2 DEFINING THE SCIENTIFIC METHOD

The invitation for those nominating candidates for the Nobel Prize in economics, the "Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel," described the award of the prize as being "based solely on scientific merit." No criteria for judging scientific merit were provided, but nominators were directed to "consider origin and gender" of the nominees. Without clear criteria for the award, to what extent can one be confident that the prize was based on the scientific merit of the findings?

In this chapter we provide an aspirational definition of the scientific method. The definition is in the form of eight criteria that are based on the writings of key figures in the development of the scientific method. We then expand on each of the criteria, describing their source – where appropriate – and the reasons for their importance for the scientific method.

2.1 An Aspirational Definition

We sought to define the scientific method in such a way that most researchers *should* aspire to the ideal the definition represents. To do so, we turned to the writings of the developers of the scientific method. Scientists have been describing elements of the scientific method since before 400 BC. White (2002) concluded that the modern scientific method owes its approach to the logical framework of hypothesis testing laid out by Socrates, with later refinements by Plato and Aristotle. Socrates in effect set out the basis of a valid approach

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

to seeking knowledge that scientists still use – the use of experiments, which came to be formally recognized as important much later, excepted.

We concluded that the key elements of the scientific method – as derived from the words of famous and pioneering scientists – could be summarized by eight criteria:

1. Study important problems
2. Build on prior knowledge
3. Provide full disclosure
4. Use objective designs
5. Use valid and reliable data
6. Use valid simple methods
7. Use experimental evidence
8. Draw logical conclusions

These criteria are also consistent with the *Oxford English Dictionary (OED)*, which defines the scientific method as:

> commonly represented as ideally comprising some or all of (a) systematic observation, measurement, and experimentation, (b) induction and the formulation of hypotheses, (c) the making of deductions from the hypotheses, (d) the experimental testing of the deductions, and (if necessary) (e) the modification of the hypotheses … The modern scientific method is often seen as deriving ultimately from Francis Bacon's Novum Organum (1620) and the work of Descartes. In the 20th century, Karl Popper's idea of empirical falsification has been important. OED Online (2018).

In practice, a study can *contribute* to making a useful scientific discovery even when it does not on its own comply with all of the criteria. For example, Einstein drew on the findings of others' experiments to develop novel hypotheses about important problems that could in turn be tested against alternative hypotheses by further experiments.

Papers might also contribute to science by identifying important problems. Others might contribute by identifying shortcomings in the papers of other researchers and resolving those issues. Another contribution is to develop objective measures of important variables, and compile data using those measures, as has been done by scientists at the University of Alabama at Huntsville in estimating global average temperatures from satellite readings (Spencer et al., 2017).

While studies that fall short on some criteria – e.g., by overlooking prior knowledge – might nevertheless turn out to provide a useful contribution to research on a problem, studies that failed to use an objective design (criterion 4) are unlikely to do so. In order to claim that a principle or method is scientific, studies of the problem would, when taken together, need to satisfy all eight criteria.

We consider that the support of meta-analyses of *objective* studies that collectively comply with *all eight criteria for science* are necessary for rational policy making. The requirement is particularly important for government laws and regulations, because they involve duress rather than voluntary transactions.

2.2 Criteria for Complying with the Scientific Method

We now expand on the eight criteria for complying with the scientific method that we described above.

2.2.1 Study Important Problems

> According to the general spirit of this book, which values everything in its relation to Life, knowledge which is altogether inapplicable to the future is nugatory.
>
> Charles Sanders Peirce
> (1958, para 56)

Scientists in the past sought to address important problems. Robert Boyle, a founder of the English Royal Society, wrote in 1646 that the founders valued "no knowledge but that it has a tendency to use" (as quoted by O'Connor and Robertson, 2004).

Some scientists argue that research that does not obviously lead to useful findings is nevertheless important because of potential future usefulness. While that may turn out to be true in some cases, identifying problems that are currently in need of solutions to research is more likely to produce useful findings than is research based on curiosity about a non-problem.

