Brit. J. Phil. Sci. **55** (2004), 1–34, axh101

Prediction Versus Accommodation and the Risk of Overfitting

Christopher Hitchcock and Elliott Sober

ABSTRACT

When a scientist *uses* an observation to formulate a theory, it is no surprise that the resulting theory accurately captures that observation. However, when the theory makes a *novel* prediction—when it *predicts* an observation that was not used in its formulation—this seems to provide more substantial confirmation of the theory. This paper presents a new approach to the vexed problem of understanding the epistemic difference between *prediction* and *accommodation*. In fact, there are several problems that need to be disentangled; in all of them, the key is the concept of *overfitting*. We float the hypothesis that accommodation is a defective methodology only when the methods used to accommodate the data fail to guard against the risk of overfitting. We connect our analysis with the proposals that other philosophers have made. We also discuss its bearing on the conflict between instrumentalism and scientific realism.

1 Introduction

In textbook presentations of the evidence that supports a scientific theory, successful prediction of novel phenomena is often given pride of place. Some examples have achieved the status of legend. In 1818, Augustin Fresnel submitted an essay to the French Académie des Sciences in which he used a wave theory of light to motivate a formal treatment of optical diffraction. Poisson derived an unforeseen consequence from Fresnel's formulae: when a circular disk is illuminated by a point source of light, the center of its geometric shadow will be illuminated as brightly as if the disk were not present. This prediction was experimentally confirmed by Arago. In 1871, Mendeleev used his periodic table to predict the existence of three unknown elements; within 15 years, gallium, scandium, and germanium had been discovered. Einstein used his general theory of relativity to predict the bending of starlight by massive objects such as the sun. Eddington later confirmed this prediction with observations made during the eclipse of 1919. Of course, these theories also fit many phenomena that were previously known. Fresnel's theory fit the known phenomenon of linear diffraction; Mendeleev's table accommodated the 62 known elements; Einstein's General Theory fit the advance of the perihelion of Mercury (known since the middle of the nineteenth century), not to mention many of the confirmed consequences of special relativity and classical gravitational theory. The prediction of a novel phenomenon is certainly more *dramatic* than the accommodation of some previously known phenomenon, but does the former provide *better support* for a theory than the latter? Is the difference genuinely *epistemological*, or merely *psychological*?

This issue has engaged philosophers of science for more than a century and a half. Whewell ([1840]) emphasized the importance of predictive novelty in his philosophy of inductive science, while Mill ([1843]) claimed that no serious scientific mind could grant more than a psychological distinction between prediction and accommodation. The issue drives a wedge between what Musgrave ([1974]) calls *logical* and *historical* theories of confirmation. According to the former, the extent to which data D confirms theory T is purely a function of the logical and mathematical relationships that connect T and D (and the relevant background knowledge); the times at which T was propounded and D was known are incidental, and have nothing to do with confirmation. Logical theories of confirmation include Hempel's ([1945]) theory of instance confirmation, Glymour's ([1980]) bootstrap theory of confirmation, and the likelihood approach of Edwards ([1972]) and Royall ([1997]). By contrast, historical theories of confirmation assert that the time at which a theory was propounded, and even the thought processes that went into its construction, can affect the theory's epistemic status. Popper's falsificationist methodology (Popper [1959]) fits this model. A theory that successfully predicts some novel phenomenon is *corroborated*, since it has survived an attempt at falsification, whereas a theory that accommodates some previously known phenomenon is not corroborated, since it did not run the risk of falsification. Classical statistical hypothesis-testing also conforms to this pattern, with its demand that hypotheses, as well as sample sizes and significance levels, be specified before the data are gathered.

In this paper, we develop a novel approach to the problem of prediction versus accommodation. We adopt a broadly instrumentalist perspective, according to which an important goal of scientific theorizing is to identify hypotheses that generate accurate predictions.[1] We do not claim that this is the only legitimate goal of science; science can have many goals. But attending to the goal of predictive accuracy provides a powerful purchase on the problem of prediction versus accommodation. Theorists who aim to accommodate some known set of data run the risk of committing a methodological sin—*overfitting* the data; and overfitting is a sin precisely because it undermines the goal of predictive accuracy. Prediction is at least sometimes better than accommodation, both because it can provide a measure of protection against overfitting, and because successful prediction can provide evidence that overfitting has not occurred. When it is known that the data have been accommodated while guarding against the risk of overfitting, however, both of these advantages are lost. Indeed, since accommodation involves making use of relevant evidence, accommodation will sometimes be superior to prediction. We develop these claims in detail using a hypothetical example. While it remains to be seen whether the vastly more complex examples mentioned in the opening paragraph can be subjected to a similar analysis, we hope our account will at least highly suggestive of how the story might go in at least some of those examples.

2 Predictivisms—a taxonomy

The view that prediction is superior to accommodation may be called ‘predictivism’. ‘Accommodationism’ is normally understood to be the denial of predictivism, but one can imagine an even stronger view (defended by no one that we are aware of) according to which accommodation is actually superior to prediction. Our own view will be a mixture of all three: in some circumstances prediction is superior to accommodation; in others there is no advantage to one over the other; and in a third class of cases, accommodation is actually superior to prediction. Nonetheless, we will initially identify ourselves as predictivists. There are two reasons for this. First, we find it helpful to approach the problem by asking the question: ‘What is wrong with accommodation?’ (as we do in Section 5). Second, it is easiest to give an overview of our position by locating it within a taxonomy of predictivisms. Our taxonomy will be based on three distinctions.

First, we distinguish *global* and *local* predictivisms. Global predictivism maintains that a theory which successfully predicts some observation will always be superior to one that accommodates the same observation. Local predictivism maintains only that prediction is sometimes superior to

[1] This version of instrumentalism has nothing to do with the thesis that theories lack truth values. See Sober ([1998], [2002]) for discussion.

accommodation. We will defend a version of local predictivism. Local predictivism is compatible with local accommodationism: the view that accommodation is sometimes as good as, or even better than, prediction. We will exploit this compatibility and adopt *both* positions.

Second, we distinguish *strong* and *weak* predictivisms.[2] According to strong predictivism, prediction is intrinsically superior to accommodation. The fact that one theory predicts a phenomenon while another merely accommodates it is, by itself, a mark in favor of the former theory. According to weak predictivism, the difference between prediction and accommodation is epistemically relevant only because it tracks or is symptomatic of other differences that are themselves of evidential import.[3] We will defend a version of weak predictivism in this paper.

Weak predictivism is consistent with a certain compromise between logical and historical theories of confirmation. It may be, as the logicist maintains, that once the content of a theory is fully known, and its logical connections with the evidence are made fully explicit, further historical information about how and when the theory was constructed are rendered irrelevant. Nonetheless, information about whether a theory predicts or accommodates some phenomenon may be indirectly relevant, by providing information about what those complete details are likely to be.[4] Our version of predictivism will take this form.

The third distinction concerns different conceptions of 'novelty'. How, precisely, are we to draw the line between predicting novel phenomena and accommodating extant phenomena? Musgrave ([1974]) canvasses three answers to this question: temporal, heuristic, and theoretical. According to the temporal answer, a phenomenon is novel for a theory only if it was unknown at the time the theory was formulated. The heuristic view (defended, e.g., by Zahar [1973] and Worrall [1985, 1989]) maintains that a phenomenon is novel for a theory if the theory was not constructed specifically to accommodate that phenomenon. Finally, Musgrave defends (and attributes to Lakatos [1968]) the theoretical answer: a phenomenon is novel for a theory if it is not predicted by any of that theory's extant rivals.

The last two distinctions cross-classify: one could maintain a strong or a weak predictivism with respect to temporal novelty, a strong or a weak

[2] Our terminology is similar to that adopted by Lipton ([1991]), who draws a distinction between the strong and weak 'advantage theses' (p. 134). Our distinction does not quite align with his, however; see notes 4 and 21 below.

[3] Mayo ([1991], [1996]) and Lange ([2001]) explicitly endorse versions of weak predictivism.

[4] Lipton ([1991]) endorses a position along these lines. In particular, he distinguishes between *actual* and *assessed* support. The actual support is 'the extent to which the data actually render the theory probable,' while the assessed support is the scientist's imperfect judgment of the actual support (p. 151). Lipton then argues that successful prediction enhances a theory's assessed support to a greater extent than successful accommodation. He considers this position to be a version of what he calls the strong advantage thesis, although in our taxonomy it would only constitute weak predictivism.

predictivism with respect to heuristic novelty, and so on. The resulting versions of predictivism need not be mutually exclusive. For example, one could maintain a strong predictivism with respect to heuristic novelty, and thereby commit oneself to a *weak* predictivism with respect to temporal novelty. That is, one might deny that temporal novelty carries any intrinsic epistemic significance, while maintaining that temporal novelty strongly correlates with heuristic novelty, which *is* inherently significant.

We will adopt the heuristic conception of novelty. Versions of predictivism formulated within this conception are subject to a standard criticism:[5] it would seem to make the relevance of evidence to theory depend on psychological factors such as the intentions of the theorist. For example, the anomalous advance of the perihelion of Mercury, known since the middle of the nineteenth century, is often taken to provide strong evidence for Einstein's General Theory of Relativity, which was first used to predict the phenomenon in 1915. Would the strength of this evidence be undermined if it were discovered that Einstein developed his theory in the hope that it would explain Mercury's orbit? (He may have; see Earman and Janssen [1993].) Must physicists and students of methodology comb through Einstein's correspondence in search of information about Einstein's intentions?

Our response to this objection is that it applies only to *strong* predictivisms formulated in terms of the heuristic concept of novelty. We do not maintain that theories which were specifically constructed to accommodate some piece of evidence are *for that very reason* inferior to theories which predict that evidence. Rather, we argue only that when theories are specifically constructed to accommodate data, there is some tendency for those theories to be defective in ways that can be assessed independently of the intentions of the theorist. If we know that a theory was deliberately constructed to accommodate data, we have a *prima facie* reason to be suspicious of that theory; however, our final assessment of the theory does not require that we determine the motivations of the theorist.

