Historical science, experimental science, and the scientific method

Carol E. Cleland
Department of Philosophy and Center for Astrobiology, University of Colorado, Boulder, Colorado 80309, USA

ABSTRACT

Many scientists believe that there is a uniform, interdisciplinary method for the practice of good science. The paradigmatic examples, however, are drawn from classical experimental science. Insofar as historical hypotheses cannot be tested in controlled laboratory settings, historical research is sometimes said to be inferior to experimental research. Using examples from diverse historical disciplines, this paper demonstrates that such claims are misguided. First, the reputed superiority of experimental research is based upon accounts of scientific methodology (Baconian inductivism or falsificationism) that are deeply flawed, both logically and as accounts of the actual practices of scientists. Second, although there are fundamental differences in methodology between experimental scientists and historical scientists, they are keyed to a pervasive feature of nature, a time asymmetry of causation. As a consequence, the claim that historical science is methodologically inferior to experimental science cannot be sustained.

Keywords: methodology, induction, history, experimental investigations.

INTRODUCTION

Experimental methods are commonly held up as the paradigm for testing hypotheses: the scientific method, widely disseminated in introductory science texts, is modeled upon them. But not all scientific hypotheses can be tested in the laboratory. Historical hypotheses that postulate particular past causes for currently observable phenomena provide good examples. Although historical hypotheses are usually associated with fields such as paleontology and archaeology, they are also common in geology, planetary science, astronomy, and astrophysics. Some familiar hypotheses are continental drift, the meteorite-impact extinction of the dinosaurs, the big bang origin of the universe, and, more recently, the hypothesis that there are planets orbiting distant stars. What all of these hypotheses have in common is explaining observable phenomena (e.g., the complementary shapes of the east coast of South America and the west coast of Africa, the iridium and shocked quartz in the Cretaceous-Tertiary (K-T) boundary, the isotropic three-degree background radiation, the wobbling reflex motion of certain stars) in terms of their past causes. As discussed herein, the use of computer simulations does not change their historical character.

Although the idea that all good scientists employ a single method for testing hypotheses is popular, an inspection of the practices of historical scientists and experimental scientists reveals substantial differences. Classical experimental research involves making predictions and testing them, ideally in controlled laboratory settings. In contrast, historical research involves explaining observable phenomena in terms of unobservable causes that cannot be fully replicated in a laboratory setting. Many experimental scientists recognize this difference, and identifying sound scientific practice with their own work, sometimes denigrate the claims of historical scientists, contending that they can't falsify their hypotheses or that their confirmatory arguments resemble just-so stories (Rudyard Kipling's fanciful stories, e.g., how leopards got their spots). The startling number of physicists and chemists who attack the scientific status of neo-Darwinian evolution provides telling examples of this phenomenon. The most trenchant criticism of historical science, however, comes from an editor of *Nature,* Henry Gee (1999, p. 5, 8), who explicitly attacked the scientific status of all hypotheses about the remote past; in his words, "they can never be tested by experiment, and so they are unscientific. . . No science can ever be historical."

This paper explains why historical science is not inferior to experimental science when it comes to testing hypotheses. First, objections such as Gee's are based upon common misconceptions about experimental practice and scientific methodology in general. Second, the differences in methodology that actually do exist between historical and experimental science are founded upon a remarkably pervasive feature of nature: a causal asymmetry between present and past events, on the one hand, and present and future events, on the other. Insofar as each practice is tailored to exploit the information that nature puts at its disposal for evaluating hypotheses, and the character of that information differs, neither practice can be held up as more objective or rational than the other.

THE SCIENTIFIC METHOD

The hypotheses tested in classical experimental research are general in character; "all copper expands when heated" provides a toy example. A conditional statement T (test implication) is inferred from a hypothesis H. T states what must happen if H is true. Test implications have the following form: if condition C (heating a piece of copper) is brought about, then event E (the expansion of copper) will occur. Test implications provide the basis for experiments. Condition C is artificially produced in the laboratory, and investigators look for an instance of E.