Addressing currently pressing problems can lead, and has led, to advances in scientific knowledge that go well beyond finding solutions to those problems, as the following quotation illustrates.

> [T]he practical sciences incessantly egg on researches into theory. For considerable parts of chemical discovery we have

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

> to thank the desire to find a substitute for quinine or to make quinine itself synthetically, to obtain novel and brilliant dyestuffs, and the like. The mechanical theory of heat grew out of the difficulties of steam navigation. For it was first broached by Rankine while he was studying how best to design marine engines. Then again, one group of scientists sometimes urges some overlooked phenomenon upon the attention of another group. It was a botanist who called van't Hoff's attention to the dependence of the pressure of sap in plants upon the strength of the solution, and thus almost instantaneously gave a tremendous impulse to physical chemistry. In 1820, Kästner, a manufacturer of cream of tartar in Mulhouse, called the attention of chemists to the occasional, though rare, occurrence in the wine casks of a modification of tartaric acid, since named racemic acid; and from the impulse so given has resulted a most important doctrine of chemistry, that of the unsymmetric carbon atom, as well as the chief discoveries of Pasteur, with their far-reaching blessings to the human species. Charles Sanders Peirce (1958, para 52)

If research on relatively narrow current problems can lead the curious scientist to such widely important discoveries as are described in the quotation from Peirce (1958) above, the case for studying non-problems at someone else's expense seems weak when researcher time is a limited resource. Of course, if there is a willing well-informed funder for such activity, including self-funding, then that is the business of the parties concerned, and good luck to them.

2.2.2 Build on Prior Knowledge

Progress in science requires that scientists become familiar with prior knowledge and methods for the given problem. Newton (1675) referred to the process as "standing on the shoulders of giants."

Despite the logical necessity of doing so, researchers often fail to comprehensively review the existing evidence, perhaps because doing so greatly increases the time needed to complete a publication. Because the reviewers used by journal editors are often unaware of relevant prior scientific findings, an author's failure to identify relevant prior research can go undetected. As a consequence, researchers are prone to making *re*discoveries.

In one example, Kahneman (2011) concluded that people process information differently depending on the nature of the decision. He referred to the phenomenon as "slow versus fast," or "System 1" and "System 2" decision-making. His was at least the third discovery of the concept. In 1913 it was called "short circuit versus long-circuit" thinking as described by Hollingworth (1913). Half-a-century later, the concept was referred to as "low involvement versus high-involvement" by Krugman (1965). Whatever name the concept is given, it has been an important condition to consider for persuasion for over a century now. For more on this, see Armstrong (2010, pp. 21–22).

2.2.3 Provide Full Disclosure

The scientific method depends heavily on replication, and replication requires full disclosure of methods. Replications are needed to help determine whether potentially useful scientific findings should be accepted and acted upon.

A paper that does not provide all necessary information for replication may, nevertheless, contribute to science if it at least addresses an important problem. Other researchers can conduct *extensions* that test the same issue. The extensions can help to allay concerns about findings that arise when disclosure is incomplete.

2.2.4 Use Objective Designs

The founders whose writings we used to develop the definition of the scientific method recognized early on that objectivity is hard to achieve. They also recommended a solution. Sir Isaac Newton, for example, described four "Rules of Reasoning in Philosophy" in the third edition of his Philosophiae Naturalis Principia Mathematica (1726, pp. 387–389). His fourth rule, in Motte's translation from Latin, states, "In experimental philosophy we are to look upon propositions collected by general induction from phenomena as accurately or very nearly true, *notwithstanding any contrary hypotheses that may be imagined*, till such time as other phenomena occur, by which they may either be made more accurate, or liable to exceptions" (Newton, 1729, vol. 2, p. 205, emphasis added).

We refer to this solution as Multiple Reasonable Hypotheses Testing, or MRHT. One should include all reasonable hypotheses or

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

describe why that was not feasible. MRHT stands in contrast to the approach that has become accepted practice in psychology and the social sciences: Null Hypothesis Statistical Testing, or NHST.

