Using History as Evidence in Philosophy of Science: A Methodological Critique

James W. McAllister
University of Leiden
j.w.mcallister@phil.leidenuniv.nl

Post-print version of 'McAllister J.W. (2018), Using History as Evidence in Philosophy of Science: A Methodological Critique, Journal of the Philosophy of History 12: 239–258,' available online at https://doi.org/10.1163/18722636-12341384

Abstract

This article offers a critical review of past attempts and possible methods to test philosophical models of science against evidence from history of science. Drawing on methodological debates in social science, I distinguish between quantitative and qualitative approaches. I show that both have their uses in history and philosophy of science, but that many writers in this domain have misunderstood and misapplied these approaches, and especially the method of case studies. To test scientific realism, for example, quantitative methods are more effective than case studies. I suggest that greater methodological clarity would enable the project of integrated history and philosophy of science to make renewed progress.

Keywords

case study methodology – history of science – philosophy of science – qualitative methods – quantitative methods – scientific realism

1. Introduction

Many philosophers of science agree that we must seek some degree of empirical support or confirmation for philosophical theories and models of science in actual science.[1] Since actual science is a historically constituted phenomenon, this project amounts to trying to evaluate philosophical claims about science by appeal to historical findings. This evidential relation is a central plank of most proposals for integrating history and philosophy of science.[2]

This article reviews some past attempts and some possible methods to use history as evidence in philosophy of science. I find some confusion in the methodological statements and practices of several writers in philosophy of science since the 1970s. In particular, I find a neglect of quantitative methods even where the evaluation of the philosophical claims at issue called for such methods; I find that several writers have invoked qualitative methods, and particularly case study methodology, in projects for which they were not appropriate; and I find that, in any event, many devices called "case studies" have not met the criteria for case studies that social scientists acknowledge.

I begin in section 2 by seeking to learn from methodological debates in the social sciences, and especially from the distinction between quantitative methods, which include experimental and quasi-experimental designs, and qualitative methods, of which case study methodology is the prime example. In section 3, I look at quantitative methods in history and philosophy of science. Whereas there are some good examples of the use of these methods in science studies, philosophers of science have been comparatively unwilling to pursue them.

In section 4, I turn to qualitative methods, and especially to case study methodology, in history and philosophy of science. Despite the fact that many writers see "case studies" as the only means to bring historical data to bear on debates in philosophy of science, I find that the device is poorly understood. I discuss the use of historical data to test scientific realism in section 5. I argue

that the project of evaluating the plausibility of scientific realism by appeal to history of science demands a quantitative approach: unclarity about the required methodology helps to explain why this project has so far been inconclusive. I draw, despite everything, some optimistic conclusions in section 6.

2. Qualitative and Quantitative Methods in Social Science

Research methodology in social science is extensively analysed and codified. Researchers in the social sciences have distinguished, in the first instance, quantitative and qualitative methods. This distinction has proved an enduring key for mapping methodological diversity in social science.[3]

Quantitative methods use cross-case analysis to perform causal inference. These methods usually involve collecting numerical or other quantitative data on relatively few variables from relatively many instances. The experimental method in social science is an example. Qualitative methods, by contrast, use within-case analysis to reconstruct causal pathways in an individual case. These methods, such as the case study method, involve in-depth study of relatively many variables in relatively few instances.

This difference becomes visible in specific research tools in the social sciences. Whereas both approaches may gather data by means of questionnaires, for example, they will use questions of different sorts. The questions in a quantitative research project will offer respondents a limited range of options, from which researchers can easily derive aggregate findings. Qualitative researchers, by contrast, will pose open-ended questions that invite exposition and elaboration, from which they will try to understand how individual respondents think, feel or behave in particular situations.

The difference between quantitative and qualitative methods manifests itself in various dimensions. Two of these, which I shall call context and direction of causal inference, are of special relevance to our topic.

"Context" alludes to the distinction between contexts of discovery and of justification proposed by Hans Reichenbach and others. There is a trade-off to some extent between effectiveness of methods in discovery and in justification. Qualitative methods are suited to generating hypotheses on the strength of in-depth knowledge of a few cases. They thus provide a basis for tentative inductive inferences and are associated with exploratory research.

Qualitative methods are less suited to testing general hypotheses, by contrast. At the extreme, a single case study is clearly a poor basis for establishing whether a given hypothesis holds across a population. A single case can refute a hypothesis, one might think; but this is possible only for hypotheses that assert that one factor is a necessary or a sufficient condition for another, and such hypotheses are rare in the social sciences. More broadly, qualitative methods are vulnerable to the criticism that the cases chosen have been cherry-picked, and therefore the representativeness and generalisability of any conclusions may be disputed. These considerations amount to saying that the results of qualitative methods have, in principle, high internal but uncertain external validity.

The opposite, to some extent, holds for quantitative methods. These are not well suited to generating hypotheses: the strategy of collecting quantitative data on relatively few variables provides a poor basis for reconstructing causal pathways. Whereas quantitative methods enable us to detect correlations, moving beyond those to causal statements is more difficult. Quantitative methods are better suited to testing hypotheses. A hypothesis that posits a certain causal mechanism will predict a certain correlation in quantitative data, and this prediction can be tested effectively in a cross-case analysis. Data in numerical form are more easily analysed for patterns of association and for strength and significance of correlations.