How are hypotheses evaluated in light of the evidence obtained in an experiment? Under the rubric "the scientific method," science texts, from grade school through college, invariably provide one (or a combination) of two accounts, scientific inductivism or falsificationism. Scientific inductivism, commonly attributed to Francis Bacon, holds that the occurrence of the predicted event E under condition C provides confirming evidence for H, and that if enough confirming evidence of the right sort is obtained, H should be accepted by the scientific community. Unfortunately, scientific inductivism runs afoul of the hoary problem of induction: no finite body of evidence can conclusively establish a universal generalization. Faced with the problem of induction, many scientists embrace falsificationism, which holds that although hypotheses cannot be proved, they can be disproved. Unlike inductivism, falsificationism receives support from logic. It utilizes a logically true inference rule called "modus tollens." According to modus tol-

© 2001 Geological Society of America. For permission to copy, contact Copyright Permissions, GSA, or editing@geosociety.org.

lens, a generalization is false if it has at least one counterexample. The hypothesis that all copper expands when heated is thus false if there is a single case in which copper fails to expand when heated. Thus, although one can never prove the hypothesis (because no amount of testing can rule out the possibility that a piece of copper will someday fail to expand when heated), it seems that it could be disproved. In philosophical circles, falsificationism is associated with the work of Karl Popper (1963), who developed the logical insight about modus tollens into a sophisticated account of scientific practice. The basic idea behind Popperian falsificationism is to subject a hypothesis to a "risky test," a test that, in the context of one's background beliefs, is judged highly likely to yield a disconfirming result. If the prediction fails, modus tollens is invoked, and the hypothesis is ruthlessly rejected. According to falsificationism, it is unscientific to try to confirm a hypothesis.

For more than 50 years philosophers have known that falsificationism is deeply flawed. There are two central difficulties. First, any actual experimental situation involves an enormous number of auxiliary assumptions about equipment and background conditions, not to mention the truth of other widely accepted theories. When these conditions are taken into consideration, the logical inference licensed by modus tollens is radically altered. The falsity of an auxiliary assumption (versus the target hypothesis) could be responsible for a failed prediction. Every science student is implicitly aware of this because repetitions of classical experiments in laboratory exercises often go wrong not because the hypothesis being tested is false, but because, for example, equipment malfunctions or the sample is contaminated. Moreover, this difficulty cannot be circumvented by varying the conditions under which a hypothesis is tested, given that the number of auxiliary conditions involved in any real-world situation is unknown and potentially infinite; it is impossible to control for them all. The famous Popperian directive to bite the bullet and reject the hypothesis in the face of a failed prediction has no logical force. Furthermore, as Kuhn (1970) pointed out, scientists almost never practice falsificationism. In the face of a failed prediction, they mount a sustained search for conditions other than C that might be responsible. This amounts to exercising the logically permissible option of salvaging a hypothesis by rejecting an auxiliary assumption. A good example is provided by the response of nineteenth century astronomers to the perturbations in the orbit of Uranus; the orbit deviated from what was predicted by Newtonian celestial mechanics. Astronomers didn't behave like good falsificationists and reject Newton's theory: they rejected the assumption that there were no planets beyond Uranus, and discovered the planet Neptune. The moral of this story is that rejecting a hypothesis in the face of a failed prediction is sometimes the wrong thing to do; it is not an accident that logic gives us the option of rejecting an auxiliary assumption instead. In short, logic does not dictate that scientists behave like good falsificationists, and scientists do not in fact behave like good falsificationists. As a consequence, falsificationism cannot be used to justify the superiority of one science over another vis-à-vis the testing of hypotheses.