The increase in productivity that arose from the English Agricultural Revolution illustrates the importance of MRHT. Agricultural productivity saw little improvement until landowners in the 1700s began to conduct experiments comparing the effects of alternative ways of growing crops. The Industrial Revolution progressed in the same manner (Kealey, 1996, pp. 47–89).

Chamberlin (1890) claimed that disciplines that conduct experiments to test multiple reasonable hypotheses progress greatly, while those that do not, progress little. Nearly three-quarters of a century later, Platt (1964) reiterated Chamberlin's conclusion because researchers in many fields of science were still ignoring the original advice.

MHRT has also led to advances in medical knowledge. For example, one study examined all papers that used MRHT that were published in the *New England Journal of Medicine* from 2001 to 2010 (Prasad et al., 2013). The study found that 146 medical treatment recommendations were reversed as a consequence of experiments using MRHT. The reversals amounted to 40 percent of all procedures tested. MRHT has also led to the growth of useful knowledge in engineering, forecasting, persuasion, and technology.

2.2.5 Use Valid and Reliable Data

Validity is the extent to which the data measure the concept that they purport to measure. Validity is not a trivial matter. Many disputes arise due to differences in how concepts are measured. For example, what is the best way to measure inequality among people? Is it best assessed only in terms of money income, or should it also include the effects of taxes, wealth, transfer payments, home production, etc.? These measures produce different findings and policies. More fundamentally, should inequality be assessed in terms of life satisfaction instead of income? Money income is, after all, only one of several means to achieve the desired end of happiness. People routinely trade off money income to do work that provides greater intrinsic satisfaction or to live somewhere that they prefer.

Reliability is established when other researchers, using the same procedures, can reproduce findings. Reliability can be improved by using all relevant data that are available such as when using a time-series. As Sir Winston Churchill said, "The longer you can look back, the farther you can look forward."

Data that has been subject to unexplained revisions should not be used. Enough said.

2.2.6 Use Valid Simple Methods

> There is, perhaps, no beguilement more insidious and dangerous than an elaborate and elegant mathematical process built upon unfortified premises.
>
> Chamberlin (1899, p. 890)

Validity requires that the method used has been tested and found to be useful for the problem at hand. Simple methods are those that can be understood by those who might have an interest in reading or replicating the paper. Complex methods make it difficult for others to understand the research, spot errors, and replicate the study.

The call for simplicity in science started with Aristotle but is usually attributed to Occam as "Occam's Razor." Yet, academics and consultants love complex methods. So do their clients. After all, if the process were simple they would ask, "Why are we paying all that money?" For a further discussion of why complexity proliferates, see Hogarth (2012).

The 1976 Nobel Laureate in Economics, Milton Friedman, stressed the importance of testing the predictive validity of hypotheses against new, or out-of-sample, observations (1953). Is there a conflict between predictive validity and simplicity? Apparently not. Comparative studies have shown the superior predictive validity of simple methods in out-of-sample tests across diverse problems. The experiments on the predictive validity of simple alternatives to multiple regression analysis by Czerlinski et al. (1999), and by Gigerenzer et al. (1999) are elegant examples.

In our review of the evidence on the predictive validity of Occam's Razor, we defined a "simple method" as one for which an intelligent person could understand: (a) procedures; (b) representation of prior knowledge; (c) relationships among the elements; and (d)

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

relationships among models, predictions, and the decisions that might be made (Green and Armstrong, 2015). We found 32 published studies that compared forecasts from simple methods with forecasts from more complex methods that had been proposed by their authors as a way to improve accuracy. We hired university students to rate complexity against the simplicity criteria listed above. Simplicity improved out-of-sample predictive validity in all 32 studies involving 97 experimental comparisons. On average, complex methods had errors for out-of-sample predictions that were 27 percent larger for the 25 papers that provided quantitative comparisons. The strength and consistency of the findings astonished us and are a caution to researchers who assume that complex data modelling methods have predictive validity.