Putting all of these pieces together, we are predictivists insofar as we hold the following position: a theory that predicts phenomena that were not used in the construction of that theory is, in certain circumstances, better than a theory that accommodates the same phenomena. This superiority arises because successful prediction of novel phenomena is symptomatic of a certain type of theoretical virtue. Theories that accommodate evidence may also have this virtue, however, and so there will be cases when accommodation is as good as, or even superior to, prediction.

[5] For criticisms of predictivisms cast in terms of the other two conceptions of novelty, see Musgrave ([1974]).

3 Observations

In this section, we will make a number of observations. We take them to be relatively uncontroversial, and claim no originality for them. These observations will help to explain why there is such a diversity of opinion on this issue, and also to provide some *explananda* for our own account to address.

1. *Accommodation is easy*. It is always possible, after the fact, to come up with some hypothesis or other that accommodates a given body of data. The only constraint is the ingenuity of the theorist. This observation is, somehow, intimately connected to the apparent superiority of prediction over accommodation.

2. *Accommodation is not always bad.* Consider the following simple example. Marsha measures the width of her desk with a tape measure. She holds the end of the tape with '0' on it at one end of the desk, and observes that the other end of the desk coincides almost exactly with the line on the tape labeled '150'. On the basis of this observation, she hypothesizes that the desk is between 149 and 151 centimeters wide. Her hypothesis accommodates the data extraordinarily well. The first observation we made above certainly applies in this case: no matter what the result of her measurement turned out to be, Marsha could have formulated a hypothesis that accommodated that result beautifully.[6] (It would not even have required much ingenuity for her to do so.) But Marsha's hypothesis does not seem to be defective in any way. In particular, her hypothesis is not completely devoid of predictive import, since it can be used to predict future desk measurements; it also predicts that she will fail to get the desk through a 140-centimeter-wide doorway without turning it.

3. *Fit with existing data is a good thing.*[7] Nobody thinks that good theories should be constructed *a priori*, ignoring existing evidence in a quest to avoid accommodation. Einstein's General Theory of Relativity, for example, was specifically constructed so that it would closely agree with Newtonian gravitational theory in those domains where the latter was known to fit the data well. Surely the solution to the problem of understanding the epistemic relevance of prediction versus accommodation does not consist in denying the Principle of Total Evidence.

4. *Simplicity matters.* When a set of data can be accommodated by a very simple theory, we are often inclined to conclude that the theory is well

[6] Howson ([1990]) and Mayo ([1996], Chapter 8) use similar examples to criticize strong predictivism.

[7] This observation is echoed in the title of a paper by Colin Howson ([1990]): 'Fitting Your Theory to the Facts: Probably Not Such a Bad Thing After All'.

supported by the data. For example, Kepler formulated his harmonic law, that $T^3 = aR^2$ (where T is the period of a planet's orbit around the sun, R is its mean distance from the sun, and a is a constant) after carefully combing through Brahe's extensive data. The harmonic law is none the worse for having been formulated on the basis of existing data. Hempel ([1966], pp. 37–8) makes a similar claim regarding the Balmer series. Lange ([2001]), tweaking an example of Maher's ([1988]), provides another illustration of this point. Suppose we observe an apparently random sequence of 99 coin tosses, let us say TTHTH … HTTHT. Tony then produces a 'theory' about the pattern of heads and tails; his theory is that the first one hundred tosses are TTHTH … HTTHTH (he tacks on one head at the end). The fact that Tony's theory successfully accommodated the first 99 tosses provides no reason to think that the 100th toss will land heads. (This is Maher's original example.) But suppose that the original sequence had been an alternating sequence of tails and heads. If, after 99 tosses, Tony were to formulate the theory that odd-numbered tosses come up tails while even-numbered tosses come up heads, we would take this theory to be well supported by the first 99 tosses. In particular, we would be willing to wager that the 100th toss will come up heads. (This is Lange's tweak.)

5. *Background theory matters.* Consider Popper's ([1962]) famous critique of Adlerian psychological theory. Suppose a man jumps into a river to save a drowning child, and that the Adlerian explains this action in terms of the man's inferiority complex. Popper complains that the Adlerian could have equally well invoked an inferiority complex to explain the man's actions had he pushed the child into the river instead of saving him. Popper's complaint is directed, not so much against the specific hypothesis that the man has an inferiority complex, as against the background theory from which this hypothesis was drawn. The reason that accommodation is easy, for the Adlerian, is that the explanatory resources of the psychological theory are so plastic. Our first observation was that accommodation is easy, but it is easier for some theories than for others. The more one's background theory makes it easy to accommodate new data, the less the success of that theory in accommodating the data redounds to the credit of the theory.

6. *Accommodational plasticity entails predictive impotence.* When a background theory is sufficiently plastic that it can accommodate any data that may come along, it is in no position to make predictions about what data will come along. For Popper, this predictive impotence renders a theory unfalsifiable, and hence unscientific. We shall put a slightly different spin on this observation. First, accommodational plasticity is not an all-or-nothing affair, but rather comes in degrees. Second, we shall argue that the problem with many highly plastic theories is not that they fail to make *any* predictions, but rather that they can often be expected to make *inaccurate* predictions.

4 Formulating the problem

Instead of talking about prediction and accommodation in the abstract, it will be helpful to formulate the problem more precisely using a hypothetical example. Consider two scientists, Penny Predictor and Annie Accommodator. In the course of their investigations, they acquire identical sets of data. Let us call this data D. Both researchers advance theories on the basis of this data, but there is a difference. Penny formulates her theory after acquiring the initial data fragment D_1, and then uses her theory to make predictions about the remaining data. Her theory predicts the remaining data D_2 (where $D = D_1 \cup D_2$) to a high degree of accuracy. By contrast, Annie formulates her theory only after she has acquired all of the data in D, constructing her theory deliberately so as to accommodate this data. Does Penny's theory, by virtue of accurately predicting data D_2, enjoy a privileged epistemic status that Annie's does not?

This hypothetical example, as described so far, is underspecified, and different ways of fleshing it out lead to different problems. Suppose we stipulate further that Penny and Annie propose the *very same theory* T, and that we *know* what this theory says. This formulation of the problem would be natural if we wished to defend (or criticize) *strong* predictivism, because the cases of Penny and Annie differ only with respect to whether T predicted or accommodated D_2. It is sound methodology to compare cases that differ only with respect to the feature that is alleged to matter. On the other hand, there is a sense in which this version of the problem lacks intellectual urgency. The problem concerns a situation in which there is no real choice to be made. There is only one theory (namely T) available for the phenomenon under investigation, and if forced to make predictions, provide explanations, and so on, we have no choice but to use it.

For this reason, we will concentrate on a different version of the problem. Suppose that Penny constructs a theory T_p that successfully predicts data D_2, whereas Annie constructs a (possibly different) theory, T_a, by accommodating the entire data set D. Does the fact that T_p predicted D_2 whereas T_a was designed to accommodate this data give us reason to believe that T_p is the better theory? We will argue that, in an interesting range of cases, the answer to this question is *yes*.

The second formulation of the problem can be sharpened further. Suppose that Penny and Annie sample n members of some population. For each member sampled, they measure the value of two quantitative variables, X and Y. For instance, they might measure the average daily caloric intake (X) and weight (Y) of American males between the ages of 30 and 60. Penny and

Annie each posit a functional relationship between X and Y that is polynomial in form: $Y = a_r X^r + a_{r-1} X^{r-1} + \cdots + a_1 X + a_0$. It is not assumed that the value of X uniquely determines the value of Y; rather, it is possible that there will be 'noise' in the data.[8] Annie performs measurements on all n individuals before positing her functional relationship T_a. By contrast, Penny measures the first n_1 individuals, and then constructs her polynomial function T_p. For the remaining n_2 individuals (where $n_1 + n_2 = n$), Penny measures the value of X and then uses T_p to predict the value of Y that will be observed. As it turns out, Penny's predictions are highly accurate. The accuracy of these predictions may be measured in different ways. For each individual sampled, we could take the difference between the predicted and the observed values of Y, called the *error*. The sum of the squares of the errors is a widely used measure of fit to data. Alternately, if T_p explicitly includes an error term U—that is, if it has the form $Y = f(X) + U$—and also provides a probability distribution for U, then T_p will specify how *probable* each observed value of Y is.[9] These probabilities may be multiplied (assuming the observations are independent of each other, conditional on T_p) to yield the overall probability of the data D_2 generated in Penny's predictive phase, on the assumption that T_p is correct. This value is called the *likelihood* of T_p with respect to D_2, and is another useful measure of how well T_p fits the data. Subject to certain conditions, it can be shown that the function with the smallest sum of squared errors is identical to the function with the highest likelihood. For technical reasons, it will be most useful for us to measure fit with data by the *logarithm* of the likelihood. This will of course be highest when the likelihood itself is highest.

Now imagine that we are about to participate in a game. More samples will be drawn from the same population, and their X values measured. (It is assumed that the underlying relationship between X and Y remains uniform during this time.) We will make predictions about the values of Y, and will be rewarded or penalized according to the accuracy of our predictions (as measured by the logarithm of the likelihood). We may use either T_a or T_p to formulate our predictions. Do we have any reason to prefer the latter over the former?

[8] There are at least three possibilities here. First, and most hypothetically, it may be that the data are not noisy at all, but conform perfectly to some true but unknown curve $Y = f(X)$, there being no measurement error at all. Nonetheless, because we are interested in comparing the predictive accuracies of different models, it may still be useful to posit probabilistic error distributions in those models. Second, it may be that the true value of Y is uniquely determined by the value of X according to the function $Y = f(X)$, in which case the 'noise' merely reflects deviations of the measured value of Y from the true value of Y. Finally, it may be the case that the true value of Y is determined by factors in addition to X, in which case the 'noise' reflects the influence of these unmeasured factors. In this case, we will follow the literature in defining the 'true curve' to be the relationship between X and Y that holds when all other relevant factors attain their mean values.

[9] If U is continuous, then T_p will not assign a finite probability to each value of Y, but rather a density distribution over values of Y.