Let us now turn to the second dimension, direction of causal inference. Both quantitative and qualitative methods are able to ground causal claims, but they tend to work in opposite directions.

Quantitative methods are particularly well suited to answering questions about the effects that follow from given causes. Researchers using these methods generally assume that a particular cause has operated in the set of instances under study, and gather data about the effect of that cause in the population. In other words, they study the consequences of a given intervention in a system. These are known as forward causal inference or effects-of-causes questions.[4]

Take as an example an experimental design that involves comparing subjects that receive a treatment with members of a control group. This approach is designed to gauge the average or typical effect of the treatment on members of a population. To do this, the researcher tries to isolate the effect of the treatment: for example, randomly assigning subjects to treatment and control groups neutralises other or confounding causes. This cross-case design does not aim to confirm that the treatment is the cause of the outcome for any particular subject, or to ascertain by what mechanism the treatment causes the outcome in that subject.

Qualitative methods, by contrast, are better suited to answering questions about the causes that have led to certain observed effects. Researchers using these methods usually start from a real-world case in which a certain effect is apparent, and inquire what caused it. In other words, one seeks to understand the causal relationship between an already observed outcome and hypothesised earlier interventions. These are known as reverse causal inference or causes-of-effects questions.

The case study is an example. John Gerring has defined this as "the intensive study of a single case where the purpose of that study is—at least in part—to shed light on a larger class of cases (a population)".[5] This will typically start by noting an outcome in the case under examination, and try to ascertain which of the many causal factors acting on the subject are responsible for causing that outcome. The researcher will carry out an over-time, processual analysis, retracing within-case causal interactions. An illustration is Theda Skocpol's case studies of the French, Russian, and

Chinese revolutions with the aim to reconstruct the causal factors determining social revolutions in general.[6] The case study method, however, is not designed to estimate a causal factor's average or typical effect in the wider population: a study of an individual case is unsuited to ascertain that.

If this simple analysis holds water, then the social science toolbox contains, among other items, two sets of approaches with complementary strengths. Quantitative methods are suited to addressing effects-of-causes questions and to testing hypotheses in the context of justification; qualitative methods, by contrast, are suited to addressing causes-of-effects questions and to generating hypotheses in the context of discovery.

Before closing, let us review a more extensive example of the use of qualitative and quantitative methods: John Snow's work to establish the transmission medium of cholera during an outbreak of the disease in London in the 1850s. This example also illustrates the productiveness of "mixed methods" or "triangulation of methods", namely the integration of quantitative and qualitative approaches in a single research project, which many recent writers in methodology of social science have emphasised.[7]

In a first phase, aiming at hypothesis generation, Snow posed a causes-of-effects question: what causal mechanisms operated in people who contracted and in people who did not contract cholera? Choosing a qualitative approach, Snow conducted a within-case causal analysis or case study of a small number of Londoners. Snow established that, in people who had contracted cholera, the causal agent seemed to attack the alimentary canal first. He further established that few residents of buildings with a private water supply showed the disease. On this basis, he formulated the hypothesis that the transmission medium was contaminated water.

In a second phase, aiming to test this hypothesis, Snow posed an effects-of-causes question: what was the incidence of cholera in households that received water from various sources? Now pursuing a quantitative approach, Snow conducted a cross-case quasi-experiment, gathering just a couple of items of information from a large number of Londoners. The results

not only confirmed the hypothesis that Snow had formulated in his qualitative investigation, but also famously pinpointed the source of contaminated water in the Broad Street pump.[8]

Writers in history and philosophy of science have, of course, extensively studied both quantitative and qualitative methods in a variety of scientific disciplines. In their work, we find in particular a clear-sighted understanding of case study methodology, its sophistication, and the reasons why it is especially suited to hypothesis generation in fields ranging from psychoanalysis to social science.[9] As we will see later, however, writers in history and philosophy of science who have discussed the use of case study methodology in their own discipline have not shown the same grasp and appreciation of the method.

3. Quantitative Methods in History and Philosophy of Science

Let us now look at the use of quantitative methods (this section) and qualitative methods (section 4) in the project to bring historical findings to bear as evidence for and against philosophical claims about science.

Sociologists and historians of science have long used quantitative methods to test general claims about science against historical evidence. Derek J. de Solla Price, for example, tested the hypothesis that science was growing exponentially against various quantitative indicators, such as numbers of publications and doctorates, since 1660.[10] He reported that the historical data confirmed the hypothesis, and that they additionally indicated a doubling period of ten to fifteen years. Price's work laid the foundations for quantitative science studies and scientometrics. Arnold Thackray and Roger Hahn surveyed similar early quantitative work in the discipline of history of science.[11] We can expect the rise of "digital humanities" approaches to foster more quantitative work in history of science in future.[12]

What about philosophy of science? We normally think of philosophy as a qualitative discipline, but in fact many claims in philosophy of science are implicitly quantitative or have quantitative implications. This is because many such claims pertain to parameters that come in degrees, such as empirical success, accuracy, simplicity, strength of evidence, degree of truth approximation, and unexpectedness of empirical findings. This feature makes it appropriate to use quantitative methods to test many philosophical claims about science against historical evidence.

