Biol Philos
DOI 10.1007/s10539-013-9373-3

Evidence in biology and the conditions of success

Jacob Stegenga

Received: 20 June 2012 / Accepted: 27 March 2013

Abstract I describe two traditions of philosophical accounts of evidence: one characterizes the notion in terms of *signs of success*, the other characterizes the notion in terms of *conditions of success*. The best examples of the former rely on the probability calculus, and have the virtues of generality and theoretical simplicity. The best examples of the latter describe the features of evidence which scientists appeal to in practice, which include general features of methods, such as quality and relevance, and general features of evidence, such as patterns in data, concordance with other evidence, and believability of the evidence. Two infamous episodes from biomedical research help to illustrate these features. Philosophical characterization of these latter features—conditions of success—has the virtue of potential relevance to, and descriptive accuracy of, practices of experimental scientists.

Keywords Evidence · Experiment · Confirmation · Error · Robustness · Avery · Water memory · Benveniste · Methodology

Introduction

Contemporary accounts of evidence, when explicated in terms of *signs of success*, specify what is achieved once one has reliable evidence. Such accounts of evidence often rely on the probability calculus ("Signs of success"), and are largely unhelpful for those scientists involved in generating and assessing evidence, since a primary concern of experimentalists is to determine whether or not some evidence is indeed reliable, rather than to determine the precise nature of what is gained, epistemically,

J. Stegenga (✉)
Institute for the History and Philosophy of Science and Technology, Victoria College, University of Toronto, Room 316, 91 Charles Street West, Toronto, ON M5S 1K7, Canada
e-mail: jacob.stegenga@utoronto.ca
URL: http://individual.utoronto.ca/jstegenga

Published online: 04 June 2013

by reliable evidence. The multiple features of methods and of evidence itself which are important for assessing evidence are *conditions of success*. Scientists assess features of methods such as quality, relevance, and transparency ("Conditions of success: methodological features"), and features of evidence such as patterns in data, concordance with other evidence, and believability of the evidence ("Conditions of success: evidential features"). When evidence is judged favorably on these desiderata it is considered truth-conducive.

I describe two cases to illustrate biologists' appeals to the conditions of success: the elucidation of the material basis of heredity in the 1940s by Avery et al. ("Illustration: material basis of heredity"), and the purported demonstration of 'water memory' by Benveniste and his colleagues in the 1980s ("Illustration: water memory"). The two cases present a nice contrast, since the first is an episode in which the evidence is now widely regarded as compelling or even conclusive, and the phenomenon for which the relevant paper was said to provide evidence is now considered generally true, whereas the second is an episode in which the evidence is now widely regarded as unreliable and the respective phenomenon very likely false. I show that the biologists in these two cases appealed to the conditions of success identified in "Conditions of success: methodological features" and "Conditions of success: evidential features" both when criticizing evidence and when praising evidence. A consideration of signs of success is totally absent in these cases: the relevant biologists do not appeal to signs of success, despite their frequent appeals to the conditions of success.

My primary aim is to compare the signs of success tradition with the conditions of success tradition, and to provide some detail to the conditions of success tradition. I focus on the conditions of success not simply because it provides a more descriptively accurate framework for characterizing how biologists assess evidence, but because it is in a sense prior to and richer than the signs of success tradition. The signs of success tradition is necessarily post hoc, in that one already must possess and have evaluated the evidence with the conditions of success in order to do business in the signs of success tradition.

Consider the following analogy. What makes for a good wine? One answer would be that a wine is good if the wine is awarded more than 90 points by a well-known wine critic. Another answer would describe methods of quality wine production by careful vineyards. A third answer would list features of wine itself, such as its bouquet, color, and taste, that one ought to consider when assessing wine. The first answer is relatively uninformative, since it simply restates what we want to know, albeit in more precise and quantitative terms. The second answer lists concrete aspects of the method of production of a particular wine that a critic could appeal to. The third answer lists concrete aspects of the wine itself that a critic could appeal to. Of course, once a critic has appealed to the latter considerations she might provide a numerical score to summarize her oenological investigation, and such a score might be useful to consumers of wine.

Characterizing evidence in terms of conditions of success is a more accurate description of how biologists assess evidence (compared with the signs of success), but the conditions of success also ought to be construed as normative. Strictly speaking the conditions of success and signs of success are not at odds: the two

traditions are not alternative normative theories of scientific inference, as, say frequentist and Bayesian conceptions of inference are. The conditions of success tradition describes the important features of evidence which ought to be considered when assessing evidence to determine whether or not it is reliable, and the signs of success tradition describes what is achieved once one has reliable evidence. The formal measures of confirmation employed in the signs of success tradition can, in principle, accommodate any of the methodological considerations employed in the conditions of success tradition. The key difference from an experimentalist's point of view is that an experimentalist has some control over at least some of the conditions of success (whereas the signs of success are simply determined by features of the methods employed and evidence generated), and an experimentalist has more immediate epistemic access to the conditions of success (since, as I argue below, the conditions of success are the daily bread and butter of experimentalists, and moreover, assessing the conditions of success is a precondition to determining the formal measures employed in the signs of success tradition).

Philosophical assessments and implications of experimentation have generated a rich literature, and several recent contributions to this literature have described experimentation in biology.[1] This paper is meant to be a contribution to this literature. In short, my focus is on the plurality and complexity of the conditions of success: assessing the conditions is complicated and there is no simple or universally agreed-upon algorithm for assessing the particular criteria. Despite their complexity, the conditions of success form the basis of the evaluation of evidence for experimentalists, in contrast with the formal accounts of evidence in the signs of success tradition. The two cases studies demonstrate the centrality of the conditions of success.

Signs of success

Many philosophical accounts of evidence describe what reliable evidence *achieves*. For instance, all compelling probabilistic accounts of evidence hold that e is evidence for some hypothesis H, given background assumptions b, if and only if $p(H|e \,\&\, b) > p(H|b)$; that is, the probability of h given e and b must be greater than the probability of H prior to having e, if and only if e is to count as evidence for H. As an example of such accounts, the difference measure of confirmation holds that the more confirming e is, the greater is the inequality between $p(H|e \,\&\, b)$ and $p(H|b)$.[2] It is standard to distinguish between the final amount of confirmation that a piece of evidence provides to a hypothesis from the change in confirmation that a piece of evidence provides to a hypothesis. The former is simply represented by the posterior probability of the hypothesis (the probability of the hypothesis after learning new evidence): $p(H|e)$. Since this is a conditional probability it can be re-

[1] Early contributions include Hacking (1983) and Franklin (1986), and recent discussions of experiment in biology include Bechtel and Richardson (1993), Burian (1993), Allchin (1996), Rheinberger (1997), Rasmussen (2001), Darden and Craver (2002), Griffiths (2002), Weber (2005), Elliott (2007), Waters (2007), and Weber (2012).

[2] Hereafter I drop reference to background assumptions (b) for notational simplicity.

written using Bayes' Theorem: p(e|H)p(H)/p(e). The latter can be represented in a number of ways. Recent literature has included defenses of the following confirmation measures.[3]

Difference measure	$p(H\|e) - p(H)$
Ratio measure	$p(H\|e)/p(H)$
Likelihood ratio measure	$p(e\|H)/p(e\|\sim H)$
Log ratio measure	$\log[p(H\|e)/p(H)]$
Log likelihood measure	$\log[p(e\|H)/p(e\|\sim H)]$

In addition to satisfying conditions like the ones above, some philosophers require that the probability of the hypothesis be above a certain threshold after receiving e, if e is to count as evidence for the hypothesis. On such views e is evidence for H only if $p(H|e) > x$, where x is some minimum threshold. Achinstein (2001), for example, requires x to be 0.5 for e to count as *veridical* evidence for H (evidence that provides good reason to believe H), and Roush (2005) requires x to be much greater than 0.5 to consider e as *good* or strong evidence for H.

