4

Experimentation

THE BASIC METHOD

In science, questions about the truth of claims are frequently settled by observation. Do raccoons salivate? Look and see. Are the glaciers of Northern Europe receding? Careful measurement from year to year should provide the answer. Observation can be indirect, coming via instruments designed to supplement our senses. Is there life on Mars? Send a probe to the planet's surface that can record and transmit its findings back to earth. Does a suspicious bit of food contain dangerous microorganisms? Examine it under the lens of a powerful microscope. Some questions, however, cannot be resolved by simply looking to nature even with the most refined and powerful of scientific tools. Observation, alone, comes up short in dealing with many interesting and important scientific questions. Most fall under one of three broad categories.

First, there are aspects of nature that cannot be observed either directly or indirectly. We know, for example, that humans possess a sense of self-identity. Is this true of other higher mammals such as elephants, dolphins, or chimpanzees? We can't ask them, and no set of observations immediately comes to mind by which we might settle this issue. In the 1930s a physicist, Wolfgang Pauli, proposed the existence of a new and as yet undetected subatomic particle, the neutrino. Such a particle, Pauli speculated, would account for a number of anomalies that had recently emerged from experimental findings about the behavior of the atom's nucleus. But the neutrino, if it existed, would carry no electric charge, be nearly mass free, extremely small, and travel at or near the speed of light. Such a particle would probably be impossible to observe. How, then, could the question of whether there really are neutrinos be answered?

Second, there are questions that pose problems because the available observational evidence is inconclusive. Many anomalous claims like those discussed in Chapter 2 present us with this sort of difficulty. Can dowsers detect

hidden sources of water with nothing more than a forked tree branch? No doubt some dowsers have had some successes, a fact to which many of their clients will attest. But can we be sure that the power of the dowsing rod is responsible for these results? Perhaps successful dowsers are just good at making educated guesses. The available observational data, it seems, provide no clear answer.

Finally, the resolution of questions about proposed explanations will often require information that goes beyond that provided by the available observational evidence. This is because a proposed explanation will "fit" what is known by observation: if correct, the explanation will tell us something about how or why the observational data are true. What we require is some additional piece of evidence that can tell us whether the explanation is right. Common household ants will not, it seems, cross a line drawn in chalk. Why? One possible explanation is that they will not enter an area where matter clings to their feet and this is what small particles of chalk dust tend to do. But is this explanation correct? The fact that ants behave in the way described does not confirm the explanation since this is precisely the phenomenon it is introduced to explain. What we need to find is something in addition to the available observational data to answer this question.

Fortunately, there is a simple and quite ingenious method for putting questions like those above to the test. Richard Feynman, Nobel Prize winner for physics, describes it in an essay, "Seeking New Laws of Nature":[1]

> In general, we look for a new law by the following process. First, we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the results of the computation to nature, with experiment or experience, compare it directly to observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science.

The method Feynman describes can be summarized as follows. First, figure out something that should happen if the claim at issue is correct. In Feynman's example the claim is a proposed law of nature. What should happen, if Feynman's guess is correct, are its predicted "consequences." Next, devise a set of circumstances in which the predicted outcome should occur. Finally, observe what actually happens in "experiment or experience." If the expected outcome occurs, we have evidence for the claim. But if the outcome does not occur, we have evidence the claim must be wrong.

Scientists have, for example, long wondered how songbirds are able to navigate thousands of miles without getting lost during their annual migrations. Do birds have an innate ability to follow migratory paths? Well, if they do—if that ability is not honed by experience—it would seem that young birds on their first migration ought to be able to navigate as well as do older, more experienced birds. To test this prediction, researchers moved a small flock of sparrows 2,300 miles in an easterly direction from the point where they normally began their yearly winter migration. As it turned out, the young birds simply headed south, while the older ones corrected for their displacement and

headed in a west-southwesterly direction toward their wintering grounds. It would seem that the claim at issue is false: migratory abilities among birds—at least among sparrows—are not wholly innate. This simple strategy—making and testing a prediction—is at the heart of the method by which questions that cannot be resolved by observation alone are put to the test in science. It is, in a nutshell, what scientists mean when they speak of *experimental method*.

CONFIRMATION AND REJECTION

Unfortunately, experiments are not always as easy to design and carry out as in our example. The goal of a decisive test is to arrange circumstances under which we can be confident that nothing unforeseen or extraneous can invalidate the experiment's outcome. There are two potential sources of error that can affect an experiment's findings. A well planned experiment will be designed to avoid both.