The theoretical problem faced by Penny and Annie is relatively simple. Most genuine scientific theorizing involves much more than fitting polynomials to data. On the other hand, there is an important respect in which our formulation of the problem is more realistic than many of the toy examples discussed in the literature. Real data are almost always noisy. Prediction and accommodation are rarely, if ever, a matter of achieving *perfect* fit to data, even when one has the true theory in hand.[10]

In addition to the issue of whether Penny and Annie formulate the same theory T, there are other dimensions along which different versions of our problem can be specified. For example: (1) What should we take to be the goal of theorizing? Do we want theories that are true, approximately true, or predictively accurate (or something else)? (2) Is it known which theories Penny and Annie have formulated, or only that one was predictively successful while the other accommodated the data? (3) What is known about the methods whereby Penny and Annie arrive at their theories? Given this plethora of problems, it is hardly surprising that the issue of prediction vs accommodation has generated a host of conflicting intuitions.

In the version of the problem that we will address, the goal of scientific theorizing is predictive accuracy. Thus our formulation of the problem is instrumentalist in character. There is a sense in which a theory whose predictions accord well with observation is 'approximately true', and we do want a theory that is approximately true in that sense. But we are not looking for a theory that is approximately true in any deeper sense. In particular, we do not care whether the theory we adopt correctly identifies the degree of the true polynomial[11] (for that matter, we do not care whether the true curve is a polynomial at all). But is predictive accuracy what we *really* want from scientific theories? It is *one* of the things we want. We may also want our theories to be true, or approximately true in some deep sense. Science may have different aims, and these aims may come into tension with one another. By adopting a broadly instrumentalist framework, we are already in a position to say something about the relative merits of T_p and T_a. If one goal of theory construction is accurate prediction, then a theory that has already yielded such predictions has already achieved a measure of success vis-a-vis this goal. From an instrumentalist perspective, predictive success is not merely *symptomatic* of scientific success; it is also *constitutive* of scientific success. Nonetheless, we wish to argue for a stronger conclusion as well: the predictive success of T_p with respect to data D_2 can sometimes give us reason to expect that T_p will fare better than T_a with respect to data that has yet to be gathered.

[10] See note 8 above on the usage of the term 'true' in this context.

[11] Note that the polynomial $Y=4X^3+2X^2+X-35$, for example, may be thought of as a polynomial of any degree greater than or equal to 3. There is, however, a *smallest* degree associated with a specific polynomial—in this case 3—and that is what we have in mind here.

In Section **6**, we will further subdivide our problem into distinct cases along the dimensions described above. Of course, in providing solutions to some versions of the problem, others may be left unsettled. Given that the different versions of the problem are rarely disentangled, however, it is not at all clear whether the observations discussed in Section **3** even pertain to other versions of the problem. To the extent that we are able to provide an explanation for all of these observations, it becomes plausible that our version of the problem may be what philosophers have been arguing about all along.

5 What might Annie be doing wrong?

Our first observation in Section **3** was that it is always possible to accommodate data after the fact. In our present statement of the problem, it is possible to formulate this claim more precisely. Data *D* includes *n* pairs of *X* and *Y* values. Assuming that no two *X* values are identical (let's assume that *X* is measured *very* accurately), it is a mathematical theorem that there exists a polynomial of degree $(n-1)$ (i.e. of the form $a_{n-1}X^{n-1} + \ldots + a_1X + a_0$) that *exactly* fits all *n* data points. Any two points can be fit perfectly by a straight line, any three points can be fit perfectly by a parabola, and so on. So Annie can always find a polynomial of degree $(n-1)$ that exactly fits the data *D*. Suppose this is how Annie goes about accommodating the data. Does this fact provide us with a reason for thinking that we will do badly by basing our predictions on Annie's hypothesis?

The answer is *yes*. The data *D* are bound to contain a certain amount of noise in addition to the information they carry about the underlying relationship between *X* and *Y*. By constructing a relatively complex curve that exactly fits the data *D*, Annie is bound to *overfit* the data. That is, she is bound to propose a theory which is too sensitive to idiosyncrasies in the data set *D* that are unlikely to recur in further samples drawn from the same underlying distribution.[12] By contrast, a polynomial of lower degree that fits the data well, but not as well, may make better predictions when new samples are drawn.

For statisticians, applied mathematicians, and other scientists who frequently construct models to make sense of data, it is a well-confirmed fact that more complex models often do poorly when predicting new data. However, this is not just a brute fact: there is a body of mathematical theory that explains why complex models frequently have this defect, and also proposes

[12] This informal characterization of 'overfitting' does not apply when the data are noise-free. However, even in this case, it may still turn out that simpler curves which fit the data less well will make more accurate predictions about new samples, assuming that the new data involve new values of the *X* variable.

remedies that cope with this problem. We will briefly discuss one important result in this literature. Akaike ([1973]) showed that an unbiased estimate of the predictive accuracy of a model can be obtained by considering both its fit-to-data and its simplicity, as measured by the number of adjustable parameters it contains. In fact, Akaike's innovation was not just to show how predictive accuracy can be estimated, but also to define predictive accuracy as a goal in model selection.

As is standard in this literature, the term 'model' is reserved for statements that contain at least one adjustable parameter. The predictive accuracy of a model is defined in terms of its average performance in a two-step process. First, the model is fitted to the data at hand (with adjustable parameters assigned values by maximum likelihood estimation), and then the fitted model is used to predict new data drawn from the same underlying distribution. The fitted model will come close to the new data or not (as measured by the logarithm of the likelihood). One then repeats this process, drawing data, finding the likeliest member $L(M)$ of the model M, and then seeing how well $L(M)$ does in predicting new data. The average (expected) fit of $L(M)$ to new data defines M's predictive accuracy.

Obviously, it is often desirable to find models that will be predictively accurate in this sense. If one knew the true curve, one could calculate the expected predictive accuracy of M.[13] But how is one to tell how well a model will do in this prediction task, given just the single data set that one possesses? Akaike's remarkable result is that (modulo certain assumptions):[14]

> An unbiased estimate of the predictive accuracy of model $M \cong \log [Pr(Data|L(M))] - k$.[15]

There thus are two factors that influence one's estimate—fit-to-data and complexity (as measured by k, the number of adjustable parameters). If a simple and a complex model fit the data equally well, the simpler model will have the higher *AIC* (Akaike Information Criterion) score. For the complex model to have the higher estimated predictive accuracy, it isn't enough that it merely fit the data better than its simpler competitors (that's pretty much inevitable); it must fit the data *sufficiently* better to compensate for the loss in simplicity that it represents.

A cautionary note is in order. Given an arbitrary curve C that belongs to model M with k adjustable parameters, we can compute an *AIC* score for C: it

[13] One would, in addition, need to know the true error distribution, and also the true distribution that determines the probability of sampling different X values.

[14] See Appendix A of Forster and Sober ([1994]) for a discussion of these assumptions.

[15] Forster and Sober ([1994]) recommend that the *AIC* estimate be expressed *per datum* as $(1/n)[\log[Pr(Data|L(M))] - k]$, where n is the number of data. Also, it is common in the literature to multiply this expression by -2 to yield positive scores. For the sake of simplicity, we will stick to the formulation given in the text.

will be log $[Pr(Data|C)] - k$. But this score need not be an estimate of the predictive accuracy of C or of M (or of anything else, for that matter). Akaike's theorem gives us an estimate of the predictive accuracy of a model, on the assumption that that model is fitted to the data via maximum likelihood estimation.

Although we do not have the space to discuss the justification for using *AIC*, we hope that the reader is starting to grasp the relevance of the Akaike framework (which says that predictive accuracy is the goal of model selection) and Akaike's theorem to the problem faced by Annie and Penny. However, there is an even closer connection that exists between Akaike's ideas and the problem of prediction versus accommodation. In addition to *AIC*, there are a variety of other measures that have been discussed in the model selection literature. One of them is termed cross-validation (Stone [1974], [1977]). In this procedure, one divides the n data points in one's data set into two groups. The size of the groups can vary, but let's consider the special case in which $(n - 1)$ data points are placed in the first group, and one data point in the second. The first group of data is used to fit the model; one then sees how well this fitted model does in predicting the single data point in the second group. One then divides the data into two groups again, but with a different data point assigned to the second group, and sees how well the model fitted to the data points in the training set does in predicting the data point in the test set. If there are n data points, this procedure can be carried out n times. The average performance in this task is the model's cross-validation score. One can then compare different models by seeing how well they perform in this prediction problem. It turns out that take-one-out cross-validation (wherein the training set contains $n - 1$ data points and the test set contains 1) is asymptotically equivalent with *AIC* (Stone [1977]). *AIC* explicitly takes the simplicity of the model into account while cross-validation does not. And cross-validation is transparently a method in which one never fits a model to *all* the data, whereas the *AIC* score is computed by considering the entire data set. In spite of these procedural differences, the two methods are very much on the same wavelength. Both procedures are devices for coping with the problem of overfitting.

It should be clear now that there is at least a superficial resemblance between the procedure followed by Penny and the procedure used in take-one-out cross-validation. In both cases, a curve is fitted to a subset of the total data, and then tested against the remainder. There are some important differences, of course. First, Penny fits her curve to the first n_1 data points, rather than to the first $n - 1$. Second, Penny performs this procedure only once, rather than n times. Third, Penny performs this test on only one curve (presumably with the intention of rejecting the curve if the fit with new data is sufficiently poor), rather than comparing the performance of many different models in this two-step process. Nonetheless, the superficial resemblance between the two procedures offers some hope that Penny's procedure will provide some protection against overfitting. We explore this possibility in detail in the next section.

As mentioned, there are other procedures now on the market in the model selection literature: for example, the Bayesian Information Criterion (*BIC*), due to Schwarz ([1978]), assigns even more weight to simplicity than does *AIC*. We will not discuss the merits and demerits of these different proposals (on which see McQuarrie and Tsai [1998], and Burnham and Anderson [1998]) because we think that what is important here is a feature that all of them share—all take account of simplicity (either explicitly or implicitly) and all are considered as devices for solving the problem posed by overfitting. They make it abundantly clear why accommodating the data in a particular way—namely, by simply finding a model that fits the data perfectly—is a bad idea if one's goal is to find models that will be predictively accurate. We choose to use *AIC* in this paper for expository reasons. We will not discuss whether or in what circumstances *AIC* is the best model selection criterion to use, nor is such a defense needed for purposes of our paper.