Another sign of good evidence is what Roush calls 'discrimination'—evidence should discriminate between a hypothesis and the negation of that hypothesis—and to measure this Roush argues that the likelihood ratio is appropriate: $p(e|H)/p(e|\sim H)$. If the likelihood ratio is greater than 1, then e discriminates between the hypothesis and its negation, and thus is evidence for the hypothesis. In other words, e should be more likely conditional on the hypothesis being true than conditional on the hypothesis being false. One of Roush's examples is the 'check engine' light in an automobile: if it is much more probable that the check engine light is on when there is engine trouble, as compared to the check engine light being on when there is no engine trouble, then the check engine light is discriminating evidence for the hypothesis that there is engine trouble. Roush also claims that good evidence should have a high probability if we are to think that the evidence is credible; that is, a high p(e) indicates that the evidence is believable.[4]

Each of these accounts provides a competing characterization of what evidence achieves. Each is compelling in different ways—I will not review the virtues and vices of these accounts of evidence. They all share one feature which renders them relatively useless to the experimenter: they indicate what is achieved once one has reliable evidence—they are signs of success—but they do not indicate how to generate or identify evidence which can then be granted these signs. They provide

[3] On these measures, see, for example, Fitelson (1999).

[4] This desideratum—requiring a high p(e) for credible evidence—departs from standard Bayesian thinking about evidence. In a standard Bayesian framework, a high p(e) indicates both that the evidence is credible but also that it provides little support to any particular hypothesis (this can be easily seen via Bayes' Theorem). A low p(e) is usually thought to represent surprising evidence, from, say, a risky prediction, and a smaller p(e) is associated with a greater increase in p(H|e) than a larger p(e), by Bayes' Theorem, and this reflects general intuition about the confirmatory power of surprising evidence. Using high p(e) to represent credible evidence sacrifices the ability to represent surprising and highly confirming evidence with p(e). A low p(e) also represents evidence generated by a high quality method (discussed in Sect. "Conditions of success: methodological features").

post hoc *characterizations* of good evidence rather than guidance on the *production* or *identification* of reliable evidence.[5] These accounts of evidence are like awards which are used to distinguish reliable evidence from unreliable evidence: they characterize the nature of the award, but scientists want to know which evidence to give the award to.

Some may think that there is nothing very general to be said about substantive evidential standards. Some may think that identifying reliable evidence is a matter best left to scientists, whereas characterizing reliable evidence is more properly a philosophical concern. The conditions of success tradition, however, has aimed at describing some of the most important substantive evidential standards. In what follows I describe general features of methods and of evidence itself which can be used to generate and identify reliable evidence ("Conditions of success: methodological features" and "Conditions of success: evidential features"), and which, as the case studies demonstrate ("Illustration: material basis of heredity" and "Illustration: water memory"), are actively employed by experimental biologists to assess evidence.

Conditions of success: methodological features

A method of generating evidence can be assessed in the abstract, independently of any actual evidence generated by the method. That is, prior to the consideration of any evidence from a method, the method itself can be (and almost always is) assessed. Three general features or desiderata of methods are freedom from systematic errors, relevance to our hypothesis of interest, and how ascertainable either of these are. I will call these quality, relevance, and transparency.

Quality

A method is high quality if and only if possible systematic errors are controlled for. If the method controls for systematic errors, then evidence generated by the method is a faithful indicator of the subject of study. The term 'internal validity' has often been used for the notion of quality, and is meant to indicate how well a study is designed and performed to avoid systematic error and bias, which can result from

[5] According to accounts of evidence associated with Williamson (2000), Neta (2008), and others, evidence is factive, and so it makes little sense to talk about the 'veracity of evidence', or 'reliable evidence'. There is, on the factive account, simply evidence, the veracity of which is taken for granted. Any account of evidence must be able to accommodate the difference between (i) evidence generated by a method with systematic errors and (ii) evidence generated by a method which controls for known errors. On my account (ii) is reliable evidence and (i) is unreliable or weak evidence (but the amount of systematic error present is presumably a degree notion). On the factive account, a proposition expressing an evidential report can accommodate the difference between (i) and (ii) by including the relevant information regarding the methodological differences between (i) and (ii) in the proposition expressing the evidential report itself. Given the methodological complexity of contemporary experiments, many evidential reports in such a factive account would be complex and cumbersome. The advantage of the conditions of success account is that an evidential report can be stated rather simply while the assessment of such reports can be as complex as need be. This is, moreover, precisely how biologists report and assess evidence.

flaws in study design, conduct, analysis, interpretation, and reporting. Quality has been a staple subject for statisticians, philosophers of science, and scientists concerned with methodology; volumes have been written on the subject. One recent account of evidential quality is what Mayo calls the 'severity principle': data x provide good evidence for a hypothesis H to the extent that H severely passes a stringent test with x (see Mayo 1996, Mayo and Spanos 2006). On this account a method is high quality if it comprises a stringent test, and a stringent test is one which severely probes for possible errors. Achinstein proposes another way to characterize quality: his notion of 'evidential flaws' refers to flaws in the evidence-generating method. On this account, quality is an absence of such flaws (2001).

The notion of quality of a method is itself comprised of numerous features. The presence of standard elements of experimental design determines quality of evidence—for instance, in a medical study random allocation of subjects, appropriate blinding, and proper use of analytical tools are factors which determine the quality of evidence. Quality can be characterized in terms of the plausibility of the background assumptions required to consider evidence generated by the method as a truth-conducive indicator of the particular subject under investigation. The relation between the numerous methodological features that comprise quality and the formal measures employed in the signs of success tradition can be characterized as follows. A well-controlled method minimizes the probability that the evidence generated by the method could have occurred for reasons other than those supposed by the hypothesis. Since the quality of a method is constituted by controlling for systematic errors, quality is meant to ensure that evidence generated by the method would be otherwise unlikely were it not for the truth of the particular hypothesis of interest. Putting this in terms of probabilities, the higher the quality of a method which generates evidence e, the lower is p(e), and so the greater is the difference between p(e|H) and p(e).[6] In the signs of success tradition, confirmation can be characterized as proceeding via the 'Bayesian multiplier'—the ratio of p(e|H) to p(e)—and so a higher p(e|H) and a lower p(e) entails greater confirmation.

Relevance

Tossing a coin several times gives some evidence regarding the fairness of the coin, since there is a clear relationship between the results of a series of coin tosses and the probability that the coin is fair. But tossing a coin several times gives no evidence regarding tomorrow's weather, since there is no relationship between the results of a coin toss and tomorrow's weather. In other words, evidence from coin tossing has no relevance to tomorrow's weather. And this, of course, is true regardless of any actual evidence generated from coin tossing. Relevance to a hypothesis is obviously a crucial feature of methods.

[6] If c_i represents the possible confounding errors of the method used to generate e, and if we assume for simplicity that H and c_i represent a total partition of the possible causes of e, then, by the principle of total probability: $p(e) = p(H)p(e|H) + p(c_1)p(e|c_1) + p(c_2)p(e|c_2) + \ldots p(c_n)p(e|c_n)$. Since the quality of a method amounts to decreasing the prior probability that any of c_i are true, quality directly influences p(e). The higher the quality of a method, the lower the p(e). For a valuable discussion of Bayesian approaches to evidence, see Strevens (2009).

The degree of relevance of a method to a hypothesis depends on both the method and the hypothesis (obviously), as the coin tossing example shows. Suppose our hypothesis is more general than simply the fairness of the single coin which we toss, but is rather about the fairness of all the coins in my pocket. The method—tossing the single coin—then would be less relevant to the hypothesis. Considering a more expansive hypothesis in this case rendered the same method less relevant. Relevance depends on the background assumptions which we are willing to entertain. Some methods will be similarly relevant to hypotheses of a range of generality. For instance, dropping a coin once will give some evidence about the tendency of this coin to fall, but it will also give some evidence about the general tendency of coins to fall in such situations, because it is reasonable to suppose that when it comes to falling, there is no relevant difference between the coin which we dropped and most other coins.

