4. The following is another version of John Wisdom's "Gardener Parable" from Carl Sagan's book, *The Demon-Haunted World* (1996): "A fire-breathing dragon lives in my garage" … "Show me," you say. I lead you to my garage. You look inside and see a ladder, empty paint cans, an old tricycle – but no dragon. "Where's the dragon?" you ask. "Oh, she's right here," I reply, waving vaguely. "I neglected to mention that she's an invisible dragon." You propose spreading flour on the floor of the garage to capture the dragon's footprints. "Good idea," I say, "but this dragon floats in the air." Then you'll use an infrared sensor to detect the invisible fire. "Good idea, but the invisible fire is also heatless." You'll spray-paint the dragon and make her visible. "Good idea, but she's an incorporeal dragon and the paint won't stick." And so on. I counter every physical test you propose with a special explanation of why it won't work.
5. Every time I snap my fingers the entire universe doubles in size, and I mean everything, so there's nothing to compare the now double-sized universe with the original universe. But I tell you, it's true.

9.6 Experiments and Other Tests

Testing is often one of the features of modern science hailed as rendering it superior to other forms of inquiry and other ways human beings advance claims about the world. Indeed, testing is crucial to empirical science. So, if empirical testing is so important to scientific methods, what is testing? Here are a few dimensions of testing critical thinkers ought to keep in mind when evaluating scientific (and non-scientific) claims.

Controls and variables

One of the most important requirements of a scientific test is to isolate the factor that is to be tested. Consider the following hypothesis: *higher levels of carbon dioxide in the atmosphere will result in higher atmospheric temperatures*. To test this, an experimenter must be able to isolate carbon dioxide levels and vary them independently of other variables. So far as possible, all potentially relevant factors should be kept the same except the CO_2 levels (the factor that's being tested). To do this, we might acquire two samples of atmosphere that are exactly the same in composition, volume, pressure, container, and the amount of sunlight to which each is exposed. We could, at that point, take an initial reading of temperature in each sample. Then we might add carbon dioxide to one sample while leaving the other unchanged. We would then take subsequent readings of temperature, comparing the samples to see if temperature changes with the addition of carbon dioxide. If the temperature rises, the hypothesis is confirmed (or at least not yet falsified). (See 6.8 and Mill's Method of Concomitant Variation.)

In this experiment, the temperature is called the *dependent variable*. It's the component of the experiment that we expect to be changed by altering the factor upon which we hypothesize it depends (the CO_2). Our hypothesis says that the temperature of the sample will change because it's dependent upon CO_2 levels. CO_2 is, in contrast, the *independent variable*. It is the factor changed by the experimenter and is not contingent or dependent upon the dependent variable (if it is, our experiment won't tell us anything new about our hypothesis). It must, of course, be independent.

This type of experiment is called a "controlled" experiment because the independent variable is varied while many other factors are held fixed. The sample where no independent variables are varied is called the *control group*. The sample in which the independent variable is varied is called the *experimental group*. To test the effectiveness of a new medicine in a controlled way, therefore, an experimenter would need two groups of patients that are alike in every way except that one receives the claimed medicine to be tested (the experimental group) and one does not (the control group). Constructing a controlled experiment isn't easy – though, as you can probably see, it's much easier to conduct controlled experiments in laboratories than in the outside world. Indeed, constructing controlled experiments is one of the principal reasons laboratories exist. Moreover, it's very difficult to conduct controlled experiments on people, given the tremendous variety among human beings. Putting together samples that are exactly the same and remain exactly the same except for a single factor is simply a tall order and is one reason science is so difficult.

Epidemiological studies

People are difficult to get into the laboratory, and they don't always comply with the rules of controlled experiment. It would be immoral to force them to stay there or to be treated in ways likely to harm them (forcing them to smoke or experience tragedy). So, in studying people it's often desirable just to examine the way they live in the world. You've probably heard about studies finding that eating low fat diets containing fish, garlic, and olive oil are good for your heart. For the most part, this wasn't discerned through controlled laboratory studies but by examining what people actually eat all over the world. Researchers noticed that large groups of people with high olive oil, low fat, and high fish consumption, such as the Japanese and those living in the Mediterranean, experience lower rates of heart disease than groups with different kinds of diets.

