whether the finding is related to the environment might be significant for the general population because manganese is the metal that has been used to replace lead in gasoline. This alerts us not to an immediate danger but to the need to know more about the neurological effects of manganese," Buchsbaum said.

The study initially looked at prisoners at Deuel Vocational Institution in San Joaquin County who had been convicted of violent crimes. Guards and townspeople were also tested as comparison groups. Although all had fairly low level of manganese, the prisoners had seven times the amount in their hair as townspeople and five times that of guards.

Two more groups of prisoners, guards and locals were then studied. The second group included convicts in San Bernardino County Jail. The third group included recently arrested inmates awaiting trial in Los Angeles County Jail in an effort to rule out long-term exposure to the jail environment as a factor. While the Los Angeles inmates had substantially lower levels than the other prisoners, they still had double the amount found in guards and local residents.

The scientists said they needed to do more studies, perhaps with other aggressive people who have not been jailed, such as boxers or other athletes. ("Metal found in criminal's hair," by Susan Peterson, *The Sacramento Bee*, April 12, 1992. The article first appeared in the *Orange County Register*.)

14. For some young men, clinical depression--a state of feeling sad and without hope--"may have more to do with drug abuse than with deep psychological problems or a genetic predisposition to depression." Psychiatrist Marc Schuckit of the University of California at San Diego studied 964 men between the ages of 21 to 25 who were affiliated with the university. He found that while 82 percent had never been depressed, 11 percent had been. An additional 7 percent had depression that seriously interfered with their lives....Schuckit reported that only 30 percent of the never-depressed young men had problems connected with drug or alcohol use (job loss, arrest, ill health, marital disruption), compared with half of the most seriously depressed group. Moreover, a majority of the depressed men said that their drug-related problems had preceded their depression. (Only about a quarter said that they had become depressed first.) ....[S]tudents would do well to consider changing their drug habits before they come to the conclusion that they are hopelessly depressed." (Phillip Shaver, *Psychology Today*, May 1983, p. 16)

15. Public school administrators across the country have been cutting costs by increasing the number of pupils per class, citing studies that assert such changes will not harm the quality of education. Teachers cite other studies that show just the opposite. An exhaustive new survey of all the research to date shows mixed results for both sides. Achievement levels, it seems, increase significantly in classes with fewer than 20 students to each adult, but for classes larger than 20, achievement levels remain relatively constant even if classes swell to over 60....Gene Glass and Mary Lee Smith, psychologists with the Laboratory of Educational Research at the University of Colorado at Boulder, turned up 80 different controlled studies of the effects of class size....Classes in those studies ranged from tutorials for just a few students to packed classrooms of more than 60. Teachers' aides were counted as teachers in determining the ratio of students to adults. Glass and Smith found that both the negative and positive effects of class size were slightly greater in secondary school than in elementary school; the subject matter also made little difference....Other things being equal, pupils in a class of 15 will achieve more than those in a class of 30. Reducing class size from 30 to 25, however, will barely bring a noticeable change, according to the research.

*16. "[Dr. Ian James of the Royal Free Hospital in London wanted to find out whether oxyprenolol, a member of the beta-blocker group, might possibly alleviate anxieties and thus allow otherwise competent drivers [of automobiles] to perform to their full capabilities [during driving tests, which they had failed because of excessive nervousness].... Beta-blockers do exactly what the name implies. They block the beta receptors on the surface of cells, preventing adrenaline from binding on these sites--a reaction that prolongs and intensifies the effects of a stressful situation....Thirty-four healthy young string players (not selected for undue nervousness) performed on separate days after receiving oxyprenolol or a placebo [an inert compound]. Their playing was then assessed by two experts who did not know which of the two compounds the musicians had received. The aim of the experiment was to determine the effect of what Dr. James termed `stage fright'--the natural anxiety and stress of performing in public. He chose string players because he felt the adverse effects of tremor would be more noticeable in them....As reported in *The Lancet* [1977, Vol. II, p. 952], the outcome was striking. Musical quality improved significantly--especially on the first occasion when players took oxyprenolol. All aspects of their playing improved: right- and left-hand dexterity, intonation, and control of tremor. Although the overall mean improvement was only 5 percent, some subjects registered 30 percent, and one registered 73 percent. As the musicians were not selected for being particularly prone to anxiety, the results suggest that some people might benefit greatly from such medication." [Dr. Bernard Dixon, "Ill-Defined Parameters," in *Omni*, July 1979, pp. 29-35.]