First, an experiment may overlook factors that can lead to *false confirmation* of the claim at issue. Here is a simple illustration. Suppose I suspect that today is a legal holiday but I'm not certain. My mail is always delivered first thing in the morning and I know that mail is not delivered on holidays. (I'm also pretty sure it's not a Sunday.) At noon, I check my mailbox and find nothing. So I conclude it must be a holiday. But suppose now that the mailman did deliver to my neighborhood earlier today but had nothing for my address. My test, it seems, has allowed me to confirm a claim that is false!

The way to minimize the possibility of a false confirmation is to set up experimental conditions that control for extraneous factors—factors other than the claim at issue that might lead to the predicted result. If, in addition to checking for mail well after the time at which it is normally delivered, I had thought to watch for the mailman throughout the morning, my test would have been on considerably firmer ground. Of course, I can't rule out everything that could lead to a false confirmation. Maybe my regular mailman was sick and her replacement was going to do her route in the afternoon. Maybe her mail truck broke down. But the more we can rule out, the more confident we can be in the results of a test. As a general rule, it is always worth asking of any experimental test: has anything been overlooked which might lead to the predicted outcome, other than the truth of the claim at issue? Unless our answer is in the negative, the test will not enable us to confirm the claim at issue under the conditions we have imposed.

A second source of experimental error involves overlooking factors which can lead to a *false rejection* of the claim being tested. Suppose, I run the test described above, though now I discover mail in my mail box. It would seem that it's not a holiday after all. But what if I had inadvertently failed to check for mail yesterday? If the mail I have found was delivered yesterday, I run the risk of falsely rejecting the claim I am testing! So, of any experimental test, we need to ask a second question: has anything been overlooked that might lead to a failure to get the predicted outcome even if the claim at issue is right?

QUICK REVIEW 4.1 Confirmation and Rejection

A well-designed experiment will control for factors that could lead to a false confirmation or rejection of the claim being tested.

False confirmation: Could the predicted outcome be due to anything other than the claim at issue?

Yes. The experiment cannot verify the claim at issue.

False rejection: Could the predicted outcome fail to occur even if the claim at issue is correct?

Yes. The experiment cannot falsify the claim at issue.

If an experiment is well designed, the answer to both questions will be "No."

As you can imagine, some very tough problems must be solved in designing a good experimental test. The key is to anticipate and then eliminate the possibility of confounding factors—things that might lead us to either falsely confirm or reject the claim at issue. To get a grasp of the problems that may be encountered in designing and carrying out a good experiment, we will turn next to a few examples from the world of science. As we move through the case studies we will want to pay special attention to the kinds of precautions that must be taken to rule out these two very real possibilities. Then in the next chapter we will take a close look at how experimental method plays out in one very common and important type of scientific research—studies designed to investigate the effects of causal factors within large groups.

DESIGNING A GOOD TEST

One of the more interesting episodes in the history of science involves the theory of spontaneous generation. As recently as the late 1800s, many people believed that living organisms could be generated from nonliving material. One physician in the 17th century, for example, claimed that mice arose from a dirty shirt and a few grains of wheat placed in a dark corner. Similarly, it was thought that maggots—tiny, white wormlike creatures which are the larvae of common houseflies—were generated spontaneously out of decaying food. In 1688, an Italian physician, Francesco Redi, published a work in which he challenged the doctrine that decaying meat will eventually turn into flies. The following passage is from Redi's *Experiments on the Generation of Insects*:

> ... I began to believe that all worms found in meat were derived directly from the droppings of flies, and not from the putrefaction of meat, and I was still more confirmed in this belief by having observed that, before the meat grew wormy, flies had hovered over it, of the same kind as those that later bred in it. Belief would be vain without the confirmation of experiment, hence in the middle of July I put a snake,

> some fish, some eels from the Arno and a slice of milk-fed veal in four large wide-mouthed flasks; having well closed and sealed them, I then filled the same number of flasks in the same way, only leaving these open.[2]

In this passage Redi proposes a novel explanation for the worms that appear on decaying meat: they are derived from the droppings of flies. He next outlines the experiment he carried out. Fill two sets of four flasks with meat and fish, seal one set and leave the other set open so that flies can enter. Though he does not explicitly set out his prediction, it seems clear from what he says: worms will appear only in the second set of flasks. Is Redi's experiment well designed? Does it, in other words, control for the possibility of a false confirmation or rejection?