In these cases, people have already experienced the independent variable; to see whether this variable is correlated with the outcome we hypothesize, we compare these groups with people who haven't experienced the independent variable. This backward-looking type of study is called an *epidemiological* or *retrospective* study.

Epidemiological studies are not as strong as experiments with more controls for isolating specific causal factors, and so it's quite possible that the lower rates of heart disease among fish- and olive oil-eating populations are the result of something else (a lurking variable). Perhaps it's just a matter of genetics. The findings of epidemiological studies, however, are a good place to start looking for hypotheses that can be tested in labs or at least in more controlled conditions. What changes might occur if some testable population of, say, Scots moved to a Mediterranean diet? Sometimes, of course, epidemiological studies offer the best that's possible given the constraints of morality and human conduct.

Personal experience and case studies

Upon visiting a health food store, Peter met a clerk who swore that eating shark cartilage had cured his cancer. How did the clerk know? It was his personal experience, he said. He took the shark cartilage, and his cancer went into remission. We greet stories like this frequently in life. Typically, personal experience, however, is not taken to be adequate proof of very much, scientifically speaking. There are lots of reasons for this. For one thing, we are subject to many kinds of biases and cognitive distortions (the subject of Chapter 7). We deceive ourselves, we see what we want to see, and we miss a lot in experience. It's also just very difficult to sort out all the different factors that may contribute to an event, and especially a medical cure. The human body is complex, and we live in environments where we are exposed to thousands of different compounds daily. The body of the clerk with whom Peter spoke may have healed itself (spontaneous or natural remission happens in a remarkably high number of cancers). There may have been some compound in his environment that helped cure him. The original diagnosis may have been in error. On a more formal level, a sample of one is much too small on the basis of which to draw conclusions as general as "shark cartilage cures cancer." As we saw in 6.4, samples to be tested must be of an adequate size, randomized, stratified, perhaps studied in longitudinal ways (across long periods of time) – and, of course, the outcomes produced by experimental samples must be contrasted with control groups. Experiments, ideally, should be repeated, too.

Strictly speaking, even the personal experiences of physicians with their patients are not adequate as the bases of scientific conclusions. Such one-off experiences are called *anecdotal evidence*, and we cannot rationally infer from one or two instances something about whole populations. Physicians are highly skilled in the arts of diagnosis. They must be deeply informed about anatomy and the many ailments to which people are subject, as well as the many ways those ailments present themselves in different people. But physicians are subject to just the same biases and distortions as the rest of us. Having done what many of us cannot and produced an accurate diagnosis, a physician should not attempt to do on the basis of personal experience alone what only research scientists are competent to do – namely conclude what treatments are most effective for that illness. A physician's prescription should be evidence based – that is, based upon the findings of controlled, repeated scientific experiments – so far as possible. Physicians' and other healthcare givers' experiences with treatments are called, when recorded and organized, "case studies." Case studies, like personal experiences and epidemiological studies, are good starting points for developing hypotheses to test in rigorous, scientific ways. And recently there has even been increasing pressure to think of case studies as a weaker but still important kind of testing. But nothing beats a controlled scientific experiment to decide questions about matters of fact.

Blinding and double blinding

It may sound strange to suggest that good science sometimes requires blinding people, even in a metaphorical sense. After all, science is normally taken to help us see the world more clearly. But blinding is often regarded as necessary when experiments involve human beings. It's an admission that human subjectivity and beliefs may have some effect upon perception.

"Blinding" involves keeping human subjects of an experiment in the dark, so to speak, about whether or not they have been subjected to the factor being tested, or whether they are part of the control group. As an example, Teri has agreed to take part in an experiment testing a new analgesic medicine. Upon coming down with a headache, Teri takes the pill that had been distributed by the experiment staff and then gives a report on its effectiveness. Teri, like the other participants in the experiment, knows she's in an experiment, but she doesn't know whether the pill she received is the medicine or a dose of sugar. She is, in a sense, blind to what she has received. That's important because of what researchers call the *placebo effect*. The placebo effect occurs when people report a positive result from a medicine they believe they have received even when they have not really received it – that is, even when they have received only a placebo. The effect is related to *wishful thinking* (see 7.2). People wish for the medicines prescribed them to be effective. That wish can affect the way people interpret what they feel subjectively. (Some critics think that the placebo effect accounts for much more of the positive regard for the drugs prescribed for mental illness than is commonly understood.)