Like quality, relevance can be characterized by the signs of success tradition. I noted above that confirmation can be construed as proceeding via the Bayesian multiplier: p(e|H)/p(e). The greater the likelihood, p(e|H), the more confirmation H receives. To the extent that a method is relevant to H, H renders e more probable, and thus the likelihood is greater. To see this, it is helpful to consider a method which is completely irrelevant to a hypothesis. In this case, H does not change the probability of e at all: $p(e|H) = p(e)$. This is an application of a simple mathematical truth: the independence of two events (or variables) A and B can be represented probabilistically as: $p(A|B) = p(A)$. So if the hypothesis is independent of the evidence—if the method that produced e is not relevant to H—then H does not change the probability of e. Alternatively, if a method that generates e *is* relevant to H, then H can partially explain why e occurs (perhaps because H is a claim about a mechanism which could bring about e in the circumstances in which the method is employed).

Both quality and relevance are features that we should want our methods to have: methods with these features are more likely to generate evidence which is truth-conducive. These are standard desiderata of methods, though they are not always weighted equally. For example, in evidence-based medicine, randomized controlled trials (RCTs) are often considered to be high quality because they are said to minimize selection bias, and this is often said to be important even if a particular RCT is less relevant to a general hypothesis of interest than is a study design which has more potential for systematic error (such as larger observational studies). In other words, in evidence-based medicine, for better or worse, quality has tended to be emphasized over relevance.[7]

Transparency

Another general feature of methods, independent of the evidence produced by them, is transparency. Some methods are easier to know how they produce their evidence, which helps with knowing if a method has systematic error, and if evidence produced from a method would be relevant to a hypothesis. In order to know if a

[7] Many now argue that this is for the worse; see for example Worrall (2002) and Cartwright (2007).

method is high quality or is relevant to a particular hypothesis, one must know the details of the method's operation. A method is transparent if we can understand how it works—that is, a method is transparent if we can understand the mechanism of the method or, to use the catchy phrase from Cartwright (1999), a method is transparent if we can understand the 'nomological machine' underlying the method. We want to be able to make transparent judgments regarding quality and relevance—judgments which can be communicated and shared such that some agreement regarding quality and relevance might be achieved.

Unlike quality and relevance, however, transparency is merely a relational fact about scientists' understanding of an experimental system, rather than an intrinsic feature of the experimental system in question. Moreover, a transparent method is not necessarily truth-conducive (that is, a transparent method is not necessarily a method with high quality and relevance), and vice versa, a truth-conducive method is not necessarily transparent.[8] In many cases, especially those in which new methods are introduced into scientific practice, the method is not transparent.[9] Because transparency is not an intrinsic feature of experimental methods, and because it is not necessarily truth-conducive, I do not explore it further and do not refer to it in my illustrative case studies below. However, it is an important feature to note, because if scientists cannot directly assess the quality and relevance of a method—if the method is not transparent—then, to help determine if the evidence produced by the method is truth-conducive (or rather is inductively risky), scientists need to appeal to something other than the methodological features (that is, features other than quality and relevance). I turn now to the other class of epistemic features that are accessible by scientists and that are especially important when a method lacks transparency, namely, features of evidence itself.

Conditions of success: evidential features

Using examples from microscopy and neuroimaging studies, Bechtel (2000) argues that when scientists assess evidence produced by novel methods (methods with low transparency regarding quality and relevance), rather than assess the quality or relevance of the method that generated the evidence, scientists tend to assess multiple features of the evidence produced by the method. In other words, rather than simply assessing features of the method, scientists also assess features of the evidence itself. Bechtel's examples are of visual evidence, but his point generalizes.

Evidence is considered compelling (or not) based on a variety of features of the evidence produced by the method, independent of prior considerations of the

[8] I am grateful to an anonymous reviewer for noting that transparency might trade off against properties of an experimental system (such as quality and relevance) and properties of the target system (its complexity, say).

[9] Collins famously argues that in such scenarios, assessing evidence involves an 'experimenters regress': good evidence is generated from properly functioning techniques, but properly functioning techniques are just those that give good evidence (1985). Even if we put aside this rarified worry, we often cannot make judgments regarding the quality or relevance of a method simply because we do not know enough about the inner workings of the method to make such judgments.

features of the method (quality or relevance). Some of these features which are often appealed to by scientists include: patterns within the evidence, concordance with evidence from other methods, and believability. I discuss each in turn.

Patterns

If data have a determinate pattern or structure, then that is suggestive that the evidence is tracking something real. For example, Bechtel (2000) discusses the strategies that neuroscientists use to assess images of the brain generated by positron emission tomography (PET), one of which was to appeal to determinate structures in the images generated by PET. The PET images were not collections of randomly distributed colors, but rather the colors (which are transformations of numerical data) were arranged in a salient structure. This structure, independent of any interpretation of the structure's relation to brain activity, was taken to indicate that the images were not merely artifacts but rather represented something of real significance. Assessing the relevance of PET was not straightforward since the method was not transparent, but the sheer existence of structure in PET data suggested that the images were not artifacts.

Similarly, the epidemiologist Sir Bradford Hill provided a list of nine criteria with which he judged causal hypotheses in medicine, and one of these was the presence of a 'dose–response relationship' between the purported cause and the purported effect, which is another example of the presence of a suggestive pattern in data. The lung cancer rate was higher amongst those who smoked more cigarettes, and Hill considered this to be better evidence than a simple correlation between smoking and lung cancer. A simple correlation could be due to a common cause of smoking and lung cancer, but the presence of a dose response between smoking and lung cancer—a highly structured pattern of correlations—is more likely due to a true causal relationship.

Bogen and Woodward (1988) have emphasized the importance of patterns in data: for them, patterns are precisely what scientists look for when examining data.[10] In response, McAllister (1997) agrees that patterns are important but argues that for any set of data an infinite number of patterns can be discerned, expressible as the sum of a relatively simple function and an error term, such as $F(x) = ax + R(x)$. For instance, any set of data could be described by the following patterns:

Pattern A + noise at *m* percent
Pattern B + noise at *n* percent
Pattern C + noise at 0 percent

Pattern C would be the (perhaps complex) pattern which exactly fits the data points. McAllister is raising the standard curve-fitting problem, which is a version of underdetermination. McAllister suggests that the choice of which pattern is the

[10] See also Woodward (1989): "The problem of detecting a phenomenon is the problem of […] identifying a relatively stable and invariant pattern of some simplicity and generality with recurrent features – a pattern which is not just an artefact of the particular detection techniques we employ or the local environment in which we operate."

salient one for a given experiment is itself a complex judgment on the part of the investigators.

Numerous algorithms have been proposed to aid in choosing between pattern descriptions (or models of data), such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[11] These algorithms attempt to balance two standard desiderata of data models commonly recognized to trade-off against one another: simplicity and accuracy. The algorithms reward patterns which have fewer terms (i.e. patterns which are simpler) and less departure from observed data (i.e. patterns which are accurate). However, there is no meta-methodological algorithm for choosing between the model selection algorithms; for instance, there is no principled way to choose AIC over BIC. Trouble arises when AIC and BIC (or any other model selection criterion) select different models as achieving the optimal balance of simplicity and accuracy. In my toy example above, AIC might choose "Pattern A + noise at *m* percent" as the best model of the data, while BIC might choose "Pattern B + noise at *n* percent" as the best model of the data. Without a methodological meta-standard, it is unclear what the superior model is.

In short, a distinctive pattern or a suggestive structure in data is an important but complex consideration when assessing evidence.

Concordance

If a particular piece of evidence displays notable patterns, this might be due a feature of the experimental intervention or an artifact of the observational apparatus, rather than a sign of a real feature of the object under investigation. Thus, a common practice is to compare the evidence from one method to evidence from other methods. Concordant evidence from multiple methods is taken to be epistemically valuable (the term 'robustness' has been used to describe situations in which evidence from multiple methods is concordant). For example, when Hacking (1983) asked 'do we see through a microscope?'—asking, in other words, if we 'observe' real unobservable entities with the aid of instruments—to which van Fraassen and other antirealists answer 'no'—Hacking answered 'yes'. One of Hacking's arguments was that the fact that we can observe the same microscopic entities with multiple kinds of microscopes (optical, ultraviolet, electron…) gives us good reason to think that such entities are real. Elsewhere I have more fully characterized this feature of assessing evidence, and have argued that the conditions under which such arguments are successful are more demanding than many have supposed (see, e.g., Stegenga (2009)).[12] However, it is undeniably a common practice for scientists (and philosophers of science) to appeal to the epistemic value of concordant evidence.