2.2.7 Use Experimental Evidence

> The testing of the hypothesis proceeds by deducing from it experimental consequences almost incredible, and finding that they really happen, or that some modification of the theory is required, or else that it must be entirely abandoned.
>
> These experiments need not be experiments in the narrow and technical sense, involving considerable preparation. That preparation may be as simple as it may. The essential thing is that it shall not be known beforehand ... how these experiments will turn out. Charles Sanders Peirce (1958, paras 83, 90)

Experiments emerged as a key element of the scientific method in the practice of the natural sciences in the sixteenth century. The importance of experiments was generally not recognized in medical research and the social sciences until the nineteenth century (DiNardo, 2018).

Robert Boyle and other scientists established the forerunner of the modern-day Royal Society around 1645 to acquire knowledge through experiments. The value the society placed on experiments was highlighted by the appointment of Robert Hooke as a Curator of Experiments who was tasked with "furnish[ing] them every day on which they met with three or four considerable experiments" (O'Connor and Robertson, 2004). The society translates its Latin motto, *nullius in verba*, as "take nobody's word for it." It expresses the Royal Society Fellows' determination "to withstand domination of

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

authority and to verify all statements by an appeal to facts determined by experiment" (Royal Society, 2019).

Experiments can be controlled, quasi-controlled – include some, but not all, important causal variables – or natural. Laboratory experiments allow for more control over conditions, while field experiments are more realistic. Interestingly, a comparison of findings from laboratory versus field experiments in 14 areas of organizational behavior concluded that they produced similar findings (Locke, 1986). Vernon Smith demonstrated that "laboratory" (controlled) experiments can be used to test competing hypotheses in economics. He found that very simple experiments could be devised that would replicate the relevant behaviors of participants in real markets (Smith, 2002).

Experiments have been conducted in fields of science as diverse as astronomy (e.g. Ostro, 1993, described the use of radar to conduct experiments on the scale of the solar system and gravitation, among other things), evolutionary biology (e.g., Schluter, 1994, conducted experiments to test theories about the effect of resource competition among species on evolution), geology (Kuenen, 1958, described the use of experiments in geology starting with those of Sir James Hall, who began conducting his experiments in 1790), paleontology (e.g., Oehler, 1976, described experiments that simulated fossilization in synthetic chert), and zoology (e.g., Erlingsson, 2009, described the rise of experimental zoology in Britain during the 1920s).

Darwin is most famous for his theory of evolution, but he also devoted much time to testing hypotheses with experiments. For example, he hypothesized, contrary to then current belief, that plants move, and designed experiments that tracked plant movement (Hangarter, 2000). But not all research problems are amenable to testing by way of experiments that are controlled by the researcher, as Mayr (1997) described in his book on the science of biology: "Much progress in the observational sciences is due to the genius of those who have discovered, critically evaluated, and compared ... natural experiments in fields where a laboratory experiment is impractical, if not impossible" (p. 29).

Natural experiments have been used to test competing theories in the physical sciences; for example, Maupertuis's expedition to Lapland over the winter of 1736–1737 to undertake observations that would test the Cartesian theory that the earth is taller than it is broad against Newton's theory that the opposite is the case. More

famously, Eddington's 1919 expeditions were mounted to determine whether Einstein's or Newton's gravitation theories provided the better prediction of phenomena by taking advantage of the natural experiment provided by a solar eclipse (Sponsel, 2002).

Hypotheses on the distribution of plants from Darwin's speculations and findings from experiments on the survival and dispersal of plant seeds (Carlquist, 2009) were tested by the natural experiment of the 1883 eruption of the island of Krakatoa (Krakatau). The eruption sterilized what was left of the island such that most plant life – with the possible exception of some grasses – would have to have arrived on or over open sea. Nine months after the eruption, there was no sign of plant life, but by 1930 the whole island was covered with dense forest (Went, 1949).