Blinding does not eliminate wishful thinking, but the procedure makes it possible for testers to factor it out of experimental results. When the results come in, rather than simply count positive outcomes, experimenters only need look for whether or not there is a significant difference between the results of the subjects who received the placebo and the subjects who really received the tested factor. This, of course, assumes the placebo effect and wishful thinking are distributed evenly across the experimental population.

Another form of blinding is called "double blinding." In double blinding, both the subject and the person administering the factor are kept in the dark about who's received the experimental therapy and who's received only a sugar pill or salt water. That's because the body language, facial expressions, and tone of voice of the person administering the therapy might give away the truth to patients. You may have heard, for example, about the famous case of "Clever Hans." Hans was a horse whose owner (and many others) was convinced could do basic arithmetic. But as it turned out, careful observation discovered that the owner was making quite unconscious facial gestures that Hans was correctly interpreting. Hans would stroke out the answer with his hoof, pawing one, two, three, etc., until the answer to the math problem was reached.

But when the answer was reached the owner would physically react, raising his eyebrows, widening his eyes, leaning forward, etc. Hans would stop pawing there and receive a welcome reward – perhaps a sugar cube! Sometimes people are as smart as horses, and researchers need to protect their experiments from the sorts of errors the reading of body language can introduce.

In vitro studies

One thing case studies and epidemiological studies have going for them, however, is that they deal with people as they live in the world. Labs are artificial environments – necessarily so. And as such they may miss something of the synergies and interfering, magnifying, intersecting causal networks that compose the world and that produce the phenomena in which scientists are interested. So, some therapies that work in the lab may be stifled by factors in play out in the world. Popular household cleaners may react with, say, anti-cancer drugs.

Experiments done in the lab are, therefore, often done "*in vitro*" (from the Latin for "in glass") – that is, in petri dishes, test tubes, vials, and beakers – in isolation from the hundreds of thousands of compounds that compose the human body and its normal environment. The strength of *in vitro* studies is that scientists can bracket out the possible influence of those other potential factors and isolate the independent variable. The weakness, however, is that a drug that effectively diminishes the activity of cancer cells in a petri dish may not do so in the human body precisely because of the way it interacts with those many other factors in the real world – and vice versa. *In vitro* studies may, therefore, seem a good place to start, but remember that a drug that doesn't work "in the glass," so to speak, might actually work in a living body because of those same unknown but crucial interactions. In cases like that, *in vitro* experiments may actually mask a drug's beneficial effects.

Non-human animal studies

Similar problems (and advantages) arise with studies conducted on non-human animals such as rats and mice. It's a lot easier to work with mice, and since most people don't mind mice dying in the course of studies the way they'd object to human mortality, rats and mice can be subjected to various compounds and treatments we would find immoral (and illegal) to administer to people. But the very thing that has made it permissible to conduct experiments on non-humans also raises questions about the scientific value of studies involving them – namely, that non-humans are different from humans. As any veterinarian will tell you, drugs that may be effective in treating diseases afflicting mice may be ineffective in treating even the same disease in human beings. And, vice versa, treatments that fail in non-humans may succeed in humans. Aspirin can be toxic to cats, for example, and chocolate in substantial quantities is toxic for dogs. Researchers using non-humans must be very careful, therefore, that the sorts of biochemical mechanisms they examine are, in fact, sufficiently similar across species and that no intervening factors are at play in one subject species but not in others.

There are, moreover, continuing controversies about the moral propriety of experimenting on living non-humans, too, even in cases where it's scientifically valuable to use them. Good critical thinkers will consider that issue. Remember that Nazi experiments on their prisoners did yield real scientific results, but it doesn't for that reason make those experiments or others like them morally permissible. A number of European countries (Netherlands, Austria, Sweden, the UK, Germany) and New Zealand have banned the use of "great apes" or, more precisely, family *hominidae* (chimpanzees and bonobos, gorillas, orang-utans) as well as humans in various forms of experimentation. The United States seems to be reducing the scope of its use of non-humans in experimental contexts, as well. The US National Institutes for Health (NIH), for example, announced in 2013 its decision to comply with the recommendations of the Institute of Medicine (IoM) and dramatically reduce its use of primates in scientific studies. Science is a human practice like any other, and human practices are governed by moral considerations. What counts as moral and immoral scientific practice, as well as what counts as strong and weak science, is a complex and rich subject with which critical thinkers should remain engaged.