[11] For discussion of such algorithms, see Sober (2007).

[12] See also Weber (2005), who calls such appeals to concordant evidence 'arguments from independent determinations'.

Believability

Evidence can only provide some confirmation or disconfirmation to a hypothesis to the extent that that evidence is believable. Evidence can be believable or not for several reasons. Evidence is theoretically believable if and only if *that* the evidence occurred can be explained by, or at least be consistent with, broadly accepted theories. Evidence is mechanistically believable if and only if *how* the evidence occurred can be explained by, or at least be consistent with, plausible or broadly accepted mechanisms of both the target system and the method.[13] A metaphysician's example illustrates the distinction. If an astronomer reported observing a gold sphere one mile in diameter, this would hardly be mechanistically believable, since we have no plausible explanation for *how* such a sphere could have originated. Nevertheless, a one-mile radius gold sphere is at least theoretically possible, since it would not contradict broadly accepted physical theories or laws. On the other hand, if an astronomer reported observing a sphere of uranium one mile in radius, this would contradict broadly accepted and fundamental physical theories, since a one-mile radius sphere of uranium would be far larger than the critical mass of uranium (which is reached by a sphere of uranium of roughly 15 cm).[14] If evidence is mechanistically believable, then presumably it is also theoretically believable, since assessments of possibility are made conditional on background theories; however, the reverse is not true (just because evidence is theoretically believable, it does not follow that the evidence is mechanistically believable).[15] Mechanistic believability, is, then, in a sense a stronger notion than theoretical believability.

Hill's nine criteria for judging causal hypotheses in epidemiology, discussed above, included 'plausibility'—that is, if epidemiological data were plausible on independent theoretical grounds, this was further reason to consider the data as truth-conducive. To continue the smoking example, if we know that cigarette smoke contains certain toxins, and similar toxins are otherwise known to cause adverse health effects in laboratory animals, then the finding that smoking is correlated with adverse health effects is more indicative of a true causal relationship than if there were no independent grounds for thinking that the toxins in cigarette smoke could possibly cause adverse health effects.

Summary

I have discussed the conditions of success which scientists appeal to when assessing evidence. When general features of methods are assessed, the question asked is "Would any evidence from this method be truth-conducive?" When general features of evidence are assessed, the question asked is "Is the evidence which was

[13] Such mechanisms include, but are not limited to, mechanisms of the method (since mechanisms of the method could be assessed under the rubric of quality). On the role of mechanisms in discovery, see Darden and Craver (2002), Bechtel (2006), and Tabery (2009).

[14] The criterion of believability is similar to Quine's (1951) claim that one can be justified in rejecting a certain observation if that observation strongly conflicts with one's background theories, while in the absence of such theories the same observation might be more plausible.

[15] I am grateful to an anonymous reviewer for suggesting this point.

actually produced truth-conducive?" The table below is a list of the evidential features discussed above. I give each an abbreviation for quick reference in the illustrations that follow.

General features of methods

(Q) Quality
(R) Relevance
(T) Transparency

General features of evidence

(P) Patterns
(C) Concordance
(B) Believability

This account of evidence—in terms of conditions of success, rather than signs of success—has the advantage of reflecting more closely the assessment of evidence by scientists.[16] The following two cases illustrate the ubiquitous practice of assessing evidence in terms of conditions of success rather than signs of success. When generating, assessing, and criticizing evidence, scientists do not ask "Is $p(H|e) > p(H)$?" or "Is $p(H|e) > 0.5$?", but rather "What is the (R) and (Q) of this method, and does the evidence have features (P), (C), and (B)?" These features of evidence are normative—as I have tried to briefly indicate in the last two sections, philosophers and scientists argue for (and occasionally challenge) the importance of each of these features as conditions for truth-conduciveness.

Illustration: material basis of heredity

Determining the material basis of heredity occupied the interest of some scientists in the 1930s through 1950s. I will use the assessment of evidence presented in the 1944 paper by Avery, Macleod, and McCarty (hereafter *AMM1944*) as an illustration of the appeal to what I have identified as conditions of success. This is a good illustrative case, since some biologists have retrospectively called the results in *AMM1944* "the pivotal discovery of 20th-century biology" (Lederberg 1994), and yet both the methodology and evidence presented in this paper were widely and critically assessed, and as I show, such assessments were based on the features of methods and evidence identified above "Conditions of success: evidential features".

First, some history.[17] In 1928 Griffith published his work on the 'transformation' of pneumococcal types. He had injected heat-killed, virulent, "smooth" pneumococci and live, non-virulent, "rough" pneumococci into mice. The mice died, and from their blood Griffith isolated live, virulent, "smooth" pneumococci. The live bacteria had changed virulence and morphology (from non-virulent to virulent, and

[16] See also Franklin (1986), Franklin (2002), and Mayo (1996) for discussions of these strategies and others employed by scientists to minimize the risk of bias or error.

[17] I describe this case in more detail in Stegenga (2011).

from rough to smooth). Avery's own critical assessment of Griffith's results was based on (Q): "For many months, Avery refused to accept the validity of this claim [transformation] and was inclined to regard the finding as due to inadequate experimental controls" (quoted in Dubos 1956). The phenomenon of transformation was surprising, and if it was real it was possibly a kind of hereditary phenomenon.

Soon Avery's colleagues at the Rockefeller Institute had replicated Griffith's results, and had isolated the substance responsible for transformation (e.g. Dawson 1928). Alloway (1933) provided an early clue to the chemical identity of the "transforming substance" (TS): when he added the TS to alcohol, "a thick syrupy precipitate formed." Commenting on this, Avery said that "the transforming agent could hardly be carbohydrate, did not match very well with protein," and so Avery is reported to have "wistfully suggested that it might be a nucleic acid" (Hotchkiss 1965). However, it was assumed by most that the TS was a protein: proteins were known to be highly variable, whereas nucleic acids were thought to be a repetitive structural molecule, like collagen. This was partly the legacy of one of Avery's colleagues, Levene, who had proposed the tetranucleotide hypothesis for the structure of nucleotides (Levene 1921). The structure of TS was assumed to be complex, because the phenotypic features transferred between pneumococcal types were complex: a complex function, it was thought, must be caused by a complex structure.

By late 1940 MacLeod and Avery were attempting to improve the isolation and preservation of the TS, and in 1941 they had begun experiments to determine its chemical identity.[18] In February 1944 their paper was published providing evidence that the TS was DNA. This evidence was 'multimodal'—that is, the paper reported evidence from multiple methods for the hypothesis of interest. A reconstruction of the hypothesis and evidence of *AMM1944* is as follows:

Hypothesis H_1 the molecule which causes transformation (the TS) is DNA.

Evidence e:

method 1: chemical analysis of TS
$\mathbf{e_1}$: the amounts of carbon, hydrogen, nitrogen, and phosphorous were close to the theoretical values for DNA

method 2: application of trypsin, chymotrypsin, and ribonuclease—protein and ribonucleic acid degrading enzymes—on TS
$\mathbf{e_2}$: protein and ribonucleic acid degrading enzymes had no effect on TS

method 3: application of DNA-degrading enzyme on TS
$\mathbf{e_3}$: DNA-degrading enzyme inactivated the TS

method 4: ultraviolet absorption of TS
$\mathbf{e_4}$: ultraviolet absorption of TS was characteristic of DNA

method 5: electrophoretic movement of TS
$\mathbf{e_5}$: electrophoretic movement of TS was characteristic of DNA

method 6: molecular weight analysis of TS
$\mathbf{e_6}$: molecular weight of TS was characteristic of DNA

[18] See e.g. MacLeod and Avery, 22 October 1940. "Laboratory Notes"; MacLeod and Avery, 28 January 1941. "Laboratory Notes".

