RESEARCH · PHILOSOPHY AND PARADIGMS

The Scientific Method: Hypothesis, Test and Revision

How the scientific method actually works: forming hypotheses, deriving predictions that could fail, testing them fairly, and revising ideas in light of the results.

ModQu research article · 1,385 words · 15 references (APA 7th edition)

School textbooks often present the scientific method as a fixed recipe: observe, hypothesise, experiment, conclude. Working scientists rarely follow such a neat sequence, and philosophers have argued for a century about whether there is a single method at all (Chalmers, 2013). Yet beneath the variety there is a recognisable core. Researchers propose explanations, work out what those explanations predict, check the predictions against evidence that could prove them wrong, and change their ideas when the evidence demands it. This article describes that cycle, the reasoning behind each step, and the ways in which it can go wrong.

The cycle of inquiry

Box (1976) described science as an iterative process in which theory and data continually correct one another: a conjecture suggests what data to collect, the data reveal flaws in the conjecture, and a better conjecture follows. The cycle below is a simplified picture of that process.

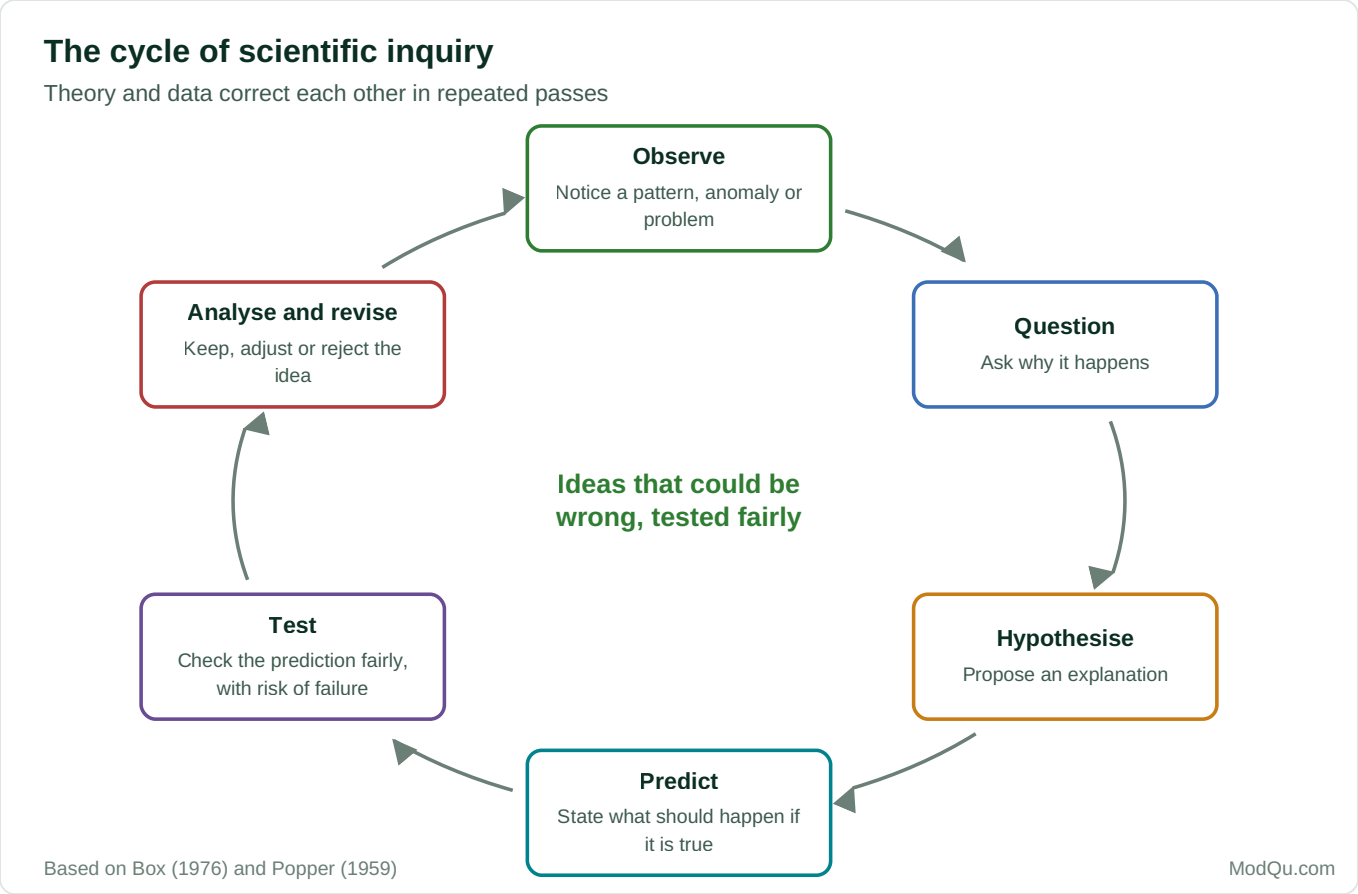


Figure 1. The cycle of scientific inquiry. Each pass refines the hypothesis or replaces it. Based on Box (1976) and Popper (1959); diagram by ModQu.

Observation and question

Inquiry usually begins with something that needs explaining: an anomaly, a pattern or a practical problem. Hempel (1966) used the example of Ignaz Semmelweis, a physician in 1840s Vienna who noticed that women in one maternity ward were dying of childbed fever far more often than those in a neighbouring ward. The observation was clear; the question was why.