Let us look more closely at what experimental scientists actually do when they test a hypothesis. The test condition C, specified by the target hypothesis, is held constant (repeated) while other conditions are varied. When this activity is preceded by a failed prediction, it resembles the activity condemned by Popper, namely, an ad hoc attempt to save a hypothesis from refutation by denying an auxiliary assumption. However, there is an alternative interpretation: it may be viewed as an attempt to protect the hypothesis from misleading disconfirmations. It is significant that the same process of holding C constant while varying auxiliary conditions also occurs upon a successful test of a hypothesis. Moreover, C itself may be removed for the purpose of determining whether it was required for the successful result. While these responses to successful tests superficially resemble attempts at falsification, a little reflection reveals that this can't be what is going on, because they do not conform to Popper's requirement that the tests performed be "risky." The hypothesis has survived similar tests, and no one expects it to fail this time. Even if it does, it won't automatically be rejected. Viewed from this perspective, the activity more closely resembles an attempt to protect the hypothesis from misleading confirmations. In other words, a close look at the work of experimental scientists suggests that they are primarily concerned with protecting their hypotheses against false negatives and false positives, as opposed to ruthlessly attempting to falsify them. This makes good sense because, as discussed earlier, any actual test of a hypothesis involves many auxiliary conditions that may affect the outcome of the experiment independently of the truth of the hypothesis.

In this light, let us turn to the reputedly problematic differences between historical and experimental science. Historical scientists are just as captivated by falsificationism as experimental scientists; as three eminent geologists (Kump et al., 1999, p. 201) counsel in a recent textbook discussion of the extinction of the dinosaurs, "a central tenet of the scientific method is that hypotheses cannot be proved, only disproved." Nevertheless, there is little in the evaluation of historical hypotheses that resembles what is prescribed by falsificationism. The big bang theory of the origin of the universe provides an excellent example. It postulates a particular occurrence (a primordial explosion) for something we can observe today, i.e., the three-degree background radiation, first detected by satellite antennas in the 1960s. Traces, such as the three-degree background radiation, provide evidence for historical hypotheses, just as successful predictions provide evidence for the generalizations tested in experimental science. There is little or no possibility of controlled experiments, however, because the time frame required is too long and/or the relevant test conditions too complex and dependent upon unknown or poorly understood extraneous conditions to be artificially realized.

This doesn't mean, however, that hypotheses about past events can't be tested. As geologist T.C. Chamberlin (1897) noted, good historical researchers focus on formulating multiple competing (versus single) hypotheses. Chamberlin's attitude toward the testing of these hypotheses was falsificationist in spirit; each hypothesis was to be independently subjected to severe tests, with the hope that some would survive. A look at the actual practices of historical researchers, however, reveals that the main emphasis is on finding positive evidence—a smoking gun. A smoking gun is a trace that picks out one of the competing hypotheses as providing a better causal explanation for the currently available traces than the others.

The meteorite-impact hypothesis for the extinction of the dinosaurs provides a good illustration (Alvarez et al., 1980). Prior to 1980 there were many different explanations for the demise of the dinosaurs, including disease, climate change, volcanism, and meteorite impact. The discovery of extensive deposits of iridium in the K-T boundary focused attention on the impact of a meteor; iridium is rare at Earth's surface, but high concentrations exist in Earth's interior and in meteors. The subsequent discovery of shocked quartz in the K-T boundary cinched the case for the impact of a large meteorite, because there was no known volcanic mechanism for producing that much shocked quartz. The causal connection between the impact and the extinction, however, required a bit more work (Clemens et al., 1981). It wasn't until it became clear that the dinosaurs had died out fairly quickly around the time of the impact that the iridium and shocked quartz took on the character of a "smoking gun" for the meteorite-impact hypothesis. In short, of the available hypotheses and in light of the existing evidence (e.g., fossil record, iridium, shocked quartz, crater), the me-

teorite-impact hypothesis supplied the most plausible causal mechanism for understanding the demise of the dinosaurs.