The final sentence of the discussion in *AMM1944* reads: "If the results of the present study on the chemical nature of the transforming principle are confirmed, then nucleic acids must be regarded as possessing biological specificity the chemical basis of which is as yet undetermined." In a letter to his brother (1943) Avery asked "Who could have guessed it?" Supporting H_1 was surprising.

Assessments of the evidence presented in *AMM1944* were based on both the general features of methods—(Q), (R), and (T)—and the general features of evidence—(P), (C), and (B). The main experimental concern was that the TS was likely impure, and could have had trace amounts of protein in it which caused the transformation. The chemical tests available at the time were not sensitive enough to detect the presence of up to 5 % protein, and the enzymatic experiments could conceivably have been ineffective in degrading an active protein, especially if it was covered by structural nucleic acids. This criticism, directed at (Q), was voiced by Mirsky, one of Avery's colleagues, "frequently in personal conversations" (cited in McCarty 1986) and later in print: "…it is not yet known which the transforming agent is—a nucleic acid or a nucleoprotein. To claim more, would be going beyond the experimental evidence" (Mirsky and Pollister 1946). But *AMM1944* did not claim more: the final paragraph of *AMM1944* itself suggested that (Q) might be problematic:

(Q) "It is, of course, possible that the biological activity of the substance described is not an inherent property of the nucleic acid but is due to minute amounts of some other substance."

The last three sentences begin with "If," "Assuming…," and, again, "If." This cautious rhetoric suggests that, publicly at least, Avery, MacLeod, and McCarty were not "going beyond the experimental evidence."

Although the hypothesis tested in *AMM1944* was specific to the chemical identity of the TS, some considered the phenomenon of transformation as an exemplar of heredity more generally, and thus some thought that the chemical identity of the TS could be the material basis of hereditary phenomena more generally: the TS could be a gene. Avery himself considered this possibility. If this were the case, then the evidence in *AMM1944* could be taken to support hypothesis H_2:

Hypothesis H_2: the class of molecules responsible for heredity is DNA.

Against this, critics noted that transformation had only been demonstrated in bacteria, and it was not clear that bacteria had genes comparable to eukaryotic (non-bacterial) organisms. Even if the TS was DNA, such criticism went, this would not mean that H_2 was true. Thus (R) was an important factor in assessing the evidence in *AMM1944* with respect to H_2: the chemical identity of TS was thought irrelevant to H_2, since there was little reason to believe an auxiliary assumption (bacterial genetics) that was necessary to relate e to H_2. Commenting in retrospect, McCarty claimed that many geneticists "did not consider the bacteria, with their simple life cycles, presumably devoid of any element of sexual reproduction, as suitable for genetic study" (McCarty 1986). Similarly, Morange (1998) put this worry as follows: "the pneumococcus was poorly understood in terms of both its make-up and its biochemical nature. Prior to Avery's work, the only nucleic acid that had

been characterized in this bacterium was RNA. The existence of genes in bacteria was not universally accepted." In short, the relevance of the evidence presented in *AMM1944* to hereditary phenomena more generally was doubted:

(R) Scientists were uncertain if e was relevant to H_2 because they did not know if bacteria had genes

Transformation was a central topic at the 1946 Cold Spring Harbor Symposium, but the interpretations of transformation by some of those involved were non-committal. One author wrote that the TS is "difficult if not impossible to distinguish from viruses" (Anderson 1946). Another participant at this conference, when referring to *AMM1944*, defined transformation as "transmission of genetic material," without mentioning the molecule responsible as either DNA or protein (Hershey 1946). Still another referred to the TS as a "nucleoprotein" (Spiegelman 1946). Although the term "nucleoprotein" seemed to be the most appropriate for the transforming factor, given the concerns regarding (Q), in the two years following publication of *AMM1944* Avery's group had become more confident in their identification of the TS: "accumulated evidence … has established beyond reasonable doubt that the active substance responsible for transformation is a specific nucleic acid of the desoxyribose type."[19] In other words, the group's increase in confidence in H_1 came from an appeal to (C).

The sheer plausibility of the evidence was also considered. For decades it had been assumed that proteins were the hereditary molecule (genes), given their complex structure, and DNA was thought to be a repetitive molecule supporting the transmission of genetic protein. DNA was considered too regular a molecule, with no informational content to be able to provide genetic changes, whereas proteins were known to be diverse in structure and function. It was thought that any phenomena that resembled heredity must be due to complex molecules like protein. That is, it was thought that H_2 is false and instead H_2' was thought to be much more plausible:

Hypothesis H_2' the class of molecules responsible for heredity is protein.

Commenting on *AMM1944* in retrospect, Stanley (1970) suggested that the evidence presented in *AMM1944* was unbelievable at the time of its publication:

(B) "Perhaps of major importance was the fact that the discovery was quite contrary to the dominant thinking of many years."

In the terms discussed in "Conditions of success: evidential features", it was difficult to envision a mechanism in which DNA—thought to be a structurally simple molecule—could play such a central role in functionally complex hereditary phenomena.

The variety of methods used by Avery and his colleagues seemed to permit a favorable assessment based on (C). However, at the time of their publication there was no other evidence for H_1 independent of their own work with which the evidence in *AMM1944* could be concordant. Moreover, all of the various kinds of

[19] From McCarty et al. (1946). See also McCarty and Avery (1946a), McCarty and Avery (1946b).

evidence presented in *AMM1944* relied on the same method of isolating the TS. But in 1952 Hershey and Chase (hereafter *HC1952*) provided evidence concordant with the evidence in *AMM1944* using completely different methods. They labeled bacteriophages (viruses of bacteria) with S^{35} (which labeled only protein) and P^{32} (which labeled only DNA), and found that when the bacteriophage infected bacteria, P^{32} entered the bacteria while most of the S^{35} remained outside the cell. Given that viruses replicate inside the cells of hosts, and apparently only the DNA of viruses entered the host cells, *HC1952* provided independent evidence for H_2.[20]

This point can be made more generally: in the years following the publication of *AMM1944*, its context of evidential assessment shifted; (R), (P), (C), and (B) with respect to the evidence in *AMM1944* became more favorable, and consequently the evidential assessments of *AMM1944* changed, in some cases by those same people who earlier were critics. Mirsky's criticism of the evidence in *AMM1944* was based on (Q); his concern was that protein had contaminated the TS. But results from experiments with DNase (DNA-degrading enzyme) after *AMM1944* further strengthened H_1 (e.g. McCarty 1945). The transformation of bacillus by Boivin provided further confirmation of H_1 using a different organism (1947), and transformation was shown on multiple genetic markers (e.g. Hotchkiss 1951)—thus, H_1 became better confirmed in virtue of a more favorable assessment of the evidence with respect to (C). Genetic recombination in bacteria was demonstrated in 1946 by Lederberg and Tatum, thereby proving that bacteria had genes, a necessary condition to consider the evidence in *AMM1944* as a general hereditary phenomenon. This rendered the evidence presented in *AMM1944* more obviously relevant to general hereditary phenomena—that is, (R) became more favorably assessed. Chargaff challenged Levene's tetranucleotide hypothesis by showing phylogenetic differences in base composition and demonstrating A:T and C:G ratios, making it at least conceivable that DNA could have the complexity necessary for the molecule causally responsible for heredity (1950, 1951)—this rendered the evidence more favorable with respect to (B).