We conclude this section with three final observations. First, *AIC* and other model selection criteria are by no means limited to 'curve-fitting problems' narrowly construed. They can be applied to any problem in which one is choosing between models that have different numbers of adjustable parameters. They apply, for example, to the task of causal modeling, and help provide insight into why a unified model covering two domains is often superior to a disunified model that provides a separate treatment of each domain (Forster and Sober [1994] and Sober [2003]; see also Burnham and Anderson [1998] for numerous applications, especially in ecology).

Second, the motivation for balancing simplicity against goodness-of-fit, given the goal of finding models that will be predictively accurate, is purely instrumental in character. Even if we know that the true curve is a polynomial of degree r, it may well be that the curve of degree r that best fits the noisy data one has at hand will fare worse in future predictions than a curve of lower degree. And even if we know that all the curves on offer are false (because all contain *idealizations*), there still can be a point to choosing among them—not by saying which is 'probably true' (none of them is), but by saying which has the best chance of making accurate predictions.

Finally, Akaike's definition of predictive accuracy assumes that the X values of the new data are drawn from the same distribution that determined the X values of the old data. It therefore does not address the problem of *extrapolation*, wherein the X values of the new data are known to fall outside the range of X values found in the old. However, attaining the goal of predictive accuracy requires that one guard against overfitting in this setting as well, and giving due weight to simplicity is a means to that end. Indeed, simulations run by Forster ([2000]) suggest that in extrapolation problems, simplicity needs to be accorded even greater weight. Accommodating the data by finding a model that fits them perfectly is a bad idea, regardless of whether one wishes to interpolate or extrapolate.

6 Solutions

We are now in a position to state our solution to the problem—should we prefer Annie's theory or Penny's? There is no one answer to this question; the answer depends on further details. For purposes of illustration, we will consider five specific cases.

Before presenting these cases, two points of clarification are needed. First, the expressions T_a and T_p should be understood to refer *non-rigidly* to the theories that Annie and Penny construct. For example, T_a does not refer rigidly to the specific polynomial that Annie actually postulates. If Annie had uncovered data D' instead of D, then T_a might have been different from what it actually was. Ambiguity on this point has led to a great deal of confusion in discussions of prediction and accommodation, as we shall see in Section **8** below. Second, unless stipulated otherwise, we assume that we are not told the precise polynomials represented by T_a and T_p; all we have to go on is information about the methods that Annie and Penny used to construct their theories.

Case 1. For whatever reason, Annie and Penny end up proposing the same hypothesis—$T_a = T_p$—and we know this to be the case. Here there can be absolutely no reason to prefer one hypothesis over the other: identical hypotheses make identical predictions. The alleged fact that the hypothesis is better supported in one case than the other has no bearing on this formulation of the problem.

Case 2. We know precisely which polynomials are proposed by Penny and by Annie, and we have the wherewithal to compute the *AIC* score of each. In this case, we should choose the polynomial with the higher *AIC* score: it has the higher estimated predictive accuracy. Once we know this, further information about how Penny and Annie formulated their theories becomes irrelevant.

There are some technical difficulties here. While we can indeed calculate the *AIC* scores of the curves in question, some additional conditions need to be in place in order for these scores to be estimates of the predictive accuracies of the underlying models. First, each curve must come from some determinate model. For example, Annie's curve may have five exponents ($x, x^2, \ldots, x^5$), but it does not follow that she accommodated the data by choosing a curve from the family of fifth-degree polynomials. This is a standard problem: in practice, one must often reconstruct the model from which a specific hypothesis was drawn. So long as the model is not overly gerrymandered, there is little problem in using *AIC* as a measure of the predictive accuracy of that model.[16] So there seems to be little harm in treating Annie's curve as though it were fitted to the data from the class of polynomials of degree k_a.

A second problem is that the curves T_a and T_p might not be the best-fitting curves from their respective models, since we have said nothing about how Penny and Annie arrived at their hypotheses, except that Penny did so through prediction, and Annie *via* accommodation. In Penny's case, it is especially unlikely that T_p fits all n data points better than any other curve from the same model, since T_p was produced on the basis of only the first n_1 data. Thus, for reasons discussed above, the *AIC* score of Penny's curve need not be an unbiased estimate of her curve's predictive accuracy (and likewise for Annie). Nonetheless, it is stipulated as part of the present formulation of the problem that Penny's hypothesis accurately predicts the data D_2. Let us interpret this as implying that the log-likelihood of T_p with respect to the total data D is close to that of the best-fitting curve from the same model. That is, let T_p be drawn from model M_p with degree k_p. If Penny were to select $L(M_p)$, the best-fitting curve from M_p, the estimated predictive accuracy would be $log[Pr(D|L(M_p))] - k_p$, by Akaike's theorem. Since T_p does not fit the data as well as $L(M_p)$, its predictive accuracy can be expected to be lower than this value. Nonetheless, if $log[Pr(D|T_p)]$ is close to $log[Pr(D|L(M_p))]$, T_p can be expected to do almost as well as $L(M)$, and the difference will be on the order of $log[Pr(D|L\ (M_p))] - log[Pr(D|T_p)]$. Thus while we cannot be certain that $log[Pr\ (D|T_p)] - k_p$, the *AIC* score of Penny's hypothesis, is an *unbiased* estimator of the expected predicted accuracy of Penny's model M_k, there is a sense in which it provides a 'reasonable' estimate. A similar line of reasoning applies to Annie's hypothesis. There are some fine points here to be addressed in the theory of mathematical statistics. But the central point is unaffected: when we know the specific curves that have been proposed by Penny and Annie, we should assess the merits of those curves directly. Further information about the methods used to formulate those curves is rendered irrelevant.

Case 3. We do not know which polynomials were proposed by either theorist, but we do know something about the methods that each used. Annie fit data D exactly, using the lowest-degree polynomial necessary to do this. With rare exceptions,[17] this will require a polynomial of degree $(n - 1)$. Similarly, Penny fit data D_1 exactly, almost always deploying a polynomial of degree $(n_1 - 1)$. Penny's theory T_p predicts the remaining data D_2 quite well.

Here both Penny and Annie are likely to be guilty of overfitting. So which hypothesis should we rely on for our predictions? Let's consider the *AIC* score of each model. Annie's hypothesis fit the data D perfectly; this might tempt us

[16] This is related to what Forster and Sober ([1994]) call the 'sub-family problem'.

[17] There is a natural sense in which the set of exceptions has measure zero.

to suppose that $P(D|T_a) = 1$, but so long as T_a allows for the possibility of error, $P(D|T_a) = p_a < 1$.[18] What does follow from Annie's perfect fit is that p_a is *maximal.* T_a comes from a family with n $(= n_1 + n_2)$ free parameters, so $AIC(T_a) = log(p_a) - n$. Penny's hypothesis has n_1 free parameters, fits data D_1 perfectly, and data D_2 very well. Let $P(D|T_p) = p_p$. We know that $p_p < p_a$. The *AIC* score for Penny's curve is $log(p_p) - n_1$. For reasons discussed above, this need not be an unbiased estimate of Penny's predictive accuracy. However, if Penny's curve predicts the data D_2 very accurately, we may assume that it will fit the total data D nearly as well as the best-fitting n_1 degree curve; hence the *AIC* score of Penny's curve will provide a reasonable estimate of Penny's predictive accuracy. Comparing *AIC* scores, we get

$$AIC(T_p) > AIC(T_a) \text{ if and only if } log(p_p) - log(p_a) > -n_2.$$

If Penny's theory covers a lot of data beyond the data she used to construct her theory (n_2 is large), and if her theory confers on that new data a very high probability ($log(p_p)$ is close to $log(p_a)$), then Penny's theory will have a higher estimated predictive accuracy than Annie's.

Intuitively, the idea is that by fitting her curve only to the first n_1 data points, Penny is providing herself with some protection against overfitting: she cannot overfit data that she does not yet have. At the same time, she is exposing herself to the risk of underfitting: her curve may simply fail to fit the remaining n_2 data points. But once we learn that Penny's curve in fact does fit the new data well, we know that this threat did not materialize. Penny has succeeded in reaping the benefit without paying the cost.

Case 4. Annie chooses the best-fitting curve from the family of curves that has the highest *AIC* score on the data D. Penny chooses the best-fitting curve from the family of curves that has the best *AIC* score on the basis of D_1. Penny's curve also fits the new data D_2 very well. In this case, T_a is the theory that has the higher estimated predictive accuracy with respect to data D. Given our goal of maximizing predictive accuracy, we can do no better than rely on Annie's theory. It is possible, albeit highly improbable, that Penny's theory T_p fits the data D_2 so well that the estimated predictive accuracy of T_p comes close to that of T_a; but if Annie chose the model with the *best AIC* score relative to *all* the data, Penny's model cannot have a higher estimated predictive accuracy.[19] Annie has done everything that can be done to avoid overfitting.

The moral to be drawn from case 4 is that just because Annie *can* commit the sin of overfitting, it does not follow that she *must* do so. Once pains are taken to avoid overfitting, there is no residual advantage to prediction over accommodation.

[18] If the error is itself treated as a parameter, then it can be set to zero, raising Annie's likelihood to 1, at the cost of introducing an extra parameter.

[19] If the true error distribution has the form of a density function, the possibility that Penny's model is equally good has probability zero.

Indeed, so long as one is not overfitting the data, one should base one's theories on *all* of the data that are available, and the larger data set D is to be preferred over the smaller data set D_1. The virtue of prediction over accommodation does not mean that there is something wrong with the Principle of Total Evidence.

Case 5. As before, Penny formulates T_p on the basis of data D_1, while Annie formulates T_a on the basis of the entire data set D. We are told that T_a fits data D to some specified degree of accuracy (or better), and also that T_p predicts D_2 to some specified degree of accuracy (or better). We do not know for sure what method either Penny or Annie is using. They may be fitting their respective data sets perfectly using relatively high-degree polynomials, as in case 3, or they may be using *AIC* to balance simplicity against goodness of fit, as in case 4. And they need not be using the same method.