Although historical investigations of past events often involve laboratory work, the purpose is different from that of classical experimental research. The main emphasis is on analyzing and sharpening traces so that they can be identified and properly interpreted. As an example, speculation that life goes back 3.8 b.y. rests upon laboratory analysis of carbon isotope ratios in grains of rock as small as 10 μm across and weighing only 20×10^{-15} g (Mojzsis et al., 1996). However, historical scientists sometimes investigate auxiliary assumptions in the laboratory. A good example is the Miller-Urey experiments (Miller, 1953), which were touted as supporting the hypothesis that life on Earth began in a primordial soup, but really supports the auxiliary assumption that some of the building blocks of life (amino acids) can be produced by electrical discharges on a mixture of methane, hydrogen, ammonia, and water. In this context it is sobering to note that most scientists now believe that the origin of life on Earth is not compatible with the conditions of the Miller-Urey experiment. It is thought that Earth's early atmosphere did not contain abundant methane or ammonia, and that life may have begun near a deep-sea volcanic vent (Orgel, 1998).

Similarly, it is important not to conflate the computer-aided modeling that has become popular in historical research with performing controlled laboratory experiments. The most a computer can do is determine the consequences of a hypothesis under a small number of explicitly represented hypothetical conditions. It cannot determine which of these hypothetical conditions actually exists in the concrete physical system being modeled, nor can it represent all of the other, possibly relevant, physical conditions present in the concrete physical system. A salient example is provided by early climate simulations of a snowball Earth, which indicated that there was nothing that could reverse a global freeze (Hoffman and Schrag, 2000). The climate modelers failed to consider the activity of volcanoes, which would continue to vent carbon dioxide during a global freeze, eventually producing a greenhouse effect that would rapidly melt the ice. The point is, modeling past events is theoretical work, and while it may yield predictions, these predictions are only as secure as the assumptions upon which the model is based. The best that can be done is to search for predicted phenomena in the uncontrollable world of nature, and there are no guarantees that they will be found, even supposing that the hypothesis is correct. This brings us to the crucial point: although computer-aided models may suggest what to look for in nature, and traces and some auxiliary assumptions may be investigated in the laboratory, one cannot experimentally test a historical hypothesis per se; to recapitulate, the time frame is too long and the test conditions too complex to be replicated in a lab.

In summary, Gee (1999) was correct about there being fundamental differences in the methodology used by historical and experimental scientists. Experimental scientists focus on a single (sometimes complex) hypothesis, and the main research activity consists in repeatedly bringing about the test conditions specified by the hypothesis, and controlling for extraneous factors that might produce false positives and false negatives. Historical scientists, in contrast, usually concentrate on formulating multiple competing hypotheses about particular past events. Their main research efforts are directed at searching for a smoking gun, a trace that sets apart one hypothesis as providing a better causal explanation (for the observed traces) than do the others. These differences in methodology do not, however, support the claim that historical science is methodologically inferior, because they reflect an objective difference in the evidential relations at the disposal of historical and experimental researchers for evaluating their hypotheses.

ASYMMETRY OF OVERDETERMINATION

Localized events tend to be causally connected in time in an asymmetric manner. As an example, the eruption of a volcano has many different effects (e.g., ash, pumice, masses of basalt, clouds of gases), but only a small fraction of this material is required in order to infer that it occurred; put dramatically, one doesn't need every minute particle of ash. Indeed, any one of an enormous number of remarkably small subcollections of these effects will do. Running things in the other direction of time, however, produces strikingly different results. Predicting the occurrence of an eruption is much more difficult than inferring that one has already occurred. There are too many possibly relevant conditions (known and unknown), in the absence of which an eruption won't occur.

Philosopher David Lewis (1991) has dubbed this time asymmetry of causation "the asymmetry of overdetermination." The basic idea is that localized present events overdetermine their causes and underdetermine their effects. Perhaps the best way to appreciate the extent of the asymmetry of overdetermination is to consider the difficulty of committing a perfect crime; i.e., footprints, fingerprints, particles of skin, disturbed dust, light waves radiating outward into space must be eliminated. It isn't enough to eliminate just a few of them; anything missed might be discovered by a Sherlock Holmes and used to convict you. Moreover, each trace must be independently undone. You cannot remove a footprint by eliminating a particle of skin or, for that matter, another footprint. In contrast, and this is the other side of the asymmetry of overdetermination, erasing all traces of a crime *before* it occurs is remarkably easy, usually requiring only a single intervention: don't fire the gun.