After these developments, the evidence in *AMM1944* could be assessed in light of other evidence generated with a variety of methods, showing consistent patterns of results, and based on new considerations of relevance and believability (bacterial genetics, DNA composition): assessment of the *AMM1944* evidence in terms of conditions of success (R), (P), (C) and (B) had changed. Consequently, the overall assessment of H_2 itself changed. Mirsky himself, once the strongest critic of *AMM1944*, exemplified this change: "that intact nucleic acids have a high degree of specificity in biological systems is evident both from the role of DNA in bacterial transformation (Avery et al. 1944)…" (Mirsky et al. 1956). In an even more striking

[20] However, the primary methodological criticism that was directed at *AMM1944* could have been directed at *HC1952:* the potential for protein contamination in the portion of the virus that entered the cell in Hershey and Chase's experiments was as great as the potential for protein contamination in Avery's TS. Such criticisms against *HC1952* were not as pronounced as they were against *AMM1944*—the evidence in *HC1952* was rapidly accepted, and Hershey went on to win a Nobel Prize. At least one way to understand this is that given *AMM1944*, scientists could then assess *HC1952* favorably by (C). And conversely, once the evidence in *HC1952* was available, the evidence in *AMM1944* could also be reconsidered on the grounds of (C).

change of retrospective assessment, Mirsky (1968) wrote "25 years ago [that is, in 1943], [DNA] was conclusively shown to be the genetic material." It is foremost the *evidence* in *AMM1944* that Mirsky re-evaluated; because of this re-evaluation of e in light of re-evaluations of (R), (P), (C), and (B), Mirsky came to accept not only H_1, but more importantly, he claimed that H_2 had been "conclusively shown". Conclusively, perhaps, but only in hindsight and a context in which (R), (P), (C), and (B) had changed.

One of the strongest retrospective supporters of *AMM1944*, Joshua Lederberg, also changed his assessment of *AMM1944* after (R), (P), (C), and (B) had changed. Lederberg used cautious rhetoric when discussing Avery's work in the mid-1950s; he claimed that the TS is only "intimately connected with the stuff of heredity" (1956)—intimately, perhaps structurally, but not necessarily causally or functionally connected. Until transformation studies were "broadened about 1951 with experiments on drug resistance and other markers, a variety of opinions were forwarded" regarding the TS. Lederberg warned the reader to take note of the valid criticisms, by Mirsky and others, against over-interpreting transformation studies: e should only be construed as weak evidence for H_1. But in Lederberg's Nobel speech (1958) he claimed that "by 1943, Avery and his colleagues had shown that this inherited trait was transmitted from one pneumococcal strain to another by DNA. The general transmission of other traits by the same mechanism can only mean that DNA comprises the genes." Thus by 1958, in the prestigious forum of his Nobel speech, Lederberg was retrospectively claiming that the evidence in *AMM1944* supported not only H_1 but also the stronger H_2.

In sum, when assessing the evidence presented in *AMM1944*, both positively and negatively, scientists appealed to general features of evidence (P), (C), and (B), and general features of methods (Q) and (R).

Illustration: water memory

That the conditions of success are appealed to when assessing evidence is apparent when considering cases of extreme criticism. Evidence may be criticized on the grounds of (P), (C), and (B), and when evidence suggests something truly surprising, the evidence can be (and often is) criticized on the grounds of (Q) or (R). The following example is, like the case described above, a good illustration, since the phenomenon under investigation was inexplicable according to paradigmatic physical chemistry, and the evidence was widely and critically assessed.

When human basophils (white blood cells involved in immune defense) are exposed to a certain antibody (anti-IgE antibodies), they become "degranulated" (due to a physiological response the cells look differently under a microscope). In 1988 *Nature* published a now infamous paper from a research group led by the well-known French immunologist Jacques Benveniste, demonstrating that such degranulation occurs after anti-IgE is diluted by a factor of 10^{120} in water (Davenas et al. 1988). At this dilution no antibody remains in the solution. Benveniste coined the term 'water memory' to explain this phenomenon:

Hypothesis H_w water can retain a 'memory' of molecules dissolved to near infinite dilution.

Evidence:

method: degranulation of basophils by solutions of high dilution anti-IgE
e: degranulation occurs by anti-IgE diluted up to a factor of 10^{120} in water

H_w, if true, would provide theoretical support to homeopathy, since it is a common practice in homeopathy to treat patients with extreme dilutions of substances, under the assumption that the solute retains a memory of the substance which can stimulate one's immune system. (R) was directly invoked, by both critics and defenders of homeopathy, because the evidence presented in this paper was relevant to general homeopathic theory. An article in the magazine Newsweek, for instance, was titled "Can Water Remember? Homeopathy Finds Scientific Support." Defenders of homeopathy took e to be relevant to homeopathy and thereby concluded that homeopathic theory had received some degree of confirmation from this evidence; skeptics of homeopathy took e to be relevant to homeopathy and, conversely, concluded that since homeopathic theory could not possibly be confirmed by any evidence there must have been a problem with this particular evidence.

The paper was accompanied by an editorial written by Maddox et al. (1988), then editor of *Nature*, titled "When to believe the unbelievable", in which Maddox made the following remarks that relate to (P), (C), and (B) respectively:

(P) "there is a *surprising rhythmic fluctuation* in the activity of the solution"
(C) "there is *no evidence of any other kind* to suggest that such behaviour may be within the bounds of possibility … when told … that the experiments should be *repeated at an independent laboratory*, he [Benveniste] arranged for this to be done"
(B) "there is *no physical basis* for such an activity"; the findings "are startling not merely because they point to a novel phenomenon, but because they strike at the roots of two centuries of observation and rationalization of physical phenomenon"

Commenting later, the deputy editor of *Nature*, Peter Newmark, claimed:

(Q) "our referees could not see what the flaw was"

Prior to publishing the paper, Maddox had requested independent replications of the results from Benveniste, and Benveniste had complied: other laboratories from around the world had confirmed his results. In short, Maddox's accompanying editorial focused on (P), (C), and (B).

Nature's readers were critical of the methodology: for them (Q) was an important factor. One letter to *Nature* wrote that:

(Q) "an important control experiment has been overlooked […] one might wonder to what extent this observation can be accounted for by contaminating" (Lasters and Bardiaux 1988)

Another reiterated this concern:

(Q) "I am puzzled by the fact that there has been no control of impurities" (Danchin 1988)

The variability of the data itself—that is, a type of pattern or absence of patterns—was noted:

(P) "one obvious flaw can be seen when looking at the standard errors …" (Fierz 1988)

One letter writer resorted to ridicule: "the paper demonstrating that dilutions of anti-IgE must be vortexed rather than stirred in order to retain an imprint of the antibody on the solvent elucidates another long-standing question: how James Bond could distinguish Martinis that had been shaken or stirred" (Nisonoff 1988).

Subsequent research has been mixed. Some attempts at replicating similar protocols to the Benveniste lab have succeeded and others have failed. Recently a paper showed that "liquid water essentially loses the memory of persistent correlations in its structure" within 50 femtoseconds (50 millionths of a nanosecond) (Cowan et al. 2005). Thus subsequent research has allowed critiques of e to appeal to (C) and (B).

The paper had been published under the agreement that a team put together by *Nature* could visit the lab. The three-person team included a science journalist and a famous magician (both known as debunkers of fringe scientific claims), in addition to Maddox himself. There was no professional biologist or immunologist in their group (Maddox was a physicist). Their report (1988) criticized the evidence from Benveniste's lab by focusing on (Q):[21]

(Q) "our investigation concentrated exclusively on the experimental system"
(Q) "the extensive series of experiments … are statistically ill-controlled, from which no substantial effort has been made to exclude systematic error, including observer bias"
(Q) "the design of the experiments … is inadequate"
(Q) "the experimental data have been uncritically assessed"

Maddox noted that in the lab's original protocol the experimenters knew which test tubes contained antibody and which test tubes were the controls containing no antibody. Similarly, in a later interview, Maddox claimed:

(Q) "what we found was that his whole team was playing a trick on itself; they very rarely made these measurements blind"

[21] Fraud would be an extreme type of criticism based on (Q). The *Nature* team was less than subtle in such a suggestion: "we were dismayed to learn that the salaries of two of Dr Benveniste's coauthors of the published article are paid for under a contract between INSERM 200 and the French company Boiron et Cie., a supplier of pharmaceuticals and homeopathic medicines, as were our hotel bills." Industry funding of scientific research is, of course, ubiquitous, as Maddox must have been aware.