Several groups of researchers have tested in this way "Planck's principle", the hypothesis that younger scientists are quicker than older ones to accept novel theories. This hypothesis is philosophically significant because, as David L. Hull, Peter D. Tessner, and Arthur Diamond pointed out, it suggests that "external" or social factors play an important role in determining scientists' beliefs—perhaps outweighing "internal" or cognitive factors, such as evidence and argument.[13]

Hull, Tessner, and Diamond tested the specific hypothesis that, in Britain in the ten years following the publication in 1859 of Charles Darwin's *The Origin of Species*, younger scientists accepted evolution of species more rapidly than older scientists. They collected two pieces of data for each of 67 scientists who were born before 1839 and lived until at least 1869: their age in 1859, and their age at the time of the earliest evidence (if any) that they had accepted evolution. Hull, Tessner, and Diamond found that the mean ages in 1859 of scientists who accepted and who did not accept evolution by 1869 were 39.6 and 48.1 years respectively (a statistically significant difference), but that age explained less than ten percent of the variance in acceptance, and that among scientists who had accepted evolution by 1869, older scientists were as quick to change their minds as younger scientists. They concluded that Planck's principle was, on balance, disconfirmed.

Hull, Tessner, and Diamond's approach clearly belongs to the suite of quantitative methods used in social science. First, it operates in the phase of hypothesis test. Second, it tackles an effects-of-causes question, namely to

what extent scientists' ages influenced their propensity to adopt new theories. Furthermore, the authors performed cross-case analysis to carry out causal inference, but no within-case analysis: they made no attempt to ascertain the process by which any individual scientist adopted a theory. Lastly, of course, it deals with quantitative variables, namely ages.

Several subsequent writers have re-opened this research question and broader questions about the ages at which scientists make creative contributions, tackling them in similar ways.[14]

4. Qualitative Methods in History and Philosophy of Science

We turn now to qualitative methods used to bring evidence from history of science to bear on claims in philosophy of science.[15] The principal form of qualitative method in the domain of history and philosophy of science is the case study. Case studies in the proper sense, we may remind ourselves, are circumscribed exploratory studies that aim to yield insights for a broader category.

This device had a distinguished place in the early development of history and philosophy of science. In 1945, James B. Conant used a pedagogy of historical case studies for his "General Education in Science" course at Harvard University: "It is my contention that science can best be understood by the layman through a close study of relatively few case histories".[16] Conant's approach may have suggested to Thomas S. Kuhn, a lecturer in this course, that our understanding of science was best furthered by an intensive study of crucial historical transitions, such as the Copernican revolution—as well as, perhaps, that scientists learned their craft by studying exemplars of good science, rather than by applying general rules.[17]

In these instances, case studies were used to gain insight into the working of science, or in hypothesis generation. Many later writers, however, have claimed to use case studies to test philosophical claims about science.

This proposal has run into two problems. First, case studies are, by and large, inappropriate to the aim of testing general claims about science. This is because an examination of a small number of cases provides a slender basis for evaluating a general claim, especially if the claim has quantitative content. The second problem—which to some degree counteracts the first—is that few things called "case studies" in history and philosophy of science have actually been case studies in the sense established in methodology of social science.

We see these twin misunderstandings take root in the development of the discipline. In the 1970s Imre Lakatos proposed a theory of scientific rationality, dubbed "methodology of scientific research programmes"; simultaneously, he proposed that one could evaluate a theory of rationality by gauging the degree of precision with which it enabled one to reconstruct the history of science.[18] The test would consist in creating a "rational reconstruction" of episodes in the history of science on the basis of the theory of rationality at issue, and comparing that to the actual historical record. The greater the proportion of the historical record a theory of rationality construed as rational, according to Lakatos, the better that theory of rationality was. Testing a theory of rationality in this framework amounts clearly to posing an effects-of-causes question: what would the trajectory of science have been if a certain model of rationality had governed the reasoning and decision making of scientists in history?

This question calls for a quantitative method that performs a cross-case analysis. It is a sign of methodological unclarity, then, that Lakatos's colleagues and students described tests of his theory of rationality as "case studies". For example, Colin Howson explained that the volume, *Method and Appraisal in the Physical Sciences*, was dedicated to assessing "the fit between Lakatos's ideas and scientific practice" by means of "case studies drawn from the history of the physical sciences".[19] But a series of within-case analyses cannot perform an effective test of Lakatos's theory of rationality. To the contrary, the case studies that Howson assembled were more suited to posing causes-of-effects questions, or conducting within-case analysis to establish

which causal factors made the protagonists in the historical episodes act in the way they did. This amounted to appraising the protagonists in the historical episodes from the viewpoint of Lakatos's theory of scientific rationality.

In his chapter on nineteenth-century theories of heat, for example, Peter Clark explained what he would show:

> that the early kinetic programme was progressive […]. Then that thermodynamics, though a progressive research programme, had a limited heuristic […], and that the kinetic programme degenerated after 1880 […]. Finally I shall show that the kinetic programme became empirically progressive after 1905, with the prediction, for example, of the existence and magnitude of the Brownian motion.[20]

The promised evaluation of Lakatos's theory of rationality by appeal to history of science has turned into an appraisal of the progressiveness of various research programmes in the history of science by appeal to the conceptual apparatus of Lakatos's theory of rationality. This inversion is a natural consequence of Clark's choice of a qualitative methodology over a quantitative one, I suggest. Whereas a quantitative approach would have required the researcher to deduce the observable implications of Lakatos's model of rationality and assess whether they accorded with the historical record, the qualitative approach invited Clark to thematise the acts of past scientists as the *explanandum* and to show how the differential working of various causal factors could account for them. Historical data collected and treated in this way could in no way reveal Lakatos's model of rationality to be inadequate. This emphasizes that a quantitative method and the case study method are not equivalent or interchangeable.

