Predictive Accuracy as an Achievable Goal of Science

Malcolm R. Forster[†][‡]
University of Wisconsin-Madison

What has science actually achieved? A theory of achievement should (1) define what has been achieved, (2) describe the means or methods used in science, and (3) explain how such methods lead to such achievements. Predictive accuracy is one truth-related achievement of science, and there is an explanation of why common scientific practices (of trading off simplicity and fit) tend to increase predictive accuracy. Akaike's explanation for the success of AIC is limited to interpolative predictive accuracy. But therein lies the strength of the general framework, for it also provides a clear formulation of many open problems of research.

1. The Problem of Scientific Achievement. I wish that science could obtain the truth, the whole truth, and nothing but the truth. It is one thing to postulate a goal for science and quite another to say that science is capable of achieving that goal. Philosophy of science should focus more on what science is capable of achieving, and not what philosophers or scientists desire on its behalf.

Realists say *the* truth about the world is the *ultimate* goal of science, but most realists also agree that every theory in the history of science is false. Popper 1963 clearly understood the difference between goals that are unreached and the actual truth-related achievement of science, which he defined in terms of verisimilitude (closeness-to-the-truth). Unfortunately, Popper's definition did not work, and there are many ways to define verisimilitude (see Niiniluoto 1998 for a recent survey).

Van Fraassen (1980) complains that Truth with a capital 'T' is an un-

[†]Send requests for reprints to the author, Department of Philosophy, 5185 Helen C. White Hall, 600 North Park Street, Madison, WI 53706; e-mail: mforster@facstaff.wisc.edu; homepage http://philosophy.wisc.edu/forster/.

[‡]I wish to thank Greg Mikkelson, Elliott Sober, and the participants at the PSA symposium for the helpful questions and comments.

Philosophy of Science, 69 (September 2002) pp. S124–S134. 0031-8248/2002/69supp-0012$10

realizable goal of science. But in its place, he says that science aims at empirical adequacy, which he defined to hold of a theory iff *all* of its observable consequences are true—past, present, *and future.* Yet it is equally clear that no scientific theory to date is *perfectly* accurate in all of its predictions. So, empirical adequacy (in this simple sense) is not an achievement of science either.

Kuhn (1970) eschews all talk of truth and verisimilitude as the goal of science and changes the subject by defining scientific progress (achievements) in terms of the consensus of the scientific community.

For any philosophy of science, it is useful to ask three questions:

1. What goal, or goals, can be *achieved* by science?
2. What possible or actual means, method, or criterion, can achieve the goal?
3. What *explanation* is provided of how the means tends to achieve the goal? Is there any account of the means-goal connection?

When applied to Popper, Kuhn, and van Fraassen (1980), only Popper addresses the problem of truth-related achievements and, at best, his theory is incomplete.

The aim of this essay is to describe an alternative approach, which not only defines 'truth-related' achievements of science precisely, but also provides a rudimentary explanation of why standard methods in science may lead to such achievements. Here is the theory in summary:

1. Predictive accuracy (Forster and Sober 1994) is a truth-related achievement of science.
2. Standard methods of theory selection can be seen as approximating a certain tradeoff between simplicity and fit (referred to as AIC) (see Sober 2002 for a brief description).
3. A theorem in mathematical statistics called Akaike's theorem (Akaike 1973, 1994) shows that the application of AIC can reduce overfitting error, which then explains (assuming that the conditions of the theorem are approximately satisfied) why hypotheses selected in this way tend to have greater predictive accuracy.

Predictive accuracy is defined in terms of the expected log-likelihood of re-sampled data. Like truth, it is not something that we can 'see'. Yet Akaike's theorem explains why AIC and predictive accuracy are connected. Akaike's theory has been explained, criticized, and defended many times. I plan to openly recognize the limitations of Akaike's theorem (and to correct some misconceptions like the alleged inconsistency of AIC), and finally demonstrate that the general framework is helpful in formulating and analyzing outstanding problems.

2. Many Predictive Accuracies. The first point is this: There is nothing that precludes the possibility that there are many truth-related goals that are simultaneously achieved in science. This even applies to the realist goal of the 'whole truth' about the world. The achievability of this goal does not imply that predictive accuracy is unachievable. So, unlike van Fraassen's constructive empiricism, the theory of predictive accuracy is compatible with scientific realism.