Thus Maddox speculated that e could be explained by observer bias (though this itself would need explanation, since the hypothesis that the experimenters influenced basophil degranulation is, like H_w, fanciful).[22] The report also criticized the findings based on an unfavorable assessment of (C):

(C) "interpretation [by Benveniste's group] has been clouded by the exclusion of measurements in conflict with the claim"
(C) "the phenomenon described is not reproducible"

By this point in the controversy, Maddox seemed to avoid assessing e on the grounds of (B) and (R), though he had clearly done so before. Benveniste had been comparing himself to Galileo, claiming that his scientific results were being unfairly persecuted in virtue of their inconsistency with widely accepted physical laws—in other words, Benveniste had been claiming that his evidence was unfairly assessed on the grounds of (B). It is possible that Maddox wanted to mitigate this kind of rejoinder by sustaining his criticism on the grounds of (Q) and (C). In any case, the report concluded "that there is no substantial basis for the claim that antiIgE at high dilution (by factors as great as 10^{120}) retains its biological effectiveness." That is, e is false and so H_w is not justified.

Predicting methodological criticism, in the original paper the authors affirmed that their evidence was "established under stringent experimental conditions, such as blind double-coded procedures involving six laboratories from four countries." Benveniste's subsequent defense against the charges of the *Nature* team was also based on (Q) (1988)—he claimed that the *Nature* visit was "a mockery of scientific inquiry" and that "the judgment is based on one dilution tested on two bloods in awful technical and psychological conditions." As discussed above, (Q) is not straightforward to assess. In a subsequent interview Benveniste complained about the stressful conditions of the visit by the *Nature* team, and he claimed that his original experiments were not replicated properly, given the lack of collegiality during the *Nature* visit.

In sum, the evidence presented in the infamous paper by Benveniste and his colleagues was assessed by features of evidence (P), (C), and (B), and by features of methods (Q) and (R).

Discussion

The signs of success tradition describes what one achieves once one has reliable evidence, whereas the conditions of success tradition describes the normative

[22] A retrospective comment by one of Benveniste's co-authors on the original paper, Francis Beauvais, lends some support to this methodological concern. He claimed that unblinded experiments usually showed a positive effect, but "the results of blinded samples were almost always at random and did not fit the expected results: some 'controls' were active and some 'active' samples were without effect on the biological system" (Beauvais 2008).

strategies scientists use to generate reliable evidence.[23] The conditions under which something is considered good evidence have been often discussed by both philosophers and scientists, but this tradition of theorizing about evidence has often lacked the aim of generality of scope that the signs of success traditions has had. In epidemiology, for instance, some claim that evidence from randomized controlled trials (RCTs) is superior to evidence from case–control studies, while others dispute this claim. Some suggest that computer simulations can provide evidence for a hypothesis, while others dispute this. Many have claimed that concordant multimodal evidence is epistemically valuable. And so on. In this paper I have attempted to gather such considerations under the umbrella I am calling "conditions of success". Although the set of features I have identified is likely incomplete, they are primary considerations when assessing evidence, as illustrated by the two cases discussed above.[24] I have highlighted both the plurality of the conditions of success and the complexity of assessing each of the individual conditions.

General features of methods include quality, relevance, and transparency, and general features of evidence include concordance, patterns, and believability. This account of evidence is meant to be in terms which are both general and based on conditions of success, rather than most previous philosophical accounts which are in terms which are highly particular or based on signs of success.

One might note that my list of desiderata for assessing evidence is akin to lists of desiderata that bear on the assessment of *theories*. Such features are sometimes called 'epistemic virtues' or 'epistemic values'. For instance, Hempel gave the following criteria of theory assessment: simplicity, support by more general theories, prediction of novel phenomena, and plausibility with respect to background knowledge (1966). Kuhn's notorious criteria were accuracy, consistency, scope, simplicity, and fruitfulness (1977). Van Fraassen includes elegance, simplicity, completeness, unifying power, and explanatory power, and urges that these are merely 'pragmatic' rather than truth-conducive (1980). Lycan lists simplicity, testability, fertility, neatness, conservativeness, and generality (1998). Longino (1994) provides a list of 'feminist theoretical virtues', which include ontological heterogeneity, mutuality of interaction, applicability to human needs, accessibility of ideas, and novelty (see also Wylie 1995). The focus of these desiderata is on theory choice. My focus is, instead, on the assessment of the evidence itself. What distinguishes the desiderata presented here is that past lists of epistemic virtues were comprised of features of theories rather than features of evidence or of methods. Prior to the appeal to epistemic virtues to aid in theory

[23] The contrast between the two traditions of accounts of evidence that I have called *signs of success* tradition and the *conditions of success* is similar to Musgrave's contrast between what he calls logical accounts of confirmation and historical accounts of confirmation (1974), and is also similar to Mayo's contrast between what she calls evidential-relationship accounts of inference and testing accounts of inference (1996). See also Love (forthcoming) for a discussion of formal versus material theories of scientific inference.

[24] For an account of the assessment of evidence in molecular biology which places emphasis on the role of theory and model building rather than on features of methods and of evidence, see Schindler (2008). Similarly, Weber (2002) presents an excellent discussion of a case from biochemistry in which incommensurability of theories was overcome empirically. On the other hand, the features of evidence discussed here may be more prominent in exploratory experimentation (see Elliott 2007).

choice, I have argued that multiple evidential features must be appealed to when assessing evidence.

The signs of success tradition and the conditions of success tradition are not necessarily at odds with each other, since they describe different aspects of evidence: the signs of success tradition describes what is achieved once one has reliable evidence, whereas the conditions of success tradition describes methods of generating and identifying reliable evidence. However, characterization of evidence in terms of the conditions of success has the virtue of accurately describing those aspects of evidence which appear to matter to scientists. I have emphasized the normative aspects of the conditions of success; that is, I have discussed why these features are conditions of *success*. Scientists appeal to such conditions because, when these conditions are met, evidence is thought to be more truth-conducive. Although I have quickly sketched some ways in which some have argued for the importance of the conditions of success, showing precisely that, and how, the conditions of success are in fact truth-conducive cannot adequately be done here. This must wait for future work.

Acknowledgments Financial support was provided by the Banting Postdoctoral Fellowships Program administered by the Social Sciences and Humanities Research Council of Canada. I am grateful for commentary from Nancy Cartwright, Eran Tal, Boaz Miller, Deborah Mayo, and two anonymous reviewers, and for discussion with audiences at the Canadian Society for the History and Philosophy of Science, the University of California San Diego, and the University of Pittsburgh.

References

Achinstein P (2001) The book of evidence. Oxford University Press, Oxford

Allchin D (1996) Cellular and theoretical chimeras: piecing together how cells process energy. Stud Hist Philos Sci 27:31–41

Alloway JL (1933) Further observations on the use of pneumococcus extracts in effecting transformation of type in vitro. J Exp Med 57:265–278

Anderson TF (1946) Morphological and chemical relations in viruses and bacteriophages. Cold Spring Harb Symp Quant Biol 11:1–13

Avery OT, MacLeod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79:137–158

Beauvais F (2008) Memory of water and blinding. Homeopathy 97(1):41–42

Bechtel W (2000) From imaging to believing: epistemic issue in generating biological data. In: Creath R, Maienschein J (eds) Epistemology and biology. Cambridge University Press, Cambridge, pp 138–163

Bechtel W (2006) Discovering cell mechanisms. Cambridge University Press, New York

Bechtel W, Richardson RC (1993) Discovering complexity: decomposition and localization as strategies in scientific research. Princeton University Press, Princeton

Benveniste J (1988) Dr Jacques Benveniste replies. Nature 334:291

Bogen J, Woodward J (1988) Saving the Phenomena. Philos Rev 97:303–352

Boivin A (1947) Directed mutation in colon bacilli, by an inducing principle of desoxyribonucleic nature: its meaning for the general biochemistry of heredity. Cold Spring Harb Symp Quant Biol 12:7–17

Burian R (1993) Unification and coherence as methodological objectives in the biological sciences. Biol Philos 8:301–318

Cartwright N (1999) The dappled world. Cambridge University Press, Cambridge

Cartwright N (2007) Are RCTs the gold standard? Biosocieties 2:11–20

Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymic degradation. Experientia 6:201–209