A second major research project in historicist philosophy of science was the collaborative initiative, "Testing Theories of Scientific Change", based at Virginia Polytechnic Institute in the 1980s. First, Larry Laudan and his team collated some 250 hypotheses about scientific change put forward by Kuhn,

Paul K. Feyerabend, Lakatos, and Laudan.[21] Subsequently, Arthur Donovan, Larry Laudan, and Rachel Laudan commissioned tests of 32 of these hypotheses in the form of what they called "case studies".[22]

However, their methodology is badly flawed. There are five main criticisms. First, the claims that Donovan, Laudan, and Laudan targeted cannot be plausibly tested by a case study approach, but only by a quantitative approach. Virtually all 32 hypotheses under test posited a correlation between quantitative variables. One example is thesis GA4.5, "During a change in guiding assumptions (i.e., a scientific revolution), younger scientists are the first to shift and then conversion proceeds rapidly until only a few elderly holdouts exist". This is a version of Planck's principle—in section 3 above we saw Hull, Tessner, and Diamond test this thesis effectively by means of a quantitative approach. A second example is thesis T2.2, "The appraisal of a theory depends on its ability to solve problems it was not invented to solve":[23] this thesis too can be properly tested only by means of a quantitative method that performs a cross-case analysis. Donovan, Laudan, and Laudan's proposal to test such claims by means of case studies was misguided. As Hull noted shortly afterwards:

> Although case studies are in principle sufficient to *refute* a general thesis, in practice they rarely do so. Too many objections can be raised to their relevance, applicability, construction, execution, etc. They are not even in principle sufficient to *confirm* a general thesis. Rarely, however, are theses about science presented in a universal form. Usually they are hedged here and there. […] Systematic, preferably quantitative, studies are required to test claims such as these.[24]

Second, Laudan, Laudan, and Donovan misunderstood quantitative methods:

> We deliberately chose a historical case-study […] method rather than experiments, surveys or ethnomethodological studies […]. Our reasons for rejecting an experimental method are so obvious as to scarcely need explaining. Given our lack of control over the events that constitute scientific change and the impossibility of creating a situation in which we could manipulate such events, an experimental study of scientific change was out of the question.[25]

This explanation for neglecting quantitative methods is not convincing, however. Control over or manipulability of events is not a prerequisite of quasi-experimental methods in social science. Indeed, Hull, Tessner, and Diamond used a quasi-experimental method applied to a nineteenth-century population, for which there can be no suggestion of control or manipulation.

Third, Laudan, Laudan, and Donovan did not clearly distinguish case studies from quantitative methods. They explained why they felt that they had to commission new case studies: although in the previous literature "there are a handful of case studies which seek to apply various theories of scientific change to selected episodes in the natural sciences", "most of the avowed 'tests' would not pass muster on even the most tolerant view of robust experimental or quasi-experimental design".[26] This criticism of the previous literature is puzzling, since experimental and quasi-experimental methods belong to the family of quantitative methods, which is separate from the qualitative methods that include case study methodology. Case studies are not held to the same requirements as experimental or quasi-experimental methods.

Fourth, the things that Donovan, Laudan, and Laudan offered as "case studies" would not, in fact, be recognized as such by users of qualitative methods in social science. Take as an example C. E. Perrin's test of thesis GA4.5.[27] Perrin tested the hypothesis that, in eighteenth-century France, younger scientists switched from the phlogiston to the oxygen theory of combustion more rapidly than older scientists. He followed the procedure of

Hull, Tessner, and Diamond. He gathered two pieces of data on each of 69 chemists: their ages in 1785 and their ages at the time of the first known documentation showing that they had adopted key components of the oxygen theory. Perrin interpreted his findings as disconfirming thesis GA4.5. To call this a "case study" is a misrepresentation: it is an orthodox quantitative study. Most of the other chapters in the volume fall foul of the team's own criticism of previous attempts: "many of the avowed case studies are not 'tests' of the theory in question at all; rather, they are applications of the theory to a particular case".[28]

Fifth, the group confused hypothesis generation with hypothesis test, and the role that case studies can play in these two contexts. In a recent look back on the Virginia Polytechnic Institute initiative, Laudan and Laudan wrote:

> A basic premise of hyphenated history-and-philosophy-of-science is that theories of scientific change have to be based on empirical evidence derived from carefully constructed historical case studies. This paper analyses one such systematic attempt to test philosophical claims, describing its historical context, rationale, execution, and limited impact.[29]

The first sentence suggests plausibly that historical case studies can be a source of evidence for formulating theories of scientific change. The second sentence, however, shifts from hypothesis formulation to hypothesis test, suggesting that case studies can serve to test theories of scientific change—a more problematic assertion.

Perhaps because of this confused track record of the category of "case study" in history and philosophy of science, confidence in the method among writers on methodology of history and philosophy of science has fallen. Joseph C. Pitt provided one of the most trenchant rejections:

> What do appeals to case studies accomplish? Consider the dilemma: On the one hand, if the case is selected because it exemplifies the philosophical point, then it is not clear that the historical data hasn't been manipulated to fit the point. On the other hand, if one starts with a case study, it is not clear where to go from there—for it is unreasonable to generalize from one case or even two or three.[30]

In this way, Pitt dismisses the sophisticated and powerful case study methodology, which social scientists have successfully applied in a wide variety of settings. In answer to Pitt, there is no "dilemma": case studies are a tool designed not to "exemplify a point", but for hypothesis generation; and the methodology of case studies involves not "generalizing" from a few cases, but conducting within-case causal analysis in order to formulate a hypothesis that may subsequently be tested by means of quantitative methods.