In fact, Redi succeeded in getting the results he wanted:

> It was not long before the meat and fish, in these second vessels, became wormy and flies were seen entering and leaving at will; but in the closed flasks I did not see a worm though many days had passed since the dead flesh had been put in them.

It may seem hard to imagine that Redi's outcome could be due to anything but fly droppings and that the possibility of a false confirmation is therefore quite low. However, many scientists of Redi's time believed in the doctrine of spontaneous generation and looked upon his results with some suspicion. They speculated that there might be some "active principle" in the air necessary for spontaneous generation. By depriving the meat and fish in the sealed containers of a sufficient flow of fresh air, they reasoned, Redi may have inadvertently prevented the spontaneous generation of worms. Thus, it seems at least a possibility that Redi's experiment has failed to account for a factor that could lead to a false confirmation of his explanation.

In light of this objection, Redi modified his experimental conditions and began again. Rather than sealing the first set of flasks, he covered them with a "fine Naples veil" that kept flies from coming into contact with the meat and fish but did allow air to circulate. Carrying out this modified experiment Redi once again obtained the expected results: worms appeared only in the uncovered flasks. By this maneuver Redi was able to rule out the possibility that something in the air—something necessary for spontaneous generation—might be responsible for his results. Consequently, the conclusion that fly dropping were responsible for the worms was on a much stronger footing.

Did Redi overlook anything that could have conceivably led to a false rejection of his explanation? What, for example, if the seals were not perfect, allowing flies to contaminate the sealed containers? The result would be "wormy" specimens in both sets of containers, an outcome that would have suggested that Redi was wrong. Of course, the way to eliminate this potential source of error is to examine the seals. If they are in working order then a negative outcome provides evidence that Redi was indeed mistaken. Assuming, then, that Redi took pains to insure that the covered flasks were properly sealed, the outcome ought to be decisive. Redi's experiment is well designed. In fact, Redi's results were

sufficiently clear to provide a foundation for further experimentation. Building on the work of Redi and others, later researchers were able to look much deeper into the processes Redi documented, using a new scientific instrument, the microscope, to observe the behavior of bacteria and other microorganisms.

Our discussion of the problem posed by the possibility of poorly sealed containers illustrates an important point. Frequently, experiments will involve some sort of apparatus: sensitive measuring instruments and devices, computers, and so on. It is always worth checking to make sure that all such apparatuses are operating properly. More than once, in the annals of experimental research, claims have been rejected on the basis of data obtained by malfunctioning equipment.

In tests of causal explanations like Redi's, experimental and control groups will often be used to rule out the possibility of a false confirmation. The members of the two groups will differ in only one respect. The experimental group but not the control group will be subject to the suspected cause. (In such experiments, the suspected cause will sometimes be called the *independent variable* and its claimed effect, the *dependent variable*.) The prediction, then, will be that only members of the experimental group will respond in the appropriate way. Thus, in Redi's second test, the experimental group was composed of the bits of meat and fish in the veil-covered flasks and the control group of specimens in the open flasks. His prediction was that worms would be found only in the latter group, the open flasks. Control groups provide an effective counter to the nagging possibility that some unknown explanatory factor may have been overlooked, something that may account for a successful outcome even if the explanation is wrong. For if the experimental and control groups are identical it is hard to imagine some factor other than the suspected cause that could be responsible for the predicted difference in outcomes between the two groups.

Next, consider a case in which experimental method is used to answer a question about an aspect of nature that cannot be directly observed. Psychologists have long wondered whether animal species other than human beings have a concept of self. Recently an experiment carried out at the Bronx Zoo provided an important clue about one species. An eight-foot square mirror was put in the enclosure where an Asian elephant named Happy lived. Initially, Happy peeked behind the mirror, touched it with her trunk, rubbed up against it, and swayed in and out of the field of view to see if her reflection did the same thing. After the first few days Happy was spending quite a bit of time in front of the mirror and would even bring her food to eat in front of the mirror. Happy was seeing an elephant in the mirror. But was Happy recognizing Happy or just another elephant? To answer this question, researchers painted an X on Happy's hide in a spot where she could not see it directly.