Chargaff E (1951) Structure and function of nucleic acids as cell constituents. Fed Proc 10:654–659
Collins H (1985) Changing order: replication and induction in scientific practice. University of Chicago Press, Chicago
Cowan ML, Bruner BD, Huse N, Dwyer JR, Chugh B, Nibbering ETJ, Elsaesser T, Miller RJD (2005) Ultrafast memory loss and energy redistribution in the hydrogen bond network of liquid H_2O. Nature 434:199–202
Danchin A (1988) Explanation of benveniste. Nature 334:286
Darden L, Craver C (2002) Strategies in the interfield discovery of the mechanism of protein synthesis. Stud Hist Philos Biol Biomed Sci 33C:1–28
Davenas E, Beauvais F, Amara J, Oberbaum M, Robinzon B, Miadonna A, Tedeschi A, Pomeranz B, Fortner P, Belon P, Sainte-Laudy J, Poitevin B, Benveniste J (1988) Human basophil degranulation triggered by very dilute antiserum against IgE. Nature 333:816–818
Dawson MH (1928) The Interconvertibility of 'R' and 'S' Forms of Pneumococcus. J Exp Med 47(4):577–591
Dubos RJ (1956) Obituary of O. T. Avery 1877–1955. Biograph Mem Fellow R Soc 2:35–48
Elliott K (2007) Varieties of exploratory experimentation in nanotoxicology. Hist Philos Life Sci 29:311–334
Fierz W (1988) Explanation of Benveniste. Nature 334:286
Franklin A (1986) The neglect of experiment. Cambridge University Press, Cambridge
Franklin A (2002) Selectivity and discord: two problems of experiment. University of Pittsburgh Press, Pittsburgh
Griffith F (1928) The significance of pneumococcal types. J Hyg 27(2):113–159
Griffiths PE (2002) Molecular and developmental biology. In: Machamer PK, Silberstein M (eds) The Blackwell guide to philosophy of science. Blackwell, Oxford
Hacking I (1983) Representing and intervening. Cambridge UniversityPress
Hempel C (1966) Philosophy of natural science. Prentice-Hall, Englewood Cliffs
Hershey AD (1946) Spontaneous mutations in bacterial viruses. Cold Spring Harb Symp Quant Biol 11:67–77
Hershey AD, Chase M (1952) Independent functions of viral proteins and nucleic acid in growth of bacteriophage. J Gen Physiol 36:39–56
Hotchkiss RD (1951) Transfer of penicillin resistance in pneumococci by the desoxyribonucleate fractions from resistant cultures. Cold Spring Harb Symp Quant Biol 16:457–461
Hotchkiss RD (1965) Ostwald T. Avery. Genetics 51:1–10
Kuhn T (1977) Objectivity, value judgment, and theory choice. In The essential tension: selected studies in scientific tradition and change. University of Chicago Press, Chicago
Lasters I, Bardiaux M (1988) Explanation of Benveniste. Nature 334:285–286
Lederberg J (1956) Genetic transduction. Am Sci 44:264–280
Lederberg J (1958) Nobel lecture. Available at www.nobel.se
Lederberg J (1994) The Transformation of Genetics by DNA: an anniversary celebration of avery, MacLeod and McCarty, 1944. Genetics 136:423–426
Lederberg J, Tatum EL (1946) Gene recombination in *Eschericia coli*. Nature 58:558
Levene PA (1921) On the structure of thymus nucleic acid and on its possible bearing on the structure of plant nucleic acid. J Biol Chem 48:119–125
Longino H (1994) In search of feminist epistemology. The Monist 77:472–485
Love A (forthcoming) Material versus formal theories in philosophy of science: a methodological interpretation. In de Regt, Okasha, Hartmann (eds) Proceedings of the second European philosophy of science association meeting. Springer, Berlin
Lycan W (1998) Theoretical (Epistemic) virtues. In Craig E (ed) Routledge encyclopedia of philosophy vol 9. Routledge, London. pp 340–343
MacLeod CM, Avery OT (1940) Laboratory Notes: Exp. 1 (T[ransforming]. P[rinciple].) Effect of Fluoride on Autolysis of Pneumococcus Type III and on Preservation of the Transforming Principle"
MacLeod CM, Avery OT (1941) Laboratory notes: effect of ribonuclease on Deproteinized Extract 5-40
Maddox J, Randi J, Stewart WW (1988) 'High-dilution' experiments a delusion. Nature 334:287–290
Mayo Deborah (1996) Error and the growth of experimental knowledge. University of Chicago Press, Chicago
Mayo D, Spanos A (2006) Severe testing as a basic concept in a Neyman-pearson philosophy of induction. Brit J Philos Sci 57:323–357

McAllister J (1997) Phenomena and patterns in data sets. Erkenntnis 47:217–228
McCarty M (1945) Reversible inactivation of the substance inducing transformation of pneumococcal types. J Exp Med 81:501–514
McCarty M (1986) The transforming principle: discovering that genes are made of DNA. Norton & Company
McCarty M, Avery OT (1946a) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. II. Effect of desoxyribonuclease on the biological activity of the transforming substance. J Exp Med 83:89–96
McCarty M, Avery OT (1946b) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. III. An improved method for the isolation of the transforming substance and its application to pneumococcus types II, III, and VI. J Exp Med 83:97–104
McCarty M, Taylor HE, Avery OT (1946) Biochemical studies of environmental factors essential in transformation of pneumococcal types. Cold Spring Harb Symp Quant Biol 11:177–183
Mirsky AE (1968) The discovery of DNA. Sci Am 218:78–88
Mirsky AE, Pollister WW (1946) Chromosin, a desoxyribose nucleoprotein complex of the cell nucleus. J Gen Physiol 30:117–148
Mirsky A, Osawa S, Allfrey V (1956) The nucleus as a site of biochemical activity. Cold Spring Harb Symp Quant Biol 21:49–74
Morange M (1998) A history of molecular biology. Harvard University Press
Musgrave A (1974) Logical versus historical theories of confirmation. Brit J Philos Sci 25:1–23
Neta R (2008) What evidence do you have? Brit J Philos Sci 59:89–119
Nisonoff A (1988) Explanation of Benveniste. Nature 334:286
Quine WVO (1951) Two Dogmas of empiricism. Philos Rev 60:20–43
Rasmussen N (2001) Evolving scientific epistemologies and the artifacts of empirical philosophy of science: a reply concerning mesosomes. Biol Philos 16:629–654
Rheinberger H-J (1997) Experimental complexity in biology: some epistemological and historical remarks. Philos Sci 64:245–254
Roush S (2005) Tracking truth. Oxford University Press, Oxford
Schindler S (2008) Model, theory, and evidence in the discovery of the DNA structure. Brit J Philos Sci 59:619–658
Spiegelman S (1946) Nuclear and cytoplasmic factors controlling enzymatic constitution. Cold Spring Harb Symp Quant Biol 11:256–277
Stanley WM (1970) The 'Undiscovered' discovery. Arch Environ Health 2:256–262
Stegenga J (2009) Robustness, discordance, and relevance. Philos Sci 76:650–661
Stegenga J (2011) The chemical characterization of the gene: vicissitudes of evidential assessment. Hist Philos Life Sci 33:105–127
Strevens M (2009) Objective evidence and absence: comment on sober. Philos Stud 143:91–100
Tabery J (2009) Difference mechanisms: explaining variation with mechanisms. Biol Philos 24(5):645–664
van Fraassen B (1980) The scientific image. Clarendon Press, Oxford
Waters CK (2007) The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. Hist Philos Life Sci 29:275–284
Weber M (2002) Incommensurability and theory comparison in experimental biology. Biol Philos 17(2):155–169
Weber M (2005) Philosophy of experimental biology. Cambridge University Press, Cambridge
Weber M (2012) Experiment in biology. In Zalta EN (ed) The stanford encyclopedia of philosophy (Spring 2012 Edition), http://plato.stanford.edu/archives/spr2012/entries/biology-experiment
Williamson T (2000) Knowledge and its limits. Oxford University Press, Oxford
Woodward J (1989) Data and phenomena. Synthese 79:393–472
Wylie A (1995) Doing philosophy as a feminist: longino on the search for a feminist epistemology. Philos Topics 23:345–358