Nevertheless, writers have continued to criticize case study methodology in history and philosophy of science for limited usefulness and unreliability. Jutta Schickore has advocated that we abandon entirely the project of confronting philosophical models of science with historical data, largely because of the problems of using historical case studies.[31] Adrian Currie has argued that case studies cannot serve as a source of inductive evidence, and that lifting the "curse of case studies" in philosophy of science involves restricting them to heuristic, rhetorical, and illustrative roles.[32] Katherina Kinzel has concluded that the question, "how can case studies from the history of science support claims in philosophy of science?", admits no strong answer: whereas historical reconstructions may serve various useful functions, the theory ladenness of case studies makes them unsuited to adjudicate between philosophical claims.[33] Most recently, despite noting that researchers in social and medical sciences use the case study method productively in hypothesis generation, Wolfgang Pietsch has underlined the

difficulties of generalizing from case studies, and advocated restricting them chiefly to the role of developing conceptual schemes.[34]

I find these judgements to be unduly despairing. Case study methodology has potentially no less productiveness and value in history and philosophy of science than it has shown over decades in social science. The necessary condition is that we acknowledge that case studies are a method primarily to generate, rather than to confirm, hypotheses, we refrain from asking case studies to carry out functions for which they are unsuited, and we avoid using the label "case study" for methodological devices that do not fall within this category.

5. Testing Scientific Realism

The issues that we have discussed up to now come together in the project of testing scientific realism, one of the most important loci of the discussion on how to use historical data to test philosophical theories and models of science. We take scientific realism to be the thesis that a scientific theory's observational success follows from its referential success, or its success in identifying real structures in the world. How can we bring historical evidence to bear in testing such a thesis?

Three preliminary considerations apply. First, both referential success and observational success are properties that admit degrees: theories can be more or less right about the structure of the world, and their predictions can be more or less accurate. Most writers have endorsed the common-sense view that, in general, later scientific theories have shown more observational success than earlier ones.[35]

Second, scientific realism posits a correlation between these two quantitative variables: theories with a high or higher degree of referential success achieve a high or higher degree of observational success. The discovery of a strong correlation between degrees of referential success and

degrees of observational success in theories in history would constitute a strong confirmation of scientific realism; the absence of such a correlation, by contrast, would disconfirm realism. Philosophers of science have discussed, partly by appeal to probabilistic arguments, how we could expect a correlation of this kind to manifest itself in the historical record.[36]

Third, since it is possible for untrue premises to entail a true conclusion, there are innumerably many conceivable theories that have low referential success but high observational success. Some scientific realists have suggested that it is unlikely—even a "miracle"—that a theory actually put forward in history could have had high observational success without some degree of referential success. No one, however, has considered this outcome impossible.

For these reasons, testing scientific realism by reference to history of science calls for a quantitative approach. Such an approach would pose the required forward causal inference question: what consequences follow from a theory's degree of referential success? To answer this question systematically, we need to formulate quantitative estimates of the degrees of referential and observational success of a large number of scientific theories in history, and gauge the strength of the correlation between referential success and observational success in this data set.

In summary, if we wish to test scientific realism against historical evidence, we should follow the approach by which Hull, Tessner, and Diamond tested Planck's principle. Admittedly, constructing quantitative estimates of the degrees of referential and observational success of past scientific theories is more difficult than harvesting biographical data. However, any historical test of scientific realism would require us, explicitly or implicitly, to form an opinion of the referential and observational success of past theories: all that a quantitative method demands is that we commit ourselves to specific estimates about a sufficiently large number of theories.

What about a qualitative approach, such as the case study method? This would be less appropriate. It would consist in choosing scientific theories

in history that have shown a high degree of observational success, reconstructing what led to their success, and checking whether referential success played a part. This approach would have two interrelated shortcomings. First, the case studies themselves would deliver no estimate of the strength of the correlation between referential and observational success: that would have to be reached separately. Second, the approach would yield only conclusions based on the individual theories examined: in order to guard against cherry-picking of cases, the conclusions would still have to be tested against a wider sample.

The empirical tests of scientific realism that philosophers of science have actually mounted have gone slightly differently. Larry Laudan presented his test as refutation: "the history of science […] decisively confutes several extant versions of […] scientific realism". He attempted to carry out this refutation on the strength of thirteen "theories which were both successful and (so far as we can judge) non-referential", such as the family of aether theories of the 1830s and 1840s, which Laudan presented as counterexamples to scientific realism.[37]

This approach is flawed, because of the mismatch between the quantitative phenomenon and the method of refutation. Laudan acknowledged that both referential and observational success came in degrees. He interpreted a theory's referential success in terms of "approximate fit" with the world and "approximate truth", which are self-evidently quantitative properties. Similarly, he wrote that calling a theory observationally successful meant that "it has functioned in a variety of explanatory contexts, has led to confirmed predictions and has been of broad explanatory scope", also accomplishments that come in degrees.[38] But a statistical hypothesis cannot be refuted in any straightforward way. Laudan, therefore, was able to proceed only by reinterpreting the two relevant terms as binary properties, rather than as showing degrees, and by reformulating scientific realism as the claim that possession of one was a necessary condition for possession of the other. He listed the thirteen theories as if they

straightforwardly had observational success but lacked referential success. This, however, goes against our understanding of these quantitative parameters.