If Happy saw the image as herself, they conjectured, then she would try to investigate her own hide at roughly the spot where the X appeared on the body of the elephant in the mirror. As it turned out, this is just what Happy did. She spotted the mark in the mirror and became instantly curious, repeatedly probing the spot on her own hide as if it were some source of irritation. One researcher, Joshua Plotnik commented, "It seems to verify for us she definitely recognized herself in the mirror." This ingenious experiment enabled researchers to test

a claim that could not be directly verified or falsified—Asian elephants have a concept of self. And the experiment seems to be well designed. It is hard to imagine that Happy would behave in the way she did if she did not recognize the elephant in the mirror as herself. Thus, the chances of a false confirmation seem quite low. And unless Happy was totally uninterested in the mark on her hide, it is hard to imagine that she wouldn't respond as predicted if she did identify with the image in the mirror. (This contingency could be checked by painting a similar mark on a part of Happy's hide that she could directly see.) With this possibility ruled out, the chances of a false rejection seem quite low as well.

REAL-WORLD EXPERIMENTS

One feature shared by the two cases we have examined bears emphasis. Under naturally occurring conditions it would probably have been impossible to test either claim. In the first case, Redi found it necessary to put his specimens in a contrived environment to ensure that only one group would be exposed to flies. In the case of Happy the elephant, unusual circumstances had to be arranged so that Happy's subsequent behavior might provide a clue about a claim that could not itself be directly checked. But experimentation does not always involve the kind of special "laboratory" conditions required in these two cases. Sometimes nature will provide the clues necessary to test a claim. Consider, for example, the test described in the following news story.

SATELLITE SUPPORTS 'BIG BANG' THEORY

Phoenix—A NASA satellite has provided powerful evidence supporting the "big bang" theory, which holds that the universe began over 15 billion years ago with the most colossal explosion ever.

John C. Mather, an astronomer with the space agency, said Thursday that precise measurements by the Cosmic Background Explorer satellite of the remnant energy from the big bang give readings that are exactly as the theory predicted.

The theory, first aired in the 1920s, posits that all matter in the universe was once compressed into an exceedingly small and super-heated center that exploded, sending energy and particles outward uniformly in all directions. At the moment of the explosion, temperatures would have been trillions and trillions of degrees and have been cooling ever since.

If the theory is correct, astronomers expected an even distribution of temperatures just fractionally above absolute zero to still exist in the universe as an afterglow from the explosion.

Mather said that a Cobe instrument called the Far Infrared Absolute Spectrophotometer has now taken hundreds of millions of measurements across the full sky and has determined that the primordial temperatures are uniformly distributed. He said the uniform temperature left from the big bang is 2.726 degrees above absolute zero, or about minus 456.9 degrees F.[3]

This story reports on the results of an experiment done to provide new evidence for the big bang theory, an explanation most astronomers and cosmologists accept. (Even the most well-entrenched explanations can benefit from further confirmation, particularly if they involve elements—like the big bang theory—that cannot be directly observed.) The theory predicts a uniform temperature throughout the universe and consists of millions of measurements taken across the full sky.

The chances of a false rejection are quite low in this experiment, unless we have some reason to suspect the accuracy of the apparatus used to take the measurements. If the big bang theory is right, there should be a uniform afterglow and it ought to be detectable using the techniques mentioned. Are the chances of a false confirmation equally low? Can we, in other words, rule out the possibility that something else might explain the predicted result? Perhaps not, if the prediction were simply that there should be a uniform temperature throughout the universe. Cosmological events other than the big bang might be able to account for the uniformity. Or a successful match between prediction and actual outcome may be a matter of happenstance. After all, the universe either has a uniform background temperature or it does not. Perhaps the match was just a bit of luck. But the actual prediction involves a bit more. The story goes on to say:

> Craig Hogan, a University of Washington astronomer, said the new research "is verifying the textbooks" by providing powerful evidence for the theory. Hogan said that the Cobe results exactly match the theoretical curve of temperature energy decay that would be expected in the big bang theory.

This new passage suggests that the chances of a false confirmation are indeed low, largely due to the specificity of the prediction. The big bang theory predicts a very specific temperature at a very specific time in the development of the universe. And as it turns out, the universe is just as advertised. The close fit between prediction and experimental outcome would be hard to explain if the big bang theory were wrong!

Of course, observing and measuring what is going on in nature might turn up evidence that a claim is wrong. Suppose in our last example that astronomers did not find an even temperature throughout the universe, or that the temperature, though evenly distributed, did not match the predicted decay rates. Either result would suggest some underlying difficulty for the theory at issue, the big bang theory. At this point, however, rejection of the theory would be premature. There is a great deal of evidence from other experiments and observations that suggests that the theory is correct in broad outline. One compelling piece of evidence is the fact that the galaxies are moving away from each other in a way and at a rate that strongly suggests a common starting point some 15 billion years ago. What would be required instead is some modification of the theory's structure to account for the observational discrepancy. Well-confirmed theories in science are rarely overturned on the basis of a single negative experimental outcome. If enough negative evidence

accumulates, a big, seemingly well-confirmed theory may need to be discarded. A single negative result, no matter how dramatic, will generally necessitate some slight modification to the theory rather than wholesale rejection.