This is perhaps the most interesting case. Let us call the methods described in case 3 *overfitting* methods, and those described in case 4 *balancing* methods. In case 5, the fact that T_p successfully predicts D_2 is evidence that Penny's theory avoided overfitting the data D_1. If Penny had drastically overfit the data in D_1, we would not expect her theory to accurately predict D_2. After all, overfitting is undesirable precisely because it leads to unreliable predictions. Thus, the very fact that Penny's theory accurately predicts the novel data D_2 provides evidence that Penny took appropriate care to balance simplicity against goodness of fit. By contrast, the fact that Annie successfully accommodated data D does not give us any reason to think that she was using a balancing method rather than an overfitting method. Indeed, we would expect an overfitting method to provide better fit with data D than a balancing method.

We may express this argument in terms of likelihoods. Let B_a and B_p be the propositions that Annie and Penny respectively obtained their theories by appropriately balancing simplicity and fit; let F_a and F_p be the propositions that they obtained their theories by finding models that perfectly fit the data they used (D in Annie's case, D_1 in Penny's). Let *Fit-via-Pred*(T_p, D_2) be the proposition that Penny's theory accurately fit the data D_2, even though Penny did not use D_2 to formulate T_p. And let *Fit-via-Acc* (T_a, D) be the proposition that Annie's theory accurately fit the data D, and that she *did* use all of D to formulate T_a. We can now assert two likelihood inequalities:

$$\Pr(\text{Fit-via-Pred}(\mathrm{T_p}, \mathrm{D_2}) | \mathrm{B_p}) > \Pr(\text{Fit-via-Pred}(\mathrm{T_p}, \mathrm{D_2}) | \mathrm{F_p})$$

$$\Pr(\text{Fit-via-Acc}(\mathrm{T_a}, \mathrm{D}) | \mathrm{B_a}) < \Pr(\text{Fit-via-Acc}(\mathrm{T_a}, D) | F_\mathrm{a})^{20}$$

[20] Note, however, that this still does not fully settle the matter in favor in T_p. If Annie and Penny both balance, then T_a will likely be better than T_p (case 4); and it *may* be the case (though it need not be) that if both merely found the best-fitting theory, then T_a will be better than T_p (case 3). Thus it could happen that the superiority of T_a *within* each case gives T_a a higher expected predictive accuracy, even though T_a is more likely to overfit the data. This just underscores our point that the problem has no uniform solution, but must be answered on a case-by-case basis.

Both inequalities are true even if, as a matter of fact, Annie and Penny happen to produce the same theory.

Our treatment of this case resembles arguments that have appeared in the literature. Patrick Maher ([1988], [1990]) has argued that prediction is superior to accommodation because the fact that a hypothesis successfully predicts further data gives us reason to believe that the hypothesis was generated by using a reliable method (and hence that it will continue to be predictively accurate). Maher formulates this argument within a Bayesian framework, carefully stating conditions that are sufficient for the argument to go through. However, Maher never says what makes a method reliable, nor why successful prediction is evidence that a method has these features. We have done this explicitly—a reliable method balances simplicity against goodness-of-fit so as to reduce the risk of overfitting, and successful prediction is symptomatic of balancing because hypotheses that drastically overfit the data are unlikely to yield successful predictions.

Lipton ([1991]) also provides an argument that parallels ours. He writes:

> When data need to be accommodated, there is a motive to force a theory [. . .] to make the accommodation. The scientist knows the answer that she must get, and she does whatever it takes to get it. The result may be an unnatural choice or modification of the theory [. . .] that results in [. . .] weak support, a choice she might not have made if she did not already know the answer she ought to get. In the case of prediction, by contrast, there is no motive for fudging, since the scientist does not know the right answer in advance [. . .]. Scientists are aware of the dangers of fudging and the weak support that results when it occurs. Consequently, when they have some reason to believe that fudging has occurred, they have some reason to believe that the support is weak. My claim is that they have such a reason when they know that the evidence was accommodated, a reason that does not apply in cases of prediction. (p. 140)

Once again, the idea is that when a theory is constructed by accommodating the data, this is evidence (not proof) that the method used to construct that theory is unreliable (in the sense that it involves 'forcing' or 'fudging'). Lipton does not define forcing or fudging, although it is clear that for him the defect of forced or fudged theories lies in their inability to provide good explanations of the data they accommodate. In the context of our problem, forcing and fudging correspond to overfitting the data by using a model that contains many adjustable parameters. Theories that have been forced and fudged are defective because they have a low expected predictive accuracy.[21]

Case 5 underwrites another very simple argument. The fact that a hypothesis has made accurate predictions about novel phenomena provides evidence

[21] The fact that this argument remains valid even when Penny and Annie happen to formulate the same theory is what leads Lipton to consider this an argument for what he calls the 'Strong Advantage Thesis'.

that it will continue to do so in the future.[22] This is because predictive accuracy provides evidence that a hypothesis has appropriately balanced simplicity against goodness-of-fit. Since we desire theories that will make accurate predictions, we have some reason to prefer theories with a track record of predictive success.

In cases 2 and 3, we offered quantitative conditions for Penny's theory to be preferable to Annie's. These conditions depend upon our use of *AIC*, but the qualitative points underscored by these cases hold regardless of how we choose to balance simplicity against goodness-of-fit. Readers are invited to construct further cases, and work out the consequences of our analysis for themselves. We hope that our discussion of these five cases brings out the key features of our account. There are certain features of hypotheses by virtue of which they are good or bad predictors. Hypotheses that drastically overfit existing data will probably do a poor job in predicting future data; hypotheses that balance simplicity against goodness-of-fit tend to do better. It is possible to accommodate data without overfitting them, but when one is accommodating data, the temptation to overfit is always present. By contrast, when one accurately predicts new data that were not used in formulating one's theory, there is no opportunity to overfit those data. It is for this reason that our account can be viewed as a version of weak predictivism: there is nothing wrong with accommodation *per se*; accommodation only provides evidence that overfitting has occurred. At the same time, our discussion shows that there is no uniform solution to the problem of prediction versus accommodation, as formulated in Section **4** above; the solution depends upon the details of the case. In particular, it depends upon what is known about the methods used by Penny and Annie to construct their hypotheses, and what is known about the hypotheses themselves.

In order to determine whether prediction is superior to accommodation in some specific scientific example, we need to know how that case assimilates to our breakdown of cases given above. Likewise, in order to determine whether prediction is superior to accommodation in science quite generally, we need to know whether the majority of scientific cases assimilate to cases 3 and 5, where prediction can offer advantages over accommodation; to cases 1 and 2, where prediction offers no advantage, or to case 4, where accommodation is actually superior to prediction. Indeed, a number of the arguments that have been offered in the literature are naturally construed as arguments over just this

[22] This assumes what the proof of Akaike's theorem also assumes—that the underlying reality generating the data remains stable throughout. There is no solution to the problem of induction on offer here! In a similar vein, Akaike's theorem provides no solution to Goodman's ([1955]) new riddle of induction. For example, in a curve-fitting problem, it is *assumed* that *X* and *Y*, and not some gruified construction therefrom, are the independent and dependent variables; for discussion see Forster ([1999]).

point. For example, Howson ([1989]) presents a critique of predictivism in which he claims that 'in science [. . .] predictions are generated from publicly available theories' (p. 384). This can naturally be construed as a claim to the effect that typical scientific examples assimilate to case 2, in which the competing theories are fully known. (See Horwich [1982], p. 117, for a similar argument.) Our analysis of case 2 does indeed show that *if* this is correct, then there would indeed be no advantage to prediction over accommodation. We take no stand on the issue of how often the various cases arise.[23] Nonetheless, it is important that arguments about the prevalence of the different cases be disentangled from arguments about the logic of each. We hope that our taxonomy provides a useful prolegomenon to future debates about predictivism.

7 Observations explained

We are now in a position to account for the observations described in Section **3**.

1. *Accommodation is easy.* It is always possible to accommodate n data if one adopts a theoretical framework (such as a family of polynomial curves) with n (or more) free parameters.

2. *Accommodation is not always bad.* So long as appropriate steps are taken to avoid overfitting the data, there is nothing intrinsically wrong with formulating one's hypothesis on the basis of the available data. Part and parcel of this idea is that it is better, *ceteris paribus*, to base your choice of theory on the more inclusive data set D than on the less inclusive data set D_1. There is nothing wrong with the Principle of Total Evidence.

3. *Fit with existing data is a good thing*. While we have focused on the dangers of overfitting, a theory may also be defective if it *underfits* the data. A theory that contains too few free parameters may be unable to fit the data well enough, as measured by log-likelihood, to receive a high *AIC* score. A theory can be expected to make accurate predictions of new data when it achieves the right balance of simplicity against goodness-of-fit. A theory that places too high a premium on simplicity will fare no better than one that places an excessive premium on fit.

[23] Thus there is a sense in which it is misleading for us to paraphrase our position, as we have done, by saying that successful prediction is symptomatic of successfully balancing simplicity against goodness-of-fit. If 'symptomatic of' means 'statistically correlated with', the paraphrase does not strictly follow from our analysis. It may be, for example, that all actual cases conform to the structure of case 4; if so there will be no correlation between accommodation and overfitting.

4. *Simplicity matters.* A sufficiently simple hypothesis, formulated on the basis of a given body of data, will not drastically overfit that data. It does not contain too many parameters whose values have been set according to the data. Thus a simple hypothesis that successfully accommodates a given body of data can be expected to make more accurate predictions about new data than a more complex theory that fits the data equally well.

5. *Background theory matters.* For a given k, we may think of the family of polynomials of degree k as a background theory. Such a theory determines the form that specific hypotheses can take on the basis of the data. It determines how the data may be used to construct specific hypotheses. A background theory that has a large number of free parameters is able to accurately fit the data more easily than one with fewer free parameters. But as we have seen, such a theory is liable to produce a hypothesis that substantially overfits the data that it accommodates.

6. *Accommodational plasticity entails predictive impotence.* Since a theory that has sufficiently many free parameters to accurately fit whatever data may come along will almost certainly overfit the data at hand, such a theory cannot be expected to generate hypotheses that will make accurate predictions. It is worth emphasizing that the idea of 'too many parameters' depends on the size of the data set. Linear models cannot perfectly fit data sets that contain three or more (non-colinear) data points, but they can perfectly fit any data set that contains just two. Every model of the sort to which the Akaike framework applies can perfectly fit sufficiently small data sets, but will be unable to do this for data sets that are sufficiently large. Our point here departs from Popper's critique of unfalsifiable theories, which he claims make no predictions. The problem with models that overfit is not that they make no predictions, but that they are apt to make inaccurate predictions.