17. Irwin K.M. Liu, associate professor of reproduction at the University of California, Davis, School of Veterinary medicine, was searching for a method to prevent infertility in domestic horses when he came upon an alternative solution to the problem

of wild horse overpopulation which is leading to overgrazing of public range lands. Currently, the government either slaughters the horses or lets people adopt them. Liu was aware that "a naturally occurring antibody found in some women blocks penetration of an egg by sperm." He reasoned that the same phenomenon might well occur in horses. He developed a vaccine which stimulates the antibody in horses. "The horses were turned loose with a fertile stallion, and when they were recaptured almost a year later, nine out of the ten mares had not conceived. The tenth was in the very early stage of pregnancy, Liu said, indicating the antibody had worn off." ("Davis professor invents, test birth-control vaccine for horses," by Kathryn Eaker Perkins, *The Sacramento Bee*, Aug. 1, 1984, B1.)

18. "A study of 900,000 Americans confirmed that the 40% to 50% of adults who have quit smoking sharply cut their risk of dying of lung cancer." Researchers at the University of Michigan analyzed health and lifestyle data of the group over a period of six years. "About half of [the 900,000] had never smoked tobacco, another quarter had quit smoking and the remaining fourth continued to smoke....The analysis showed that among each 100,000 persons in the population, fewer than 50 of those who had never smoked died of lung cancer by the age of 75. Among those who continued to smoke, the death rate from lung cancer by age 75 was about 1,250 per 100,000 men and 550 per 100,000 women....Among the men who quit smoking in their 30's, the death rate from lung cancer by age 75 was about 100 per 100,000 persons in the population, or only 7% of that of men who continued to smoke. Among women who quit in their 30's, the lung cancer death rate by age 75 was about 50 per 100,000, or about 10% that of smokers....If the smoker waited until his or her early 60's to quit smoking, the chances of dying of lung cancer by age 75 were about half those of smokers. Specifically, of every 100,000 persons in the population, about 550 men and 250 women who waited until their early 60's to quit smoking died of lung cancer by age 75....But, ...even those who quit in their 30's have a risk approximately twice that of those who have never smoked, a difference that does not decrease with age." ("When a Smoker Quits May Determine Chances of Dying From Lung Cancer," by Jerry E. Bishop, *The Wall Street Journal*, March 17, 1993, B7.)

19. A study by the Drug Safety Research Unit in Southampton, England, studied 448 women between the ages of 16 and 44 who had suffered heart attacks. They matched them by age and region with 1,728 women who had not had heart attacks. 13 percent of those who had heart attacks used birth control pills. 15 percent of those who did not have heart attacks used birth control pills. 80 percent of those who had heart attacks smoked, while only 30 percent of those who did not have heart attacks smoked. The study concluded that "there is no increased risk [of heart attack] associated with taking the oral contraceptive." The researchers also found that "among those who smoked, taking the pill did not further increase the chance of heart attack." ("Study: Birth control pills not linked to heart attacks," by Emma Ross, Associated Press, *Sacramento Bee*, June 11, 1999.)

20. "In a study of more than 1,700 North Carolina adults age 65 or older, Duke University researchers found that those who attend religious services at least once a week have healthier immune systems than those who don't." The study measured blood samples for levels of interleukin 6 (IL6) and other substances that regulate immune and inflammatory responses. (Those with AIDS, Alzheimer's, osteoporosis, and diabetes have high levels of IL6, which is also associated with stress and depression.) Those who attended religious services once a week were about half as likely as those who didn't to have elevated levels of IL6.

Dr. Harold Koenig, director of Duke's Center for the Study of Religion/Spirituality and Health, was the lead author of the study. Koenig is a psychiatrist and family doctor. The study was funded by the National Institute on Aging and was published in the *International Journal of Psychiatry in Medicine*. Said Dr. Koenig: "Maybe believing in [a] higher power…could be a strong key to people's health. We can actually show that these immune systems are functioning better." He thinks that people who go to church are "probably able to handle stress better. They are significantly less likely to be depressed." He also said, "We think that both the social aspects and the faith aspects of having a belief system help them cope." Dr. Koenig cautioned that the results might show a regional influence, since participants are from the Bible Belt South, "where religion is ingrained in the social fabric."

Chaplain David Carl claimed that the study supports the notion that "our beliefs show up, right now, in our biology." ("Attending church boosts immune system, study says," by Karen Garloch, Knight-Ridder Newspapers, *Sacramento Bee*, October 24, 1997.)