The truth is that, even if these thirteen theories had had very low referential success and very high observational success, there could still be a strong correlation between referential and observational success in history. These theories could have been outliers, perhaps cherry-picked for that reason. Even if it could be carried to fruition, therefore, Laudan's approach would not yield an estimate of the strength of the correlation between referential and observational success. For that, we need a quantitative approach that tests the hypothesis of the correlation against as wide and representative a sample of past scientific theories as possible.

Many subsequent philosophers of science have rightly regarded Laudan's article as a landmark in the use of history of science as evidence in philosophy of science. However, they have differed in the lessons that they have drawn from it.

Some writers have followed Laudan in regarding attempted refutation or *modus tollens* as the correct way to test scientific realism.[39] A second group has interpreted Laudan's argument as inductive or, more specifically, as a "pessimistic meta-induction".[40] For example, Stathis Psillos has rejected Laudan's findings on the grounds that "the inductive basis is not big and representative enough to warrant the pessimistic conclusion".[41] It remains unclear, however, how either attempted straightforward refutation or enumerative induction could test a hypothesis that two quantitative parameters are correlated.

A third group has explored quantitative methods, albeit not yet a direct test of a correlation between referential and observational success of past theories. Moti Mizrahi has argued that the historical record contained fewer abandoned theories and theoretical posits than antirealists have assumed. Citing Price's claim that science has grown exponentially, Ludwig Fahrbach has argued that Laudan's examples were not representative of

scientific theories, because they belonged to the earliest ten percent of scientific work up to now.[42]

The largest group, including Timothy D. Lyons, Juha Saatsi, and Peter Vickers, has followed a fourth project. It has embarked on detailed scrutiny of each of Laudan's historical examples, as well as similar examples that have been proposed subsequently, to gauge to what extent they truly weighed against scientific realism. This work consists chiefly in examining to what extent the observational success of past theories depended on theoretical components that scientists later abandoned. This is a valuable critical audit of the historical data, which will yield more refined estimates of the degrees of referential and observational success of some individual past theories: it can be regarded as an essential step preliminary to a quantitative test of the thesis of scientific realism.

Many writers who have contributed to this fourth project have voiced two assumptions: first, that this work consists in the production of case studies, and second, that this audit of the data itself constitutes a historical test of scientific realism.[43]

Neither of these assumptions is tenable, however. First, genuine case study methodology would be useful if we wished to generate hypotheses about the factors responsible for the observed observational success of some theories—a reverse causal inference question. All contributors to this project, by contrast, already have a clear hypothesis ready for testing: that there is a correlation between referential and observational success in theories. To describe the fourth project as producing case studies is therefore misleading: an examination of an individual historical case is not necessarily case study methodology in the strict sense.

Second, testing similar general hypotheses requires a cross-case analysis on the broadest evidential sample possible, not a study of individual examples. Because the hypothesis at issue here posits a correlation, furthermore, a quantitative approach is all the more necessary. Tests of the thesis of scientific realism must thus still consist in checking for a correlation

between degrees of referential and observational success in as broad a population of historical scientific theories as possible. Whereas the fourth project consists in a valuable preliminary refinement of the empirical basis, it does not on its own constitute a test of the thesis of scientific realism: that work remains to be done.

This, I suggest, helps to explain why the debate on the historical plausibility of scientific realism, which has proceeded since the 1980s in terms of falsification, enumerative induction, and case study methodology, has not approached resolution. We need, instead, a test of the thesis of scientific realism that uses cross-case analysis to tackle forward causal inference or effects-of-causes questions: in short, quantitative methods.

This review of the scientific realism debate confirms what we found in earlier sections: the philosophy of science community has tended to undervalue quantitative methods and to omit to apply them even when a research question called for such an approach. Instead, philosophers of science minded to test their claims by appeal to history of science have tended to recognize only three methods for doing this: refutation, induction, and the case study method—and they have imperfectly understood the latter.

6. Conclusions

Because many philosophical claims about science—including the thesis of scientific realism—are quantitative, consisting of the claim that two variables are correlated, any effective test of them requires quantitative methods. Qualitative methods, like the case study method, are less suited to this task, for at least two reasons. First, in general, the case study method is less suited to hypothesis test than to hypothesis formulation. Second, more specifically, qualitative methods are unsuited to testing claims that one variable is correlated with another.

Unfortunately, philosophers of science have been less ready to use quantitative methods to test such claims than they should be. Many philosophers of science have thought, instead, that they could use a qualitative method, specifically the case study method, to this end. In a further twist, the methods that some philosophers of science have used to test quantitative claims, which they have called "case studies", have not fulfilled the specifications of a case study method.

We could sum up the situation by saying that philosophers of science have shown considerable methodological unclarity and even confusion. This is a puzzling and painful finding. We expect philosophers of science, among all researchers, to be aware of and reflective about methodological questions; furthermore, since many philosophers of science traditionally have had affinity with quantitative and experimental sciences, we might expect them properly to value quantitative methods. The contrary seems to have been the case.