HOW NOT TO DESIGN A TEST

A good test will be designed to rule out the possibility of a false confirmation or rejection of the claim at issue. Perhaps the most effective way to underscore the importance of this strategy is by looking at the design of an experiment that fails on both counts. The experiment described in the following passage is intended to shed light on the question of whether or not animals have ESP.

> At mealtime you might put out two feedpans instead of one for your dog or cat. The feedpans should be located so that they are equally convenient to the animal. They should be placed six to eight inches apart. Both should contain the same amount of food and avoid using a feedpan the animal is familiar with. Pick the dish you wish the animal to eat from and concentrate on it. In this test, the animal has a 50% chance of choosing correctly half the time. You may want to keep a record of his responses over several weeks to determine how well your pet has done.[4]

The claim under scrutiny here is that animals are receptive to human thoughts via ESP and the prediction is that, under the experimental conditions outlined, pets will pick the dish we are thinking of more than 50% of the time. (Not a 50% chance "half the time" as the author of the passage claims!)

Is the test described in the passage a good one? First, can we rule out the possibility of a false rejection? Is there anything that could account for a failed prediction if the claim that animals have ESP is true? Suppose you were to say to your pet, in an entirely monotonous tone of voice, "Eat out of the red dish, the dish on the left, Fido." I doubt Fido would grasp the meaning of your words. Domestic animals tend to react to a complex of behavioral cues, some given by vocal inflection, but not to the meaning of words uttered in their presence. Thus if saying aloud, "Eat out of the red dish" will not do the trick, it is doubtful that thinking the same thing silently will work. Nor will it do to "picture" in your "mind's eye" the red bowl. I doubt Fido would react in the appropriate way to an actual picture of the bowl, so it seems highly unlikely Fido would react to nothing more than a "mental picture" of the red bowl. Thus, under the experimental conditions described in the passage, it seems entirely possible that Fido may fail even if he or she has some undiscovered extrasensory powers. A failed prediction, then, would not entitle us to conclude that animals do not have ESP unless we are willing to grant the entirely dubious claim that animals can understand human thoughts and words.

Second, can we rule out the possibility of a false confirmation? Is there anything that could account for a successful outcome if Fido does not have ESP? A number of things come to mind here that might explain a successful outcome.

First, suppose that our subject tended to go to one bowl instead of the other. It is possible that the experimenter, who is both sending the instructions and observing the outcome will inadvertently think of the dish the pet favors. Second, domestic animals are very good at discerning nonverbal cues. It may be that the experimenter is inadvertently looking at or standing in the direction of the dish being thought about and the experimental subject is picking up these cues. Finally there may be some bias at work on the part of the experimenter. Suppose our experimenter were convinced in advance of doing the experiment that animals have ESP. In recording or evaluating the subject's responses, the experimenter might inadvertently leave out responses that would otherwise provide evidence against animal ESP.

As you can see, the experimental test sketched in the passage is poorly designed in that it fails to help us conclude whether pets do or do not have ESP. The kind of analysis we have just completed should be done as a part of the design of any experiment. If our first attempts at designing an experiment fail to account for factors that could lead to a false confirmation or rejection of the claim at issue, we can go back to the drawing board armed with what we have discovered about potential weaknesses. Our subsequent design efforts are bound to do a more effective job of creating a decisive experiment.

CONCEPTUAL VAGUENESS

Our ESP test suffers from one other shortcoming, one that makes it difficult to see how a decisive test could be designed. The notion being investigated—extrasensory perception—is *conceptually vague*. So little is understood about what ESP might involve and how it ought to function that it is hard to say what we should expect to happen even in the most tightly controlled experiment. What should a person (or animal) with ESP be able to do and what should they not be able to do? Read the mind of another or perhaps sense their feelings? Anticipate what they are about to do? Are there factors that might inhibit ESP and, if so, how might we heed them in constructing a test? Are some people more psychically gifted that others? Are there physical conditions that impede the transmission of psychic messages? If we simply have no sense of what these factors might be, then any failure to get the expected result cannot be taken to show that the experimental subjects don't have ESP. As a general rule, the vaguer a claim is, the harder it will be to rule out the possibility of a false rejection. Recently, I came across an ad for Q-ray ionic bracelets on the Internet. The bracelets, it was claimed, could reduce the pain of arthritis by "balancing the flow of chi, the universal life force." This explanation would be difficult to test, since the notion of chi, of the "life force," is so vague that it is hard to say what we ought to expect to happen when "chi" is in or out of "balance," or how those mysterious "Q-rays" are supposed to interact with "chi."