SEE ALSO

8.1 Knowledge: The Basics
8.6 Justification: The Basics
9.7 Six Criteria for Abduction

READING

John Wright, *Explaining Science's Success* (2014)
Hans Radder, *The Philosophy of Scientific Experimentation* (2003)
Samir Okasha, *Philosophy of Science: A Very Short Introduction* (2002)

9.7 Six Criteria for Abduction

Abduction? No, we're not talking about kidnapping but, rather, an important set of considerations for deciding among multiple possible explanations. The term "abduction" was coined first by the philosopher Charles Sanders Peirce (1839–1914) for this kind of thinking. Because the word "abduction" seems so odd, people often just call these procedures *inference to the best explanation*. They have been developed because, contrary to what many people believe, evidence and scientific data don't always point to a single explanation.

Consider, for example, the serious illness called AIDS. AIDS is a "syndrome" or a collection of symptoms and conditions. The scientific community has so far settled on an infectious virus, HIV, as the explanation for the emergence of that syndrome. But other explanations are possible. Consider, for example, that some have explained the disease as a punishment by God for human sin, in particular the sin of homosexual sex. Consider, too, that rather than causing the disease, HIV might just be part of the AIDS syndrome such that HIV infection might itself be the result of a deeper common cause (see 6.3). Perhaps people's chakras or humors are out of balance. Perhaps there is a still unknown biological condition that produces illness. Perhaps some evil scientist or the CIA has developed a machine that can strike people down with this syndrome from a great distance. How is one to decide? One's immediate instinct is to appeal to the empirical evidence. But what if it's the case, as French physicist Pierre Duhem and US philosopher W. V. O. Quine have argued, that no matter how much evidence we amass there will always be possible alternative explanations. In part for reasons we saw in 9.4 and 9.5, Duhem and Quine argue that no body of evidence can fully determine our truth claims.

Don't worry, though. In the face of the limits of empirical evidence to decide among possible explanations, inquirers have developed a number of criteria you can use to decide what is the "best" explanation available. There remains some controversy about what standards are relevant, but the following set reflects a large consensus of thinkers.

1. *Predictive Power*. Explanations that offer greater predictive power are better than those that offer less. The HIV theory seems to be a better predictor of who will contract AIDS than the theory of divine punishment, since there is a much tighter correlation between HIV infection and AIDS than there seems to be between "sin" and AIDS. The evil scientist theory of AIDS doesn't even seem to allow for prediction at all, since we are unable to observe the operations of either the machine or the evil scientist. Untreated HIV infection, in fact, seems closer than any other explanation to describing a *nearly sufficient* and *necessary condition* for contracting AIDS. That is, if one contracts HIV, then in the absence of treatment one faces a rather high likelihood of AIDS; and if one has AIDS, then it is almost certain that one is infected with HIV. (You can see a bit more here about why the logical considerations of sufficient and necessary conditions we explored in 2.2 are so important.)
2. *Scope*. Theories that cover more phenomena are preferred to those that are less comprehensive. So, Einstein's physics is thought to be stronger than Newton's because it explains physical phenomena at large, medium, and subatomic scales at any speed, while Newton's theories don't seem to be able to explain the behavior of very massive or very tiny and very fast things at all. Similarly, the HIV explanation of AIDS is part of the more comprehensive germ theory of disease, a theory that has been able to explain an enormous variety of human illnesses. One might, of course, hold that all disease, or perception of disease, flows from divine or supernatural causes, as perhaps Christian Scientists do. That

would be an explanation of pretty large scope. But for most, the divine punishment theory of disease is limited in its scope to a relatively restricted number of afflictions.