The physical source of the asymmetry of causation is controversial. It has been variously explained in terms of the second law of thermodynamics (statistically interpreted), the radiative asymmetry—wave phenomena (e.g., water, light) diverge into the future from their sources—and the initial conditions of the universe (Price, 1996). There is general agreement, however, that it represents an objective and pervasive physical phenomenon at least at the macro-level of nature (e.g., volcanoes, rocks, footprints, fossils, stars).

The asymmetry of overdetermination explains the reputedly problematic differences between historical and experimental science vis-à-vis the testing of hypotheses. Just as there are many different possibilities (subcollections of traces) for catching criminals, so there are many different possibilities for establishing what caused the extinction of the dinosaurs. Like criminal investigators, historical scientists collect evidence, consider suspects, and follow leads. More precisely, they postulate differing causal etiologies for the traces they observe, and then try to discriminate from among them by searching for a smoking gun—a trace that will identify the culprit beyond a reasonable doubt.

Lewis (1991) explicitly characterized the asymmetry of overdetermination in terms of causal sufficiency. It might, however, turn out to be a probabilistic phenomenon; subcollections of traces might make their causes merely highly probable, as opposed to determining them. Human experience is consistent with either possibility. Just as experimental work is irremediably fallible—due to the uneliminable threat of unknown interfering conditions—so the traces uncovered by field work are never enough to conclusively establish the occurrence of a hypothesized past event, perhaps because we haven't discovered enough of them or perhaps because there are no causally sufficient subcollections. In either case, however, the asymmetry of (quasi) overdetermination helps to explain the methodology of historical researchers. It tells us that a remarkably small subcollection of traces is enough to confer at least high probability on the occurrence of a past event, and that there are likely to be many such subcollections. The existence

of so many different possibilities for rendering a hypothesis highly probable provides the rationale for searching for a smoking gun.

In some cases, a smoking gun may be inferred directly from the hypothesis under investigation. A salient example is the big bang theory of cosmology (Kaufman, 1977). Robert Dicke and his team of Princeton physicists predicted that if the big bang theory were true, the universe should contain an isotropic, microwave background radiation a few degrees above absolute zero. The subsequent discovery by Wilson and Penzias (Kaufman, 1977) of the mysterious three-degree background radiation was taken as providing pivotal evidence for the big bang theory over the steady state theory. Sometimes, however, one just gets lucky and stumbles over a smoking gun, as did the Alvarezes et al. (1980) in the case of the meteorite-impact hypothesis for the extinction of the dinosaurs; the existence of iridium and shocked quartz in the K-T boundary was not predicted in advance of its discovery. Moreover, with the passage of time, traces of events become more and more attenuated, and eventually they may disappear. Alternatively, they may be present but very degraded. Finding them may require advances in technology. The discovery of the three-degree background radiation depended upon the development of very sensitive antennas for communicating with satellites. Similarly, a particle accelerator (cyclotron) was used to discover the iridium in the K-T boundary. Finally, with new evidence and new explanatory hypotheses, the status of a trace as a smoking gun may change; there is no more certainty in the methodology of historical science than there is in the methodology of experimental science. The important point is that one can never rule out the possibility of finding a smoking gun, and this is a consequence of the overdetermination of the past by the localized present. Failure to search for a smoking gun deprives a historical hypothesis of empirical grounding, turning it into a dreaded just-so story.

This brings us to the practice of experimental science. The causation of an event is a complex affair. Consider a short circuit that causes a house to burn down. Take away the short circuit and the house wouldn't have burned down; the short circuit triggered the fire. But there are many other factors that are part of the total cause of the fire (e.g., the presence of flammable material, absence of sprinklers), and the absence of any of them (in the circumstances that actually existed) would also have been enough to prevent the fire. In other words, localized events (such as the short circuit) that are normally identified as the causes of later events (houses burning down) underdetermine them; considered just in themselves, they are not enough to causally guarantee the occurrence of the effect.