Conceptual vagueness can make it difficult to rule out the possibility of a false confirmation as well. Suppose our work with Fido had been a smashing

success. All extraneous factors likely to lead to predictive success were anticipated and accounted for and yet Fido consistently ate from the correct bowl. In this scenario we would have established something, but it is not clear what that something is. Because so little is understood about what ESP might involve, we can only conclude that something interesting is going on, something we don't really understand.

Nonetheless, there can be some value in working with vague claims, particularly when they point in the direction of a potential new explanatory insight. Think again about two important episodes from the history of science discussed earlier, the work of Ignaz Semmelweis (see Chapter 1) and Francesco Redi. Both dealt with vague hypotheses: in Semmelweis's case the notion of "cadavric matter" and in Redi's, the idea that fly droppings could somehow transform into worms. In each case, what was confirmed was little more than a sense of what direction research toward a fuller explanation might take. Something similar can be said about much of today's neurobiological research. Modern brain scanning techniques have enabled researchers to isolate areas of the brain that are responsible for various sorts of cognitive functioning. Once the link between activity in a certain area of the brain and a given cognitive ability is fixed, researchers have the first clue as to how the brain might produce the activity in question. Recent experiments, for example, have isolated the areas of the brain that are active when a person responds to a joke. Precisely why this is the case remains an open question but now researchers know at bit more about where to look to begin closing in on an answer. Experiments designed to investigate conceptually vague notions are sometimes said to be *hypothesis generating* rather that *hypothesis testing* since much of the point is to generate new and more refined hypotheses for further investigation.

TESTING EXTRAORDINARY CLAIMS

With a few modifications, the experimental strategy we've been following can be used to test extraordinary claims of the sort discussed in Chapter 2. Consider a claim mentioned earlier in this chapter. People known as "water witches" or "dowsers" claim they can detect water with a simple forked wooden branch. Dowsers loosely grasp one of the forks in each hand and point the branch straight ahead, parallel to the ground. When they approach a source of water, the dowsing rod, as the forked stick is called, will point in the direction of the water, much as a compass needle will point in the direction of magnetic north. Many successful dowsers claim to be able to pinpoint sources of water for purposes of well drilling and some even claim to have found water where conventional geologists have failed.

As with most extraordinary claims, the evidence for dowsing is sketchy. We must rely on the testimony of dowsers and their clients about past performance. Moreover, the fact that a dowser points to a location, a well is drilled, and water discovered does not show that the dowser actually located water with his or her

dowsing rod. That water was found at the indicated location may have been a coincidence, or there may have been visual clues to aid the dowser, such as patches of greenery near the chosen location, etc. And we have no real sense of dowsers' success rates other than what they and their clients report. How often are they mistaken? Our challenge, then, is to devise an experiment that will give us decisive evidence, one way or the other, about the dowser's claimed ability.

To rule out the possibility of a false rejection, we need to come up with a set of conditions under which nothing could explain a dowser's failure other than an inability to find water with a dowsing rod. A good rule of thumb in setting up tests of extraordinary claims is to consult the experimental subject or subjects prior to designing the experiment. We want to set up conditions under which the experimental subjects will agree, in advance, that they ought to be able to perform. Otherwise failure in the actual test may be taken to show only that the experiment is hostile to the ability we are attempting to test. But if our subjects concur that the experiment approximates conditions under which they should be able to perform, such excuses lose much of their steam. If a person says he or she can perform under a given set of conditions, it is hard to take seriously protestations to the contrary, particularly after a failed test.

To eliminate the possibility of a false confirmation, we need experimental conditions under which nothing could explain our subject's success other than a real ability to dowse. What we want to try to rule out is the possibility of cheating, coincidence, inadvertent cuing on our part, visual or audio clues as to where the water is, and so on. If we succeed in imposing controls sufficiently tight to rule out these possibilities, success by the dowser can be taken to vindicate his or her claimed extraordinary ability.