8 Mayo on severe tests

In this section and the next, we argue for the fruitfulness of our approach by applying it to two recent discussions in the philosophy of science.

Deborah Mayo ([1991]; [1996], Chapter 8) has criticized the position that we have called strong predictivism. According to Mayo, in assessing the evidential bearing of data D on theory T, what matters is not whether T was constructed after D was collected, nor even whether D was used in the construction of T. What matters, rather, is whether T has passed a *severe test* with respect to D. We may think of the procedure whereby T is compared with D as a *test*. The test must specify, *inter alia*, the procedure used to collect the data.

Mayo defends a form of weak predictivism; she holds that severity is correlated with novelty:

> What underlies the basic intuition that if the data are not novel, then they fail to test or support a hypothesis are the various impediments to severity that correlate with violating novelty of one sort or another. But the correlation is imperfect. (Mayo [1996], pp. 252–3)

What makes a test *severe* for Mayo? Here is one of her formulations. Given a specific test procedure *P*, theory *T* passes a severe test with data *D* just in case:

> ***ST:*** There is a very low probability that test procedure *P* would yield so good a fit, if *T* is false. ([1996], p. 180; our notation.)

As an illustration, consider an example of ordinary statistical hypothesis testing. Let *T* be the hypothesis that a coin is biased. Let *P* be the procedure of tossing the coin twenty times and recording the number of heads. Let *D* be the result that the coin landed heads on sixteen of twenty tosses. The possible data that would fit *T* as well as *D* does include results with sixteen or more heads, or sixteen or more tails. If *T* were false—i.e., if the coin were fair—the probability of obtaining sixteen or more heads or sixteen or more tails is .046. This is a measure of the severity of the test; the lower the probability, the more severe the test. (Note that Mayo's approach here differs from that of classical statistical hypothesis testing, as usually understood. Classical statistics normally divides problems into two sorts—estimation and hypothesis-testing. In estimation problems, one uses data to formulate a hypothesis about the value of some quantity. In hypothesis-testing, one specifies the hypothesis, as well as the significance level, before collecting the data. Mayo's project involves, *inter alia*, an attempt to unify these two branches of classical statistics.)

As worded here, however, *ST* is ambiguous. It could mean:

> ***ST$_r$:*** There is a very low probability that test procedure *P* would yield so good a fit *with T*, if *T* is false,

where *T* refers rigidly to the theory that was, in fact, being tested. Let us call this the *rigid* construal of *ST*. But *ST* could also be construed to mean:

> ***ST$_{nr}$:*** There is a very low probability that test procedure *P* would yield so good a fit *with whatever theory would have been proposed*, if *T* is false.

In *ST$_{nr}$*, the first occurrence of *T* in *ST$_r$* has been replaced with a description whose referent varies depending upon the data *D* produced. (The one occurrence of '*T*' in *ST$_{nr}$* still refers rigidly.) If *T* is formulated independently of the results of test procedure *P*—if it is formulated prior to the collection of the data, for example—then the two formulations are equivalent. *T* is the theory that would have been proposed in *ST$_{nr}$*. On the other hand, if the theorist formulated *T* on the basis of data produced using procedure *P*, then we can expect that a different theory would have been obtained had *P* produced different data.

To see why the difference between these two construals matters, consider the example, from Section **3**, in which Marsha measures the length of her desk. Here, *P* is the procedure of stretching the tape measure along the side of the desk and noting which number accompanies the slash that is closest to the edge of the desk; *D* is the result that the closest slash is adjacent to the number 150; and *T* is the claim that the desk is between 149 and 151 centimeters wide. Assume, moreover, that Marsha is very reliable in her use of tape measures; it is very unlikely that her measurement will be off by more than one centimeter. It is clear from Mayo's critique of strong predictivism that she would deem this to be a case in which *T* has passed a severe test with *D*. ST_r yields this result: if the desk were *not* between 149 and 151 centimeters long, it is very unlikely that she would have observed the slash marked '150' to be closest to the edge of the desk. By contrast, ST_{nr} is false in this example. Regardless of the length of the desk, we can be almost certain that Marsha will postulate a desk length that fits the result of her measurement as well as *T* fit the actual result.

We are not suggesting that there is a *pervasive* ambiguity in Mayo's treatment of this subject. She offers other formulations (e.g., [1996], p. 181) which make it clear that ST_r is intended; and as just noted, her critique of strong predictivism is clearly consistent with the rigid reading only. Nonetheless, there do appear to be places where she tacitly trades on the ambiguity.

Consider a different example, of a sort that Mayo herself discusses at some length in a different context ([1996], Chapter 9). Suppose that we have twenty coins, and that we are interested in whether the coins are fair or biased. We flip each coin some number of times and determine if the frequency of heads differs from one half by a significant amount. One such coin—suppose it is coin number twelve—comes up heads with a frequency that is statistically significant at the 0.05 level. We therefore hypothesize that coin number twelve is biased. There is obviously something *ad hoc* about this procedure. Even if all of the coins are fair, there is a probability of 0.64 ($= 1 - [0.95]^{20}$) that at least one of the coins will yield a frequency of heads that is significant at the 0.05 level. Mayo argues that in cases such as this we need to draw a distinction between the *computed* significance level (in this case 0.05), and the *true* or *actual* significance level (0.64). This would seem to be a case in which accommodating the data leads to bad results.

But does Mayo's *ST* yield the correct result in this case? The hypothesis under consideration is T_{12}, that coin twelve is biased. If this hypothesis were false, then coin twelve would be fair. If the coin is fair, the probability that the procedure would yield a number of heads on coin twelve that fit T_{12} as well as the actual data did is less than 0.05 (this is just what it means to say that the number of heads was significant at the 0.05 level). So it looks as though the severity of this test of T_{12} has the computed significance value of 0.05. By contrast, what Mayo calls the *true* or *actual* significance value of 0.64 is precisely the probability associated with the *non-rigid* construal of *ST*. That is, 0.64 is the probability that we will formulate *some* theory or other about the bias of a coin that fits the data at the 0.05 level. So in order for Mayo to conclude that T_{12} does not pass a severe test on this data, she must employ ST_{nr} rather than ST_r. However, this is just the construal that yields the wrong answer in the desk-measuring example.

What is going on here? If one were to formulate hypothesis T_{12} first, and then test it, one would do so by tossing coin twelve some number of times and counting the number of heads. One wouldn't need to toss the other nineteen coins, but it would be harmless to do so. Intuitively, there is nothing wrong with how T_{12} was *tested*. Apparently, the problem has to do with the *ad hoc* procedure used to formulate T_{12}. But how could history matter in this way?

Here is how this example would be handled within the framework we have presented. Consider two models. The first model M_1 generates hypotheses that assign the same bias to each coin; the best-fitting hypothesis from this model is determined by estimating the value of a single adjustable parameter. The second model M_{20} generates hypotheses that assign biases to each of the twenty coins independently; this model is fit to the data by setting the values of twenty different parameters. Suppose the data are such that coin twelve exhibits a statistically significant frequency of heads, while none of the other coins do. The best-fitting hypothesis from M_1 will posit the same tiny bias for each coin. The best-fitting hypothesis from M_{20} will posit a large bias for coin twelve, and little or no bias for any of the other coins. The latter hypothesis will certainly fit the data better than the former. Nonetheless, the twenty-parameter family might still receive an inferior *AIC* score, reflecting the ease with which the data can be overfitted when there are twenty adjustable parameters. In other words, we may have reason to believe, on the basis of the data, that the one parameter model which says that the coins do not differ in their biases is likely to make more accurate predictions about future tosses than is the model which says that each coin has its own bias.[24] For that matter,

[24] One might raise the following objection. Consider the one-parameter family of hypotheses whose members all attribute no bias to the other nineteen coins, but differ only in their assignment of bias to coin twelve. The best-fitting hypothesis from this family will be one that attributes a substantial bias to coin twelve (and of course, no bias to any of the other coins). This will fit the data almost as well as the best-fitting member of the twenty-parameter family described in the text, but will not receive the penalty for having twenty adjustable parameters. This objection is just a version of the so-called 'subfamily problem', discussed at length in Forster and Sober ([1994]). When the competing models are gerrymandered in this manner, overfitting may go undetected by formal methods such as *AIC* scores. Nonetheless, we can still have reason for believing that the data will be overfitted by the best fitting hypothesis from a gerrymandered model.

the zero-parameter model which takes all the coins to be fair might score even better; this model corresponds to the 'null hypothesis'. Note that none of this presupposes that we have any independent reason for thinking that the coins all have the same bias (or lack of bias). The name of the game is predictive accuracy, not true description of an underlying reality.

9 The miracle argument and scientific realism

Scientific realism is a cluster of views about the nature of scientific theories and theorizing. While no two scientific realists espouse exactly the same tenets, a common core might include the following:

1. The aim of scientific inquiry is to produce theories that provide descriptions of the world that are literally true.
2. Theories in the 'mature sciences' are usually approximately true, and the entities postulated by those theories usually exist.

One of the most popular arguments proffered by proponents of scientific realism is the so-called 'miracle argument'. According to this argument (see, e.g., Putnam [1975]), scientific realism is capable of explaining why a predictively successful theory is predictively successful, whereas the success of the theory would be miraculous if scientific realism were not true. We reconstruct this as a likelihood argument; the claim is that $Pr(T$ is predictively successful $\mid T$ is true$) >> Pr(T$ is predictively successful $\mid T$ is false). Hence, the predictive success of *T strongly favors* the hypothesis that *T* is true over the hypothesis that *T* is false (Sober [1990]). This argument might seem to lack force if the theory in question was constructed by using the data that it is said to predict successfully. In this instance, we might expect the theory to be predictively successful whether or not it is true. So let us consider the argument with respect to *new data*; what else, besides the truth (or approximate truth) of theory *T* can be said to explain *T*'s success in making novel predictions? (See, e.g., Leplin [1997] for a version of this argument.)