In recent decades, the project of testing philosophical claims by appeal to history of science has become less fashionable. Part of the reason is, I think, that inappropriate choice of methods has doomed the project to failure. To revitalise the project, I suggest the following measures for philosophers of science. First, cultivate more willingness to learn from methodological debates in social sciences, and from the ability of these disciplines to combine qualitative and quantitative approaches. Second, for clarity, confine the term "case study" strictly to circumscribed exploratory studies that aim to yield insights for a broader category, as the term is used in methodology of social science. I do not regard this as excessive terminological purism: social scientists have done so much to refine case study methodology that I think we should defer to their definition. Third, refocus attention onto the specifically quantitative aspects of philosophical claims about science, including the thesis of scientific realism. Fourth, promote quantitative research in history of science. To adjudicate the debate about scientific realism, for example, estimates of the degrees of observational and referential success of past

theories would be very useful. Fifth, integrate qualitative and quantitative methods in history and philosophy of science: then, case studies can revert to their proper role of permitting within-case causal analysis and suggesting hypotheses, without shouldering the burden of testing general quantitative claims about science.[44]

Notes

[1] David L. Hull, "Studying the Study of Science Scientifically", *Perspectives on Science*, 6 (1998), 209–231.

[2] Seymour H. Mauskopf and Tad M. Schmaltz (eds.), *Integrating History and Philosophy of Science: Problems and Prospects* (Dordrecht: Springer, 2011).

[3] Alan Bryman, *Quantity and Quality in Social Research* (London: Unwin Hyman, 1988); Gary Goertz and James Mahoney, *A Tale of Two Cultures: Contrasting Qualitative and Quantitative Paradigms* (Princeton, N.J.: Princeton University Press, 2012).

[4] Andrew Gelman, "Causality and Statistical Learning", *American Journal of Sociology*, 117 (2011), 955–966, on 955–957.

[5] John Gerring, *Case Study Research: Principles and Practices* (Cambridge: Cambridge University Press, 2007), 20.

[6] Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China* (Cambridge: Cambridge University Press, 1979).

[7] Manfred Max Bergman, "The Straw Men of the Qualitative-quantitative Divide and Their Influence on Mixed Methods Research", in Manfred Max Bergman (ed.), *Advances in Mixed Methods Research: Theories and Applications* (London: Sage, 2008), 11–21.

[8] Dana Tulodziecki, "A Case Study in Explanatory Power: John Snow's Conclusions about the Pathology and Transmission of Cholera", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42 (2011), 306–316.

[9] John Forrester, "If *p*, then What? Thinking in Cases", *History of the Human Sciences*, 9 (1996), no. 3, 1–25; Mary S. Morgan, "Case Studies: One Observation or Many? Justification or Discovery?", *Philosophy of Science*, 79 (2012), 667–677.

[10] Derek J. de Solla Price, *Little Science, Big Science* (New York: Columbia University Press, 1963), 1–32.

[11] Arnold Thackray, "Measurement in the Historiography of Science", in Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray, and

Harriet Zuckerman (eds.), *Toward a Metric of Science: The Advent of Science Indicators* (New York: Wiley, 1978), 11–30; Roger Hahn, *A Bibliography of Quantitative Studies on Science and Its History* (Berkeley: University of California, 1980).

[12] Manfred D. Laubichler, Jane Maienschein, and Jürgen Renn, "Computational Perspectives in the History of Science: To the Memory of Peter Damerow", *Isis*, 104 (2013), 119–130.

[13] David L Hull, Peter D. Tessner, and Arthur Diamond, "Planck's Principle: Do Younger Scientists Accept New Scientific Ideas with Greater Alacrity than Older Scientists?", *Science*, 202 (1978), 717–723.

[14] K. Brad Wray, "Is Science Really a Young Man's Game?", *Social Studies of Science*, 33 (2003), 137–149; Frank J. Sulloway, "Openness to Scientific Innovation", in Dean K. Simonton (ed.), *The Wiley Handbook of Genius* (Chichester: Wiley, 2014), 546–563; Pierre Azoulay, Christian Fons-Rosen, and Joshua S. Graff Zivin, "Does Science Advance One Funeral at a Time?", National Bureau of Economic Research Working Paper 21788 (Cambridge, Mass.: National Bureau of Economic Research, 2015). http://www.nber.org/papers/w21788.

[15] Susann Wagenknecht, Nancy J. Nersessian, and Hanne Andersen (eds.), *Empirical Philosophy of Science: Introducing Qualitative Methods into Philosophy of Science* (Cham: Springer, 2015).

[16] James B. Conant, *On Understanding Science: An Historical Approach* (New Haven, Conn.: Yale University Press, 1947), 1; James B. Conant (ed.), *Harvard Case Studies in Experimental Science* (Cambridge, Mass.: Harvard University Press, 1957).

[17] John Forrester, "On Kuhn's Case: Psychoanalysis and the Paradigm", *Critical Inquiry*, 33 (2007), 782–819.

[18] Imre Lakatos, "History of Science and Its Rational Reconstructions", in Roger C. Buck and Robert S. Cohen (eds.), *PSA 1970: Proceedings of the 1970 Biennial Meeting of the Philosophy of Science Association* (Dordrecht: D. Reidel, 1971), 91–136.