Now that we have a sense of what a good experiment ought to involve, let's try our hand at actually designing one. Imagine we have contacted a group of the country's most well-known and successful dowsers and all have agreed to take part in our experiment. We propose the following test. We will place before each dowser 10 identical large ceramic jars with covers, arranged in a straight line equidistant from one another. Only one of the jars will contain water. The other nine will be empty. The dowser will be allowed to approach each jar but not to touch any jar. We will only test subjects who agree that they should be able to find the single jar with water. (We might give them a chance to dowse a jar they know contains water to insure that the experimental conditions meet their approval.) If a dowser is successful, he or she will be retested once the jars are rearranged. Of course, our subject will be asked to leave the room while the jars are being rearranged. As an additional precaution, no one who knows the location of the jar containing water will be allowed to be in the room while a dowser is being tested.

With all of the precautions we have built in, our experiment is well designed to provide unambiguous results. If a dowser can perform under such conditions we have strong evidence for dowsing. The odds of choosing the right jar in the first run are one in ten, in both the first and the second, one in a hundred. It is hard to imagine anything other than dowsing that could explain such results in our tightly controlled experiment. If, instead, the dowsers fail, it

would be hard to explain away the results given that the subjects have agreed that they should be able to perform under the test conditions.

No matter how well they are designed, tests of extraordinary abilities face a further hurdle. Suppose we run our test and all of our dowser fail. Believers in dowsing are likely to explain away our results on the ground that we have tested the wrong people, that our experiment is flawed in ways neither we nor they understand, or even that dowsing only works "in the field" under noncontrolled conditions. They will probably go on to point out that dowsing has been practiced for hundreds of years, and this is true: the earliest record of a successful dowsing dates to 1586, in Spain. Such objections are nearly impossible to counter but for this reason they lack any real credibility. They boil down to nothing more than the claim that dowsing cannot be tested. We need only reply that if it cannot be tested than we have no reason to believe it works! Dowsing is something of an anomaly and as we found in Chapter 2, the burden of proof lies with the believer, not the skeptic. Lacking any clear experimental evidence for dowsing, then, it is reasonable to assume that dowsing does not work.

PREDICTIVE CLARITY

One feature of our dowsing test deserves special note. We have been careful to arrive at a prediction that sets a clear line of demarcation between success and failure. If our dowser can find the jar containing water in two successive trials, he or she is successful; anything less constitutes failure. In designing controlled tests it is important to avoid predictions that blur the line between success and failure. Imagine, for example, we had decided to test our dowser by burying containers of water a few feet below the surface of a vacant lot. The dowser would then be instructed to place markers where he or she believed the containers to be located. Suppose the dowser placed markers within three or four feet of the location of one of the containers. Does this constitute a hit or a miss? Just how far off must a marker be before we consider it a miss? Or suppose markers are placed at ten locations when only five containers were buried and that seven of the markers are within a few feet of one or the other of the containers. How do we evaluate these results? Has our dowser succeeded or failed?

The line between success and failure can be very difficult to draw when a prediction involves some sort of subjective impression on the part of the experimental subject. Imagine, for example, we were to test a telepath, someone who claims to be able to read the thoughts of another. As part of our experiment we instruct the telepath to sketch a simple picture that someone in another room is concentrating on. Suppose the person in the other room is looking at a postcard of a small sailboat moored at a marina and that the telepath produces a simple drawing that includes a vertical straight line and a narrow triangular shape that might correspond to a boat hull or sail. To make matters worse, several of the drawing's details conform clearly to nothing we can discern on the postcard.

Is the telepath's impression accurate or inaccurate? Presuming we can decide what constitutes a detail or feature of the picture on the card, how many features or details must the telepath get right to be a clear indication of success?

To take another example, imagine if a tarot card reader were to give a personality analysis, based on the position and order of the cards, of someone unknown to the reader. The reading might indicate that the person in question, "tends to be optimistic despite occasional moments of depression or pessimism" or "makes friends easily" or "displays clear leadership ability." How do we evaluate such claims? The problem here is not only with the generality of the predictions but with the lack of a clear basis for judging them. We must first arrive at an accurate personality profile of the person in question. Presuming we could do this, what objective basis do we have for comparing our profile with that of the tarot card reader? No doubt any two sets of subjective impressions about a person's character will contain some words and phrases in common. How much similarity is required to put some stock in the analysis of the tarot card reader?