Gould (1970) advocated greater use of experiments in paleontology – "we must include the experimental approach ... and not remain tied to the observational mode of traditional natural history" (p. 88) – and described prior studies that used natural experiments. He quoted Seilacher on the topic: "One cannot make experiments with organisms that became extinct hundreds of million years ago. Still, isn't it an experimental approach if the belemnites' habits were tested through the reactions of its commensals? The fact that the actual test was made long before man's existence does not alter the principles of its evaluation" (Gould, 1970, p. 89).

Variations between the societies of different countries, regions, states, and communities, and changes over time provide natural experiments against which researchers can test hypotheses from alternative theories. Diamond and Robinson's (2010) edited book *Natural Experiments of History* includes seven analyses of political and social arrangements and their economic outcomes or causes using natural experiments from history. Alternative arrangements for managing common pool resources provide natural experiments that allowed testing of hypotheses on whether sustainable management arrangements can arise by trial and error, or whether they must be imposed by a political authority (Ostrom, 1990). Variations in regulations between US counties and states, and over time, allowed Lott (2010) to test hypotheses on the relationship between gun control and crime.

Note that some scientists consider the term "natural experiments" to be only a metaphor for studies that literally test hypotheses by making observations, or "observational studies," and not *true*

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

experiments. We prefer to use the term "natural experiments" in order to distinguish studies that are properly designed to test alternative hypotheses by identifying situations in which observations might turn out to falsify them, and reserve the term "observational studies" for studies that do not test hypotheses or that develop hypotheses to fit observations.

For ideas and guidance on designing experiments see Shadish, Cook, and Campbell's (2001) book *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. They describe diverse and creative ways to conduct experiments. Another resource is Dunning's (2012) book *Natural Experiments in the Social Sciences: A Design-Based Approach*, the first part of which is devoted to "discovering natural experiments."

Experiments guided by sound theoretical reasoning provide the only valid and reliable way to establish *causal relationships*. Causality cannot be identified by "machine learning" methods, known by names such as artificial intelligence, data mining, factor analysis, and stepwise regression. We described the lack of evidence that the models that are the product of machine learning methods have any predicted validity in our 2018 and 2019 co-authored papers.

Machine learning models violate the scientific method because they fail to incorporate prior knowledge from experimental studies and coherent theory. The models are also vulnerable to including variables that have no known causal relationship to the variable of interest. As economist Friedrich Hayek warned in his Nobel Prize lecture, "in economics and other disciplines that deal with essentially complex phenomena, the aspects of the events to be accounted for about which we can get quantitative data are necessarily limited and may not include the important ones" (Hayek, 1974).

Meta-analyses of experimental data are the gold standard of evidence. Meta-analyses combine the results of all experimental studies on the issue being studied, no matter the type of experiment. For example, a meta-analysis of 40 experiments on how communication affects persuasion found the conclusions from field and laboratory studies were similar (Wilson and Sherrell, 1993).

Findings from experimental studies do, however, often differ from those based on non-experimental data. For example, expert judgments and non-experimental research typically conclude that consumer satisfaction surveys improve consumer satisfaction. However,

well-designed experiments showed that they *harm satisfaction* because customers look for bad things to report. They also create dissatisfaction among those providing the services. The problems went away when people were asked *what they liked about the product or service* (Ofir and Simonson, 2001).

Non-experimental data from hundreds of thousands of users showed that female hormone-replacement therapy helped to preserve youth and ward off a variety of diseases in older women. The findings were replicated. However, subsequent experimental studies found that the treatment could actually be harmful. The favorable findings from the non-experimental data occurred because the women who used the new medicine were generally more concerned about their health and sought out ways to stay healthy (Avorn, 2004).

Kabat's (2008) book on environmental hazards – examining such topics as DDT, electromagnetic fields from power lines, radon, and second-hand smoke – concluded that analysis of non-experimental data in studies on health had often misled researchers, doctors, patients, and the public.

Non-experimental data analyses lend themselves to advocacy studies. They allow researchers to produce "evidence" for almost any hypothesis by attributing causal relationships to correlations in survey data.