Our treatment of the problem of prediction versus accommodation goes some way towards undermining this argument. Consider the curve-fitting problem characterized in Section **4**. Recall that curve-fitting is a two-step process. First, one must choose a *model*—e.g., the degree of the polynomial that is to be fitted to the data. Then one chooses the best-fitting specific curve from that model. It is natural to think of a model as being analogous to the ontological framework of a scientific theory. A model determines the overall form that the final (fitted) theory will take. A model is true when it correctly picks out the underlying form of the true specific curve. Suppose, for example, that a specific curve *C* is not only successful at accommodating the data D_1

which was used to construct that curve, but also at predicting the novel data D_2. It would be unreasonable to conclude from this that C is *exactly* true. Even if C has the right form, the precise values of the parameters in C are bound to be wrong by at least a small amount. However, it would not be unreasonable, *prima facie*, to infer that C is *approximately* true.

Matters change when we turn our attention to the smallest model M that contains the specific curve C. Is the success of C evidence that M is true or approximately true? There are different things we might mean here. We might mean that M, when fitted to old data, will yield predictions that are close (as measured, say, by log-likelihood) to new data generated by the true curve. To concede that scientific theories are 'approximately true' in this sense would provide little comfort to the realist, however; even the anti-realist can concede that theories are approximately true in this sense. Alternatively, we might mean that M correctly identifies the general form of the true curve—in the case of polynomials, this means that M has the correct (smallest) degree. Does the success of C in predicting the novel data D_2 give us reason to think that M is approximately true in this second sense?

No. The predictive accuracy of a model has little to do with whether the model is true. It may be, for example, that the true specific curve which, together with random noise, generates data is very complex, containing k (non-zero) parameters. This does not rule out the possibility that fitting data with a model that has fewer parameters will be more successful in yielding accurate predictions of novel data. The success of AIC as a strategy in model selection has nothing to do with correctly identifying the true (smallest) number of parameters, and everything to do with the need to avoid overfitting the data. In this case, at least, the connection between approximate truth and predictive success is severed (Sober [1998], [2002]).

We so far have shown that the realist's explanation is not the only one possible. Whereas the realist claims that a model's success in predicting novel phenomena should be explained by the hypothesis that the model is true, we have suggested that it might be explained by the hypothesis that the model is false, although it successfully balances fit-to-data and simplicity. However, we want to say something more. It is a pervasive fact about scientific practice that models known to be false often make more accurate predictions than models known to be true. The point of idealization in science is not just to make calculation tractable; idealizations—by virtue of assuming that certain known causes have no influence on an effect—are typically simpler in a way that often enhances predictive accuracy. It isn't just that an instrumentalist interpretation of a model is possible; the point is that it is often vastly more plausible than a realist construal.

The observant reader will have noticed that we have suggested that realism fares much better in connection with fitted models than it does with respect to models properly so called (i.e., models that contain adjustable parameters). A fitted model that accurately predicts a novel phenomenon has a different epistemological status from an unfitted model that does so when fitted to an old data set. The slogan for this difference might be 'instrumentalism for models, realism for fitted models' (Sober [2002]). One realist intuition that underlies the miracle argument is correct—no fitted model will be more predictively accurate, on average, than the true fitted model. If only we could identify that proposition, we would have no need for false theories. The problem is that the true fitted model is typically epistemically inaccessible. What we can do is assess unfitted models for the fit-to-data of their likeliest members and for their simplicity. If our goal is predictive accuracy, we must abandon our demand for models that are true.

It would be a mistake to think that the miracle argument can be retained by claiming that the instrumentalist ideas we have advanced pertain only to banal inference problems which are data-driven. Curve-fitting often seems to have this character, but model selection encompasses a wide range of inference problems, including the testing of causal models (Forster and Sober [1994]). Models that contain adjustable parameters can be suggested by theories; they needn't be pulled out of the air. The Ptolemaic and Copernican world pictures provide salient examples. Each begins with the simple idea that a certain object—the earth or the sun—is at the center. Articulating these ideas involves introducing models that contain adjustable parameters (Forster and Sober [1994]).

We illustrate this last point with a brief discussion of the case of Augustin Fresnel and the prediction of the bright spot. (For a more detailed discussion in the context of the problem of prediction vs accommodation, see Worrall [1989]; for a thorough historical treatment, see Buchwald [1989].) At the beginning of the nineteenth century, the dominant theory of the nature of light was the corpuscularian or 'emission' theory originally propounded by Newton; the wave theory of Huyghens had fewer adherents. The French Académie des Sciences announced a prize for the best treatise on the phenomenon of diffraction, to be awarded in 1819. Unlike the phenomenon of polarization, diffraction had received very little scientific attention. The wave theory of light could offer a rough qualitative account of diffraction, but neither theory offered a detailed quantitative treatment. It was the hope of the Académie's members that the prize would encourage bright young physicists to work on the problem of diffraction. In this respect it failed: the Académie received only two submissions. The prize was awarded to Fresnel's 'Mémoire sur la diffraction de la lumière' (Fresnel [1819]), which contained both an experimental part and a theoretical part. In the experimental part, Fresnel described a novel method for observing linear diffraction fringes and measuring the distances between them, and recorded the results of numerous observations. In the theoretical part, he described a new quantitative method, called the ‘method of integrals’, motivated by the wave conception of light, and showed that it fit the observational data to a high degree of accuracy. Poisson, a member of the prize commission, demonstrated that Fresnel’s theory entailed the occurrence of a bright spot in the center of the shadow cast by an illuminated disk; Arago, another member of the prize commission, conducted the experiment and confirmed the prediction.

If the miracle argument is correct, then we would have to conclude that the successful prediction of the bright spot (a previously unknown phenomenon) provided a particularly strong reason to believe in the underlying ontological commitments of Fresnel’s theory, namely the wave theory of light. Indeed, as the story is often told, that is exactly what happened. The historical record, however, tells another story. In the report of the prize committee, read aloud to the members of the Académie, we have a record of the reaction to Fresnel’s memoir of five of the era’s greatest physicists. In addition to Arago and Poisson, the committee included Biot, Gay-Lussac and Pierre Simon de Laplace, by far the most influential figure in French physics. The report was written by Arago (Arago [1819]), Fresnel’s strongest supporter on the committee, and no doubt represents a composite of the reactions of commissioners. The report is striking for what it leaves unsaid. First, it omits any reference to the wave theory of light, presenting only an ontologically sanitized version of Fresnel’s mathematical treatment. This suggests that the members of the commission were not at all persuaded of the truth of the underlying picture of the nature of light. Indeed, of the three members of the commission originally sympathetic to the emission theory of light, Biot remained loyal until converting some time in the 1830s, Poisson *may possibly* have undergone a conversion around the same time, although certainly no earlier, and Laplace died in 1827 without ever converting. Second, there is almost no mention of the much-vaunted bright spot. Two sentences near the very end of the long report describe the prediction and note that it was confirmed. There is no commentary indicating that this piece of evidence was of particular import.

So what was it about Fresnel’s memoir that so impressed the prize commission? First, several pages of the report are dedicated to describing and praising the ingenious new technique for observing diffraction fringes. The report also relates in detail the extreme quantitative accuracy of Fresnel’s method of integrals. Finally, and most interestingly from our perspective, the commissioners were impressed that Fresnel’s method required that only one parameter (corresponding to the wavelength of the light used in the experiment) needs to be adjusted to fit the data. Thus, the leading physicists of the day were particularly impressed that Fresnel was able to achieve excellent fit with the existing linear diffraction data using a one-parameter model.

Our account makes sense of the response of the commissioners. Since Fresnel was able to achieve such an excellent fit with the extant linear diffraction data using a model with only one free parameter, he could not have been overfitting that data. Since the main advantage afforded by the successful prediction of novel data is protection against overfitting, there was no reason to accord special status to the novel prediction in this case. Moreover, at least three of the commissioners were able to welcome an account of diffraction that could be expected to be predictively accurate while withholding belief in the entities postulated by that theory—waves in an optical medium.

10 Concluding comments

Annie found a model by accommodating the entire data set D (where $D = D_1 \cup D_2$). Penny found a model by accommodating D_1, and then used that model to accurately predict D_2. This simple story may seem to show that Penny's theory is better than Annie's, but that conclusion is premature. *How* did Annie accommodate D and *how* did Penny accommodate D_1? And *how accurately* did Penny's theory predict D_2? The narrative about Annie and Penny leaves these details unspecified.

The following two-by-two table summarizes some of our main conclusions (Table 1). Accommodating data simply means *using* them to formulate a theory. Our table contrasts two strategies of accommodation—finding a theory that fits the data perfectly and finding a theory that balances the conflicting *desiderata* of fit-to-data and simplicity with the goal of maximizing predictive accuracy. The table also contrasts two data sets that might be accommodated—the whole data set D, or part of it, D_1. We cite *AIC* as our method for balancing, but, as noted earlier, we use *AIC* in this way only for illustrative purposes; perhaps there are inference problems in which some other method would be preferable.

How might the models obtained by these four possible procedures be compared? We have suggested that historical information about the methods used is relevant only in so far as it elucidates the logical properties (including relational properties) of the models obtained. Accommodation by maximizing

Table 1

		Accommodate D		Accommodate D_1 and predict D_2
How to accommodate data?	Maximize Fit	I	$<$	II
		$\wedge$		$\wedge$
	Use *AIC*	III	$>$	IV

fit relative to the entire data set *D* (strategy I) is bad because it tends to yield models that overfit the data. But there are two ways out of this difficulty, not just one. You can accommodate part of the data and then, with luck, successfully predict the remaining data. This is strategy II. Alternatively, you can follow strategy III and accommodate all of the data, but do so by using *AIC*. We hope that by now it is intuitive that both II and III are better strategies than I. But there are further questions that need to be addressed. How accurate must the predictions be for the model obtained by strategy II to be better than the model obtained by strategy I? And how can strategy II make sense at all, if it violates the Principle of Total Evidence? How is strategy IV related to the other three? And what criterion for evaluating models can be applied to the four models obtained by these four different procedures?