In designing a test, then, it is crucial that we arrive at a prediction that clearly spells out the difference between success and failure. If in evaluating the results of a test we are unable to say precisely whether our subject has succeeded or failed, then our test has very little point. Fortunately, however, the prediction in our dowsing test seems to be clear and unequivocal; success and failure are clearly spelled out.

BIAS AND EXPECTATION

Experiments are run by people and often their subjects are people as well. It should come as no surprise, then, that bias and expectation can have an unwelcome influence on experimental outcomes. Imagine if we were to set up the following test of the claim that spinal manipulation of the sort done by chiropractors can alleviate back pain. We interview a number of people suffering from various degrees of lower back pain, asking them to rate their pain on a scale of one to ten. Next we have a chiropractor provide an objective measure of their pain by noting how the subjects respond to various movements of the back. All the subjects are told to avoid strenuous activity and to take it easy for the next 30 days. Half of the patients are also provided a semiweekly chiropractic spinal manipulation over the 30-day test period. At the end of the test, subjects are again asked to rate their back pain and the chiropractor is asked to repeat his assessment of all of the subjects. Presumably, if there is more improvement in the experimental group, spinal manipulation is the reason. Or is it? There is a very real possibility that bias and expectation, not the power of chiropractic manipulation, account for the results.

Experimenter Bias. Chiropractors no doubt believe in the efficacy of the treatment they provide. It may be, then, that our chiropractor's improvement ratings will be influenced by his or her beliefs given that there is an element of

subjectivity to reports of pain. The way around this problem is to ask another chiropractor or other qualified health care professional to do the end-of-experiment assessments, with the stipulation that the new evaluator will not know whether various subjects were from the experimental or control group. With this precaution in place, experimenter bias can be ruled out as the source of any difference that emerges. Experiments in which experimenters are unaware of whether subjects are from the experimental or control group are sometimes called *single-blind* experiments.

Experimental Subject Expectations. At the end of our experiment, subjects are asked to rate their own improvement. If the members of our experimental group believe they are receiving treatment that will help their back problems, they may tend to over estimate just how much they have improved. Any improvement, that is, may be due to a placebo effect, the belief that they are being treated. The way around this potential problem is to find a way to insure that both experimental and control subjects believe they are receiving identical treatment. This might involve providing some sort of sham spinal manipulation for the control subjects. If both groups believe they are being treated for their back problems, any difference in outcomes could not be attributed to the expectations of the experimental subjects. Experiments in which subjects are unaware of whether they are members of the experimental or control group are another kind of single-blind experiment.

Experiments in which neither experimenter nor experimental subject is aware of which subjects are members of the experimental and control groups are said to be *double-blind*. Much medical research, for example, is double-blind. Experimental subjects might be given a substance which is thought to prevent a particular condition. Control subjects will often be given a placebo—an inert substance—to control for the possibility of suggestibility; experimenters who work with the subjects and who evaluate the results of the experiment, will not be told which subjects are in which groups. The rationale for keeping the experimenter "blind" is to control for the possibility that subjects may be treated differently during the course of the experiment and to insure that the evaluation of the subject's condition at the conclusion of the experiment will be unbiased.

Psychologists have long known that an experimental subject's knowledge that he or she is taking part in an experiment can influence that subject's performance. And this is another potential way in which subject expectation can influence the outcome of an experiment. Psychologists call this the *Hawthorne effect*. The *Hawthorne effect* got its name from a series of experiments conducted at the Hawthorne plant of Western Electric Company in Illinois during the 1920s and 1930s. Researchers were interested in isolating factors that might increase productivity, factors like rest periods and lengthened or shortened work days. What they found was that just about any change seemed to increase productivity, leading them to conclude that the Hawthorne effect was in part responsible for the increases; the fact that the workers knew they were being observed led them to work more efficiently. Ironically, a reevaluation of the

QUICK REVIEW 4.2 Experimental Design Checklist

A well-designed experiment must anticipate and resolve any issues suggested by these questions:

1. Can the possibility of a false rejection be ruled out?
2. Can the possibility of a false confirmation be ruled out?
3. Is the claim at issue conceptually clear?
4. Is the difference between predictive success and failure clearly specified?
5. Have controls been imposed to eliminate the influence of experimenter or experimental subject expectations?

An experiment designed to generate new hypotheses need not make a specific prediction.

data from the original experiments many years later suggested the increased productivity of the workers at the Hawthorne plant was not due to the Hawthorne effect! Rather it was due to the fact that the workers had improved their job skills over the months during which the experiments took place. Though perhaps ill named, the Hawthorne effect has been well documented in many other experimental settings.