Holly Nelson, in a letter to the editor, asked, "Was it ever considered that these people [attending church] already have healthier immune systems, or lead healthier lives, than those who cannot attend?" She also questioned whether the important thing was getting up and moving about, rather than faith or belief. "This incomplete study said almost nothing of importance," she wrote.

Exercise 8-4

Find several articles from newspapers or magazines that are based on causal reasoning. Evaluate the reasoning in the articles. Before you begin your evaluation, you might find it useful to state the main conclusion(s) of the article and the main premises. Evaluate the size of samples and the methods of selecting samples.

Further Reading - Chapter Eight

Alcock, James E. et al. (2003). Eds. *Psi Wars – Getting to Grips with the Paranormal*. Imprint Academic.

Brignell, John (2000). *Sorry, Wrong Number! The abuse of measurement*. Brignell Associates.

Carroll, Robert T. (2003). *The Skeptic's Dictionary: A Collection of Strange Beliefs, Amusing Deceptions, and Dangerous Delusions*. Wiley & Sons.

Hyman, Ray (1989). *The Elusive Quarry – A Scientific Appraisal of Psychical Research*. Prometheus Books.

Giere, Ronald (2004). *Understanding Scientific Reasoning*. 5th ed. Wadsworth.

Gilovich, Thomas (1993). *How We Know What Isn't So*: *The Fallibility of Human Reason in Everyday Life*. The Free Press.

Radin, Dean (1997). *The Conscious Universe – The Scientific Truth of Psychic Phenomena*. HarperSanFrancisco.

Rosenthal, Robert (1998). "Covert Communication in Classrooms, Clinics, and Courtrooms," *Eye on Psi Chi*. Vol. 3, No. 1, pp. 18-22.

Shopland, D.R., Eyre and Pechacek (1991). "Smoking-attributable mortality in 1991. Is lung cancer now the leading cause of death among smokers in the United States?" *Journal of the National Cancer Institute*. 83(16):1142-1148.

Notes – Chapter Eight

[1]"Cancer scare: How Sand on a Beach Came to Be Defined As a Human Carcinogen," by David Stipp, *Wall Street Journal*, March 22, 1993, p. A4.

[2] Establishing whether a correlation is likely due to chance depends, in part, on the size of the sample, the method of selecting the sample, and the statistical formula one uses. This text is intended as an introduction to the material presented and is purposely non-technical. However, I suggest you do further reading on the matter of statistically significant correlations in either a statistics text book or in an excellent book on critical thinking in the sciences: *Understanding Scientific Reasoning*, 5th ed., Ronald N. Giere (Wadsworth, 2004).

[3] Martin, Bruce. "Coincidences: Remarkable or Random?" in *Skeptical Inquirer*, September/October 1998.

[4] Gawande, Atul. "The Cancer-Cluster Myth," *The New Yorker*, February 8, 1999, pp. 34-37.

[5] *ibid.*

[6] Gilovich, T., R. Vallone, and A. Tversky, "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17, 295-314.

[7] Gawande, loc. cit.

[8] Tversky, A. and D. Hahneman (1971). "Belief in the law of small numbers," *Psychological Bulletin*, 76, 105-110.

[9] To assume that the whole must be exactly like the parts is to commit the *fallacy of composition*. For example, just because a part of a person's brain is not functional, it does not follow that the person's brain itself is not functional. Just because the individual players on a basketball team are exceptional does not mean that the team is exceptional. Likewise, just because a team is exceptionally good does not mean that each of the players is exceptionally good.

[10] An analysis of this study is given in the Congressional Office of Technology Assessment report *Cancer Testing*, and is discussed in Giere, op. cit., pp. 274-284.

[11]"Berkeley study doubts value of cancer tests in mice," Deborah Blum, *The Sacramento Bee*, August 31, 1990.

[12]*ibid.*

[13]"Animal Tests as Risk Clues: The Best Data May Fall Short," Joel Brinkely, *The New York Times*, March 23, 1993,

[14]*ibid.*

[15]*ibid.*

[16]As a farfetched possibility, we might imagine that just the 35 who got ill also drank from a public water fountain which was contaminated. Or, maybe 20 of them were suffering from food poisoning, 10 from intestinal flu, and 5 from psychosomatic symptoms--in which case our assumption that there was a single cause would be in error. The point is that there are numerous possibilities we are likely to overlook.

[17] "Cancer scare: How Sand on a Beach Came to Be Defined As a Human Carcinogen," by David Stipp, *Wall Street Journal*, March 22, 1993, p. A4.