Just as the causal overdetermination of past events by localized present events explains the practice of historical science, so the causal underdetermination of future events by localized present events explains the practice of experimental science. The test conditions brought about in the laboratory are only partial causes of what subsequently occurs. There is a need to ferret out and control for additional causal factors; otherwise, the ostensible confirmations and disconfirmations of the target hypothesis may be mistaken. This is why experimental scientists spend so much time methodologically rejecting auxiliary assumptions that they previously accepted. They are not trying to disprove their hypotheses or to save them from falsification. They are trying to identify false positives and false negatives, which are always a threat because the test conditions brought about in the laboratory are normally only a small part of the total cause of an experimental result. In brief, the activity of experimental scientists is best interpreted as an attempt to circumvent the inevitable causal underdetermination of experimental results by test conditions derived from target hypotheses.

SUMMARY

When it comes to testing hypotheses, historical science is not inferior to classical experimental science. Traditional accounts of the scientific method cannot be used to support the superiority of experimental work. Furthermore, the differences in methodology that actually do exist between historical and experimental science are keyed to an objective and pervasive feature of nature, the asymmetry of overdetermination. Insofar as each practice selectively exploits the differing information that nature puts at its disposal, there are no grounds for claiming that the hypotheses of one are more securely established by evidence than are those of the other.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the National Aeronautics and Space Administration to the University of Colorado's Astrobiology Institute. I thank Sheralee Brindell, Bruce Jakosky, and Gifford Miller for helpful discussions and comments on an earlier draft of this paper.

REFERENCES CITED

Alvarez, L.W., Alvarez, W., Asaro, F., and Michel, H.V., 1980, Extraterrestrial cause for the Cretaceous-Tertiary extinction: Science, v. 208, p. 1095–1108.

Chamberlin, T.C., 1897, The method of multiple working hypotheses: Journal of Geology, v. 5, p. 837–848.

Clemens, W.A., Archibald, J.D., and Hickey, L.J., 1981, Out with a whimper not a bang: Paleobiology, v. 7, p. 293–298.

Gee, H., 1999, In search of deep time: New York, The Free Press, 267 p.

Hoffman, P.F., and Schrag, D.P., 2000, Snowball Earth: Scientific American, v. 282, p. 2–9.

Kaufman, W., 1977, The cosmic frontiers of general relativity: Boston, Little, Brown, 306 p.

Kuhn, T., 1970, The structure of scientific revolutions: Chicago, Illinois, University of Chicago Press, 210 p.

Kump, L.R., Fasting, J.F., and Crane, R.G., 1999, The Earth system: Englewood Cliffs, New Jersey, Prentice-Hall, 351 p.

Lewis, D., 1991, Counterfactual dependence and time's arrow, *in* Jackson, F., ed., Conditionals: Oxford, UK, Oxford University Press, p. 46–75.

Miller, S.L., 1953, A production of amino acids under possible primitive Earth conditions: Science, v. 117, p. 528–529.

Mojzsis, S.J., Arrhenius, G., McKeegan, K.D., Harrison, T.M., Nutman, A.P., and Friend, C.R.L., 1996, Evidence for life on Earth before 3,800 million years ago: Nature, v. 384, p. 55–59.

Orgel, L.E., 1998, The origin of life—A review of facts and speculations: Trends in Biochemical Science, v. 23, p. 491–495.

Popper, K., 1963, Conjectures and refutations: London, Routledge, Kegan Paul, 431 p.

Price, H., 1996, Time's arrow and Archimedes' point: Oxford, UK, Oxford University Press, 306 p.

Manuscript received February 20, 2001
Revised manuscript received June 11, 2001
Manuscript accepted June 28, 2001

Printed in USA