[19] Colin Howson, "Editorial Preface", in Colin Howson (ed.), *Method and Appraisal in the Physical Sciences: The Critical Background to Modern Science, 1800–1905* (Cambridge: Cambridge University Press, 1976), vii.
[20] Peter Clark, "Atomism versus Thermodynamics", in Howson, *Method and Appraisal in the Physical Sciences*, 41–105, on 44.
[21] Larry Laudan, Arthur Donovan, Rachel Laudan, Peter Barker, Harold Brown, Jarrett Leplin, Paul Thagard, and Steve Wykstra, "Scientific Change: Philosophical Models and Historical Research", *Synthese*, 69 (1986), 141–223.
[22] Arthur Donovan, Larry Laudan, and Rachel Laudan (eds.), *Scrutinizing Science: Empirical Studies of Scientific Change*, 2nd ed. (Baltimore, Md.: Johns Hopkins University Press, 1992).
[23] Rachel Laudan, Larry Laudan, and Arthur Donovan, "Testing Theories of Scientific Change", in Donovan, Laudan, and Laudan, *Scrutinizing Science*, 3–44, on 16, 31.
[24] David L. Hull, "Testing Philosophical Claims about Science", in David L. Hull, Micky Forbes, and Kathleen Okruhlik (eds.), *PSA 1992: Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association* (East Lansing, Mich.: Philosophy of Science Association, 1993), vol. 2, 468–475, on 470.
[25] Laudan, Laudan, and Donovan, "Testing Theories of Scientific Change", 11.
[26] Laudan, Laudan, and Donovan, "Testing Theories of Scientific Change", 5, 6.
[27] C. E. Perrin, "The Chemical Revolution: Shifts in Guiding Assumptions", in Donovan, Laudan, and Laudan, *Scrutinizing Science*, 105–124.
[28] Laudan et al., "Scientific Change", 159.
[29] Larry Laudan and Rachel Laudan, "The Re-emergence of Hyphenated History-and-philosophy-of-science and the Testing of Theories of Scientific Change", *Studies in History and Philosophy of Science*, 59 (2016), 74–77, on 74.
[30] Joseph C. Pitt, "The Dilemma of Case Studies: Towards a Heraclitian Philosophy of Science", *Perspectives on Science*, 9 (2001), 373–382, on 373.
[31] Jutta Schickore, "More Thoughts on HPS: Another 20 Years Later", *Perspectives on Science*, 19 (2011), 453–481.

[32] Adrian Currie, "Philosophy of Science and the Curse of the Case Study", in Christopher Daly (ed.), *The Palgrave Handbook of Philosophical Methods* (London: Palgrave Macmillan, 2015), 553–572, on 559–560.

[33] Katherina Kinzel, "Narrative and Evidence: How Can Case Studies from the History of Science Support Claims in the Philosophy of Science?", *Studies in History and Philosophy of Science*, 49 (2015), 48–57.

[34] Wolfgang Pietsch, "Two Modes of Reasoning with Case Studies", in Tillman Sauer and Raphael Scholl (eds.), *The Philosophy of Historical Case Studies* (Dordrecht: Springer, 2016), 49–67.

[35] Ludwig Fahrbach, "Theory Change and Degrees of Success", *Philosophy of Science*, 78 (2011), 1283–1292.

[36] Marc Lange, "Baseball, Pessimistic Inductions and the Turnover Fallacy", *Analysis*, 62 (2002), 281–285; P. D. Magnus and Craig Callender, "Realist Ennui and the Base Rate Fallacy", *Philosophy of Science*, 71 (2004), 320–338.

[37] Larry Laudan, "A Confutation of Convergent Realism", *Philosophy of Science*, 48 (1981), 19–48, on 19, 33.

[38] Laudan, "A Confutation of Convergent Realism", 23.

[39] Timothy D. Lyons, "Scientific Realism", in Paul Humphreys (ed.), *The Oxford Handbook of Philosophy of Science* (New York: Oxford University Press, 2016), 564–584, on 564–571.

[40] K. Brad Wray, "Pessimistic Inductions: Four Varieties", *International Studies in the Philosophy of Science*, 29 (2015), 61–73; Juha Saatsi, "Historical Inductions, Old and New", *Synthese*, in press.

[41] Stathis Psillos, *Scientific Realism: How Science Tracks Truth* (London: Routledge, 1999), 105.

[42] Moti Mizrahi, "The History of Science as a Graveyard of Theories: A Philosophers' Myth?", *International Studies in the Philosophy of Science*, 30 (2016), 263–278; Ludwig Fahrbach, "Scientific Revolutions and the Explosion of Scientific Evidence", *Synthese*, in press.

[43] Juha Saatsi, "Reconsidering the Fresnel–Maxwell Theory Shift: How the Realist Can Have Her Cake and EAT It Too", *Studies in History and Philosophy*

of Science, 36 (2005), 509–538; Timothy D. Lyons, "Scientific Realism and the Stratagema de Divide et Impera", *British Journal for the Philosophy of Science*, 57 (2006), 537–560; Juha Saatsi and Peter Vickers, "Miraculous Success? Inconsistency and Untruth in Kirchhoff's Diffraction Theory", *British Journal for the Philosophy of Science*, 62 (2011), 29–46.

[44] I thank Jouni-Matti Kuukkanen, Director, Centre for Philosophical Studies of History, University of Oulu, for the invitation to the workshop, "Testing Philosophical Theories Against the History of Science", September 2015. I thank the participants, two unnamed referees of this journal, and Attilia Ruzzene for valuable comments on successive drafts.