Hypothesis

A hypothesis is a proposed explanation that goes beyond the observations. Semmelweis considered and rejected several possibilities, including overcrowding and the effect of a priest's procession, before proposing that doctors who came straight from dissecting corpses carried "cadaveric matter" on their hands into the first ward. Hypotheses can arise from theory, from analogy, from a lucky observation, or from systematic reasoning about what would best explain the facts. Haig (2005) argued that this last route, often called abduction, deserves a formal place in research methods because generating good hypotheses is as much a part of science as testing them.

Prediction

A hypothesis becomes testable when it implies something specific that can be observed. If cadaveric matter caused the deaths, then washing hands in chlorinated lime before examinations should reduce them. The prediction must be derived before the test is run, and it must be one that could turn out to be false. Popper (1959) made this the centre of his account of science: a theory is scientific only if it forbids certain outcomes, and the more it forbids, the more it says about the world.

Test

The test is a deliberate attempt to see whether the prediction holds. Semmelweis introduced handwashing, and mortality in the first ward fell sharply. In modern research, tests are usually designed to rule out alternative explanations. Fisher (1935) showed how random assignment of treatments allows a fair comparison, because it prevents the researcher's choices or hidden differences between groups from producing the result.

A famous test in physics shows how a theory can be put at risk. Einstein's general theory of relativity predicted that starlight passing close to the sun would be bent by a specific amount, about twice the value expected from Newtonian theory. Observations during the solar eclipse of 1919 were reported to match Einstein's figure (Dyson et al., 1920). The prediction could have failed, and that is why its success was taken so seriously.

Analysis and revision

Results rarely settle a question cleanly. A test may support a hypothesis, contradict it, or produce an ambiguous result that calls for a better test. When results contradict a prediction, researchers must decide what to revise. Quine (1951) pointed out that a hypothesis is never tested alone: every test also relies on assumptions about the instruments, the sample and the conditions. A failed prediction shows that something in this bundle is wrong, but not necessarily which part.

Falsification and its limits

Popper's emphasis on falsification remains influential, but it has been qualified in important ways. Kuhn (1962) observed that scientists do not abandon a successful theory at the first contrary result; they treat anomalies as puzzles to be solved within the existing framework, and only a build-up of unresolved anomalies leads to a change of framework. Lakatos (1970) proposed a middle position: scientists work within research programmes that protect a core of central assumptions with a belt of auxiliary hypotheses. A programme is progressive if its adjustments lead to new predictions that are confirmed, and degenerating if its adjustments merely patch over failures.

For practical researchers, the lesson is not that hypotheses can never be rejected, but that rejection requires judgement. A single failed test may reflect a weak measure or an unusual sample. A consistent pattern of failures across well-designed tests is a strong reason to revise.

Strong inference

Platt (1964) argued that the fields making the fastest progress were those that practised what he called strong inference: devising several alternative hypotheses, designing a crucial experiment whose possible outcomes would exclude one or more of them, carrying it out cleanly, and repeating the process. The approach is simple but demanding. It forces researchers to state what each explanation predicts differently, rather than gathering data that would fit any of them.

The same idea applies in applied settings. A business that sees sales falling might consider several explanations: a competitor's price cut, a change in customer needs, or a problem with distribution. Each predicts a different pattern in the data. A price-driven decline should be concentrated where the competitor is active; a distribution problem should appear as stock shortages regardless of competition. Designing the analysis to tell these apart is far more informative than collecting general data and hoping the answer emerges.

When the method breaks down

The cycle works only when each step is carried out honestly. Two common failures undermine it.

The first is changing the hypothesis to fit the results while presenting it as if it had been predicted. Kerr (1998) called this HARKing: hypothesising after the results are known. It turns an exploratory finding into what appears to be a successful test, so that the evidence seems stronger than it is.

The second is analytical flexibility. If researchers try many ways of analysing data and report only the one that supports the hypothesis, the test is no longer a genuine risk to the idea. Ioannidis (2005) argued that such practices, combined with small studies, mean that many published findings are likely to be false. Large-scale replication projects have since found that many published effects are smaller or less consistent than originally reported (Open Science Collaboration, 2015).

One response has been pre-registration, in which researchers record their hypotheses and analysis plans before seeing the data. Nosek et al. (2018) argue that this restores the distinction between testing predictions and exploring data, both of which are valuable but support different kinds of conclusions.

Good hypotheses

Not every hypothesis is equally useful. A good hypothesis is specific enough to be tested, states a relationship between identifiable variables, and implies results that would count against it. "Customer service affects loyalty" is too vague to test well. "Customers whose complaints are resolved within 48 hours are more likely to renew than comparable customers whose complaints take longer" names the variables, the direction of the effect and the comparison, and it could clearly turn out to be false.

Good hypotheses also connect to wider theory. A prediction that follows from an established explanation tells us something about that explanation when it is tested, while an isolated guess tells us only about itself. Popper (1959) argued that bold hypotheses, which risk more, are more valuable than cautious ones because they have more ways to fail and therefore say more when they survive.

Conclusion

The scientific method is less a fixed procedure than a set of commitments: propose explanations that could be wrong, derive predictions from them before looking, test them fairly, and revise in light of the results. Semmelweis's handwashing, the eclipse observations and modern clinical trials all follow this logic. The commitments apply equally to commercial and survey research. A study that states its expectations in advance, designs its comparisons to distinguish between explanations, and reports what it finds whether or not it was hoped for is using the scientific method, whatever field it belongs to.

References

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Chalmers, A. F. (2013). *What is this thing called science?* (4th ed.). Hackett.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, 220*, 291–333. https://doi.org/10.1098/rsta.1920.0009

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*(1), 20–43. https://doi.org/10.2307/2181906