Vernon Smith, a pioneer of experimental economics and a 2002 Nobel Laureate in Economics, suggested that what can be learned from well-designed laboratory experiments is only limited by the ingenuity and creativity of the researcher.

> What are the limits of laboratory investigation? I think any attempt to define such limits is very likely to be bridged by the subsequent ingenuity and creativity ... of some experimentalist. Twenty-five years ago I could not have imagined being able to do the kinds of experiments that today have become routine in our laboratories. Experimentalists also include many of us who see no clear border separating the lab and the field. Vernon Smith (2003, p. 474, n. 27).

There may be problems or situations for which experiments are not possible. In such cases, analyses of non-experimental data may be useful for helping to identify *whether hypothesized causal relationships are plausible*. For situations in which causal relationships have been established, analyses of non-experimental data can help to assess effect sizes.

Some philosophers of science have theorized that experiments cannot do what scientists expect them to: contribute to knowledge by rejecting or supporting hypotheses. As we hope is clear from this book, we disagree, strongly. Philosopher of science Deborah Mayo and practitioner of science Vernon Smith have also disagreed, as follows.

> In principle the D-Q problem[1] is a barrier to any defensible notion of a rational science that selects theories by a logical process of confrontation with scientific evidence. This is cause for joy not despair. Think how dull would be a life of science if, once we were trained, all we had to do was to turn on the threshing machine of science, feed it the facts and send its output to the printer. In practice the D-Q problem is not a barrier to resolving ambiguity in interpreting test results. The action is always in imaginative new tests and the conversation it stimulates. My personal experience as an experimental economist since 1956, resonates well with Mayo's critique of Lakatos:
>
> Lakatos, recall, gives up on justifying control; at best we decide – by appeal to convention – that the experiment is controlled … I reject Lakatos and others' apprehension about experimental control. Happily, the image of experimental testing that gives these philosophers cold feet bears little resemblance to actual experimental learning. Literal control is not needed to correctly attribute experimental results (whether to affirm or deny a hypothesis). Enough experimental knowledge will do. Nor need it be assured that the various factors in the experimental context have no influence on the result in question – far from it. A more typical strategy is to learn enough about the type and extent of their influences and then estimate their likely effects in the given experiment. Vernon Smith (2002, p. 106, quoting Mayo, 1996, p. 240)

2.2.8 Draw Logical Conclusions

Francis Bacon (1620 [1863]) reinforced Aristotle's assertion that the scientific method involves logical induction from systematic

[1] The Duhem-Quine problem is the assertion that designing an experiment to test a hypothesis is not possible without making assumptions or involving additional hypotheses that may themselves be the cause of the experiment's support for or rejection of the hypothesis (the authors).

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press

observation. Conclusions should follow logically from the evidence provided in a paper.

How might logic be used to compare competing hypotheses? Here is an example: compare the hypothesis that people in a given community in a rich country will be happier if the government redistributes money income from higher income people to those with lower incomes (Hypothesis #1), with the hypothesis that people in a community who are happier are more productive and earn more money (Hypothesis #2), and with the hypothesis that the happiness of people within a community is more affected by their relative status than by their absolute money income (Hypothesis #3). The latter hypotheses lead to policy conclusions that are opposite to the those from the first. Frey's (2018) summary of evidence from happiness research provides support for Hypothesis #2 and #3, and cautions against Hypothesis #1.

If the research addresses a problem that involves strong emotions, consider writing the conclusions using symbols in order to check the logic. For example, the argument "if *P*, then *Q*. Not *P*, therefore not *Q*" is easily recognized as a logical fallacy – known as "denying the antecedent" – but recognition is not easy for contentious issues, such as the relationship between guns and crime.

Violations of logic are common in the social sciences. We suggest asking researchers who have different views on the problem you are studying to check your logic. Logic does not change over time, nor does it differ by field. Thus, Beardsley's (1950) *Practical Logic* continues to be useful. For an additional discussion of logical fallacies, see the website www.logicalfallacies.org.

https://doi.org/10.1017/9781009092265.003 Published online by Cambridge University Press