3. *Coherence with Established Fact*. We've already seen in our discussion of "paradigms" that sometimes claims and hypotheses that do not fit with current science – e.g., Galileo's conclusions – end up later being accepted as true. In general, however, what fits with established fact and theory is preferable to what does not. Science often assumes that the order of reality or nature is coherent, systematic, and logically consistent. If that's true, then a patchwork of incompatible theories to explain the world isn't desirable. Better to work toward a comprehensive science where various sectors of explanation all fit together to compose one great, consistent, and unified theory of the world. The HIV theory seems to fit better than others with what people have discovered about other similar diseases – namely, that viruses or bacteria cause them. The evil scientist theory and the divine punishment theory fit less well.
4. *Repeatability*. Science, as we've seen, relies on various forms of testing. All things being equal, theories confirmed through tests that are repeated and repeatable are preferable to those that are not. One might even say as a general principle that: "A single test does not a fact establish." Single experiments have found correlations between prayer and recovery. Repeated and multiple tests have produced inconsistent results. The relatively tight correlation between HIV infection and AIDS, however, has been observed over and over.
5. *Simplicity*. This criterion is often associated with medieval philosopher William of Ockham (1288–1347) and the principle known as "Ockham's Razor." For our purposes, the idea is that the best explanation is the simplest. By simplicity, we mean the fewest kinds of causal factors (whether physical or metaphysical) or the fewest unproven theoretical commitments. So, the HIV theory of AIDS implies at most the existence of the natural, biological world. The divine punishment theory implies the existence of that world plus the existence of God. The former theory is therefore simpler.

   Simplicity, however, may also apply even to the positing of entities within a given metaphysical framework. So, the evil genius argument requires adding to the world technologies that have been so far unknown, while the HIV theory requires no additions.

   Of course, the weakness of this standard can be seen in the thought that the world just might not be a simple place. As a matter of characterizing good explanation, then, we might better think of the simplicity standard as one of caution, as a caveat. Ockham's principle requires not that one should never add complications to an explanatory theory but only that one should not do so unless it's rationally necessary or there's some other compelling scientific reason to do so. In other words, keep theories simple and limited until the evidence compels you do otherwise. When thinking critically, look for excesses in a theory under scrutiny. What can it deal without? Are there other, simpler theories that might explain the phenomena just as well? When Napoleon asked Simon Laplace (1749–1827) why his book on celestial mechanics did not refer to God, the scientist did not answer with the metaphysical claim that God does not exist. Instead, appealing to the principle of explanatory simplicity, Laplace is reported to have simply said, "*je n'ai pas eu besoin de cette hypothèse*" ("I had no need of that hypothesis").

6. *Fruitfulness*. Theories are more "fruitful" when they make possible more hypotheses to test and when they encourage more testing. The germ theory of infection has made possible immense fields of scientific research into microbiology, biochemistry, and pharmaceuticals. The divine punishment theory has opened up fewer possibilities for testing. Yes, it's possible to test whether refraining from sinning changes one's likelihood of being afflicted with AIDS, but the HIV theory in contrast leads to vastly more avenues of investigation into the biological and natural worlds. Moreover, the HIV theory not only opens the possibility of new lines of inquiry, it encourages them. The divine punishment theory, however, instead resists additional testing and encourages what we've seen is called "falsification resistance." While the emergence of AIDS in populations that don't engage in homosexual sex and its control among homosexuals by practicing safe gay sex might be seen to falsify the idea that AIDS is a punishment for sin, those committed to the punishment theory readily find ways to explain the falsification away. Perhaps God accomplished His objectives already; perhaps God created HIV as a warning of the sort of thing that would be coming in the afterlife. Perhaps everyone who contracts AIDS deserves it, but not only for homosexuality. The facility with which explanations that appeal to the divine evade falsification renders them difficult and perhaps impossible to assess in scientific ways. That resistance to falsification may be something important for critical thinkers to remember about religious controversies.

SEE ALSO

6.3 Fallacies about Causation
9.4 Scientific Method
9.6 Experiments and Other Tests

READING

Douglas Walton, *Abductive Reasoning* (2014)
Mark Battersby, *Is that a Fact? A Field Guide to Statistical and Scientific Information*, revised edn (2013)
Atocha Aliseda, *Abductive Reasoning: Logic Investigations into Discovery and Explanation* (2006)
Peter Lipton, *The Inference to Best Explanation* (2004)