This last question is the key to answering the others. The analysis of Section **6** shows how *AIC* can be used to compare the models generated by the four strategies shown in the table. We know that the model recommended by strategy III will have the highest estimated predictive accuracy, relative to the entire data set *D*. This means that, in expectation, it can't be a worse model to use than the ones obtained by any of the other strategies. In particular, III is superior to I because the former balances simplicity against goodness-of-fit so as to avoid overfitting. The same point applies to the question of how strategies II and IV are related. Regardless of whether you're going to accommodate *D* or only D_1, you're better off using *AIC* than merely finding a model that fits the data perfectly. In comparing strategies III and IV, the Principle of Total Evidence asserts that we should find the best model by considering *all* the data available. For this reason, the model obtained by strategy III can't be worse, in expectation, than the model obtained by strategy IV.

Returning to strategies I and II, we note that there is certainly no rock-hard guarantee that strategy II will fare better than strategy I. When predicting new data, one is always at the mercy of mother nature. We argued in Section **6**, however, that strategy II will fare better than strategy I so long as (i) the data set D_2 predicted by strategy II is sufficiently large, and (ii) the predictions are sufficiently accurate, as measured by log-likelihood.

Notice that the ordering described in the Table has strategy I in a valley. Moving clockwise, we ascend to a single peak, which is the use of strategy III. Prediction is better than accommodation in the first row—when your method of accommodation is to find a model that fits the data perfectly. However, when you're using *AIC*, you are better off accommodating the more inclusive data set. There is, however, a feature of our account that is not made transparent by our table. The columns and rows are not independent of one another. Given a sufficiently good fit with data, the information that the entire data set was accommodated provides evidence that a model was chosen on the basis of maximal fit, whereas the information that the data D_2 were successfully predicted provides evidence that simplicity was balanced against goodness-of-fit. Thus there is an evidential correlation between the first row and the first column, and between the second row and the second column.

Acknowledgements

We thank Jed Buchwald, Mark Colyvan, Ellery Eells, Malcolm Forster, Mathias Frisch, Alan Hájek, Matthias Hild, Brian Keeley, Peter Lipton, Cornelius Menke, Jim Woodward, John Worrall and an anonymous referee for useful discussion.

Christopher Hitchcock
Division of Humanities and Social Sciences
California Institute of Technology
Pasadena CA 91125
USA
cricky@caltech.edu
Elliott Sober
Department of Philosophy
Stanford University
California CA94305
USA
esober@stanford.edu

References

Akaike, H. [1973]: ‘Information Theory as an Extension of the Maximum Likelihood Principle’, in B. Petrov and F. Csaki (*eds*), *Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267–81.

Arago, F. [1819]: ‘Rapport fait par M. Arago à l’Académie des Sciences, au nom de la Commission qui avait été chargée d’examiner les mémoires envoyés au concours pour le Prix de la diffraction’, *Annales de chimie et de physique*, **XI**. Reprinted in Senarmont *et al*. (1866), pp. 229–46.

Buchwald, J. [1989]: *The Rise of the Wave Theory of Light: Optical Theory and Experiment in the Early Nineteenth Century*, Chicago: University of Chicago Press.

Burnham, K. and Anderson, D. [1998]: *Model Selection and Inference: A Practical Information-Theoretic Approach*, New York: Springer.

Earman, J. and Janssen, M. [1993]: ‘Einstein’s Explanation of the Motion of Mercury’s Perihelion’, in J. Earman, M. Janssen and J. D. Norton (*eds*), *The Attraction of Gravitation: New Studies in the History of General Relativity*, Boston: Birkhäuser, pp. 129–72.

Edwards, A. [1972]: *Likelihood*, Cambridge: Cambridge University Press.

Forster, M. [1999]: 'Model Selection in Science—The Problem of Language Variance', *British Journal for the Philosophy of Science*, **50**, pp. 83–102.

Forster, M. [2000]: 'Key Concepts in Model Selection: Performance and Generalizability', *Journal of Mathematical Psychology*, **44**, pp. 205–31.

Forster, M. [2002]: 'Predictive Accuracy as an Achievable Goal in Science', *Philosophy of Science*, **69**, pp. S124–S134.

Forster, M. and Sober, E. [1994]: 'How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions', *British Journal for the Philosophy of Science*, **45**, pp. 1–36.

Fresnel, A. [1819]: 'Mémoire sur la diffraction de la lumière'. Reprinted in Senarmont *et al.* [1866], pp. 247–382.

Glymour, C. [1980]: *Theory and Evidence*, Princeton, NJ: Princeton University Press.

Goodman, N. [1955]: *Fact, Fiction and Forecast*, Indianapolis, IN: Bobbs Merrill.

Hempel, C. G. [1945]: 'Studies in the Logic of Confirmation', *Mind*, **54**, pp. 1–26, 97–121. Reprinted in *Aspects of Scientific Explanation*, New York, NY: Free Press, 1965.

Hempel, C. G. [1966]: *Philosophy of Natural Science*, Englewood Cliffs: Prentice-Hall.

Horwich, P. [1982]: *Probability and Evidence*, Cambridge: Cambridge University Press.

Howson, C. [1989]: 'Accommodation, Prediction, and Bayesian Confirmation Theory', in A. Fine and J. Leplin (*eds*), *PSA 1988, Vol. 2*, East Lansing, MI: Philosophy of Science Association, pp. 381–92.

Howson, C. [1990]: 'Fitting your Theory to the Facts: Probably Not Such a Bad Thing After All', in W. Savage (*ed.*), 1990, *Minnesota Studies in the Philosophy of Science*, Vol. 14: *Scientific Theories*, Minneapolis, MI: University of Minnesota Press, pp. 222–24.

Lange, M. [2001]: 'The Apparent Superiority of Prediction to Accommodation as a Side Effect: A Reply to Maher', *British Journal for the Philosophy of Science*, **52**, pp. 575–88.

Lakatos, I. [1968]: 'Changes in the Problem of Inductive Logic', in Lakatos (*ed.*), *The Problem of Inductive Logic*, Amsterdam: North-Holland, pp. 315–417.

Leplin, J. [1997]: *A Novel Defense of Scientific Realism*, Oxford: Oxford University Press.

Lipton, P. [1991]: *Inference to the Best Explanation*, London and New York: Routledge.

Maher, P. [1988]: 'Prediction, Accommodation, and the Logic of Discovery', in A. Fine and J. Leplin (*eds*), *PSA 1988*, Vol. I, East Lansing: Philosophy of Science Association, pp. 273–85.

Maher, P. [1990]: 'How Prediction Enhances Confirmation', in J. M. Dunn and A. Gupta (*eds*), *Truth or Consequences: Essays in Honor of Nuel Belnap*, Dordrecht: Kluwer, pp. 327–43.

Mayo, D. [1991]: 'Novel Evidence and Severe Tests', *Philosophy of Science*, **58**, pp. 523–52.

Mayo, D. [1996]: *Error and the Growth of Experimental Knowledge*, Chicago, IL: University of Chicago Press.

McQuarrie, A. and Tsai, C. [1998]: *Regression and Time Series Model Selection*, Singapore: World Scientific.

Mill, J. S. [1843]: *A System of Logic*, London: George Routledge and Sons.

Musgrave, A. [1974]: ‘Logical Versus Historical Theories of Confirmation’, *British Journal for the Philosophy of Science*, **25**, pp. 1–23.

Popper, K. [1959]: *The Logic of Scientific Discovery*, London: Hutchinson.

Popper, K. [1962]: *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York, NY: Basic Books.

Putnam, H. [1975]: *Mathematics, Matter and Method*, Cambridge: Cambridge University Press.

Royall, R. [1997]: *Statistical Evidence: A Likelihood Paradigm*, Boca Raton, FL: Chapman and Hall.

Savage, W. (*ed.*) [1990]: *Minnesota Studies in the Philosophy of Science*, Vol. 14: *Scientific Theories*, Minneapolis: University of Minnesota Press.

Schwarz, G. [1978]: ‘Estimating the Dimension of a Model’, *Annals of Statistics*, **6**, pp. 461–5.

Senarmont, H. de, Verdet, E. and Fresnel, L. (*eds*) [1866]: *Œuvres complètes d'Augustin Fresnel*, Vol. 1, Paris: Imprimerie Impériale.

Sober, E. [1990]: ‘Contrastive Empiricism’, in W. Savage (*ed.*), *Minnesota Studies in the Philosophy of Science*, Vol. 14: *Scientific Theories*, Minneapolis, MI: University of Minnesota Press, pp. 392–412. Reprinted in *From a Biological Point of View*, Cambridge: Cambridge University Press, 1994.

Sober, E. [1998]: ‘Instrumentalism Revisited’, *Critica*, **31**, pp. 3–38.

Sober, E. [2002]: ‘Instrumentalism, Parsimony, and the Akaike Framework,’ *Philosophy of Science*, **69**, pp. S112–S123.

Sober, E. [2003]: ‘Two Uses of Unification’, in F. Stadler (*ed.*), *The Vienna Circle and Logical Empiricism*, Vienna Circle Institute Yearbook 2002, Kluwer, pp. 205–16.

Stone, M. [1974]: ‘Cross-Validatory Choice and Assessment of Statistical Predictions (with Discussion)’, *Journal of the Royal Statistical Society* B, **36**, pp. 111–47.

Stone, M. [1977]: ‘An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion’, *Journal of the Royal Statistical Society* B, **39**, pp. 44–7.

Whewell, W. [1840]: *History of the Inductive Sciences*, London: John W. Parker and Son.

Worrall, J. [1985]: ‘Scientific Discovery and Theory Confirmation’, in J. Pitt (*ed.*), *Change and Progress in Modern Science*, Dordrecht: Reidel, pp. 301–32.

Worrall, J. [1989]: ‘Fresnel, Poisson, and the White Spot: The Role of Successful Prediction in the Acceptance of Scientific Theories’, in D. Gooding, T. Pinch and S. Schaffer (*eds.*), *The Uses of Experiment: Studies in the Natural Sciences*, Cambridge: Cambridge University Press, pp. 135–57.

Zahar, E. [1973]: ‘Why Did Einstein's Programme Supersede Lorentz's, Parts 1 and 2’, *The British Journal for the Philosophy of Science*, **24**, pp. 95–125, 223–62.