Not only does the goal of predictive accuracy allow the achievement of other goals, but predictive accuracy is not a single goal in the first place. If one considers its definition carefully, for example in the context of curve-fitting, it is clear that predictive accuracy is always relative to a *domain of prediction.* Suppose we want to predict the volume of air trapped in a J-shaped tube in the left (quantity y) on the basis of the height of the column of mercury on the right (quantity x). We may pose the prediction problem for x-values within the range of values tested by Boyle (ca. 1660), or we could predict the value of y for higher values of x. Or we could consider the prediction of volumes with *negative* values of x (i.e. pressures less than atmospheric pressure). (Boyle did this experiment as well—see Shamos 1959.) The domains of prediction in these three experiments are different, so the goal of predictive accuracy within each domain is different even if degrees of predictive accuracy are roughly the same.

This is philosophically important. For example, Kruse (1997) exploits the distinction between wider and narrower domains of prediction in discussing the age-old problem of why varied evidence is more valuable than unvaried evidence (see also Kruse 1999, 2000).

The main point of this essay is to argue for the importance of another distinction: *Interpolative* predictive accuracy is defined roughly as predictive accuracy within the same domain of prediction as that from which the observed data were sampled. *Extrapolative* predictive accuracy is predictive accuracy in a domain outside of the interpolative domain (the interesting case being when it is *entirely* outside).

More exactly, suppose that we define a *domain of prediction* as a probability distribution over values of the variable x. It may be that the distribution picks out a discrete set of x-values and gives them equal weight, or the distribution may be an interval distribution of a range of x-values, or it may be a Gaussian (normal) distribution with a specified mean value and variance. When the range of the distribution lies with the range of x-values from which the data in hand was sampled, then it is an interpolative domain, and when it is outside, then it is an extrapolative domain. This definition does not draw a sharp dividing line in all cases, but there are cases in which the difference is clear-cut and useful.

Akaike designed AIC as a method for estimating interpolative predictive accuracy. It is therefore an open question whether it succeeds as a means to the goal of extrapolative predictive accuracy as well. Later in this essay, I shall present an argument to the conclusion that it does not. While this imposes a limit on the effectiveness of AIC, it is a good illustration of the usefulness of the general framework.

3. Models and Rules for Their Selection. Let a *model* be a set of equations, which has adjustable parameters, plus an error distribution assigned to the 'dependent' variable y (the quantity to be predicted). For example, Boyle's Law in the form Volume × Pressure = Constant + U is a model in this sense if Constant is an adjustable parameter and U is Gaussian error term that has mean zero and an adjustable variance. Models are *families* of hypotheses that have precise likelihoods, in the sense that given the particular x-values measured, each hypothesis assigns a precise probability, or probability density, to the observed y values. The adjustable parameters are estimated from the observed data by finding the best-fitting version of the model (defined as the one that has the highest likelihood, i.e., makes the observations most probable). This maximum likelihood member of the family (sometimes called the *fitted model*) will assign a well defined probability value to any *new* set of data. Therefore, a fitted model has a well defined predictive accuracy in any well defined domain of prediction independent of whether we know its value or not. Unfitted models do not have well defined predictive accuracies in the sense defined.

AIC (Akaike Information Criterion) not only provides a way of estimating the predictive accuracy of fitted models, but it also defines a rule for selecting the best *model.*

AIC Rule: Select the fitted model that that has the greatest AIC score,

$$(1/N)\ \text{Maximum log likelihood} - k/N,$$

where N is the number of data and k is the number of adjustable parameters.

The AIC score written here differs from Akaike's AIC by a constant factor and is equivalent to it in the sense that a constant factor makes no difference to the AIC ordering of the rival models. Notice that everything mentioned in the rule is directly accessible to us. The maximum log-likelihood is the maximum probability (or probability density in the case of continuous quantities) of the *observed* data given particular members of the model. N is the number of data, and k is the number of adjustable parameters in the set of equations.

BIC (Bayesian Information Criterion) is derived on the basis of an idea that originally appears in Rosenkrantz 1977. The idea is that the likelihood of a model is the average likelihood of its members. Most members of the model fit the model very poorly and therefore the larger and more complex the model, the more the average is watered down from its maximum value. Schwarz (1978) went one step beyond Rosenkrantz in actually deriving a approximate formula to quantify the effect. There are many objections to the derivation (Forster and Sober 1994; Wasserman 2000) and the viability of the idea behind it (Forster and Sober 1994; Forster 2000). However, within the Akaike framework, any rule can be considered as the means to any goal, regardless of the intent of its authors or the soundness of its derivation.

BIC Rule: Select the fitted model that has the greatest BIC score

$$(1/N)\text{ Maximum log likelihood } - (\log N/2)\, k/N,$$

where N is the number of data and k is the number of adjustable parameters.

Notice that both of these rules uses the same first term. When the data are independent, according to the model, the log-likelihood is the sum of the log-likelihoods of each datum. Therefore the log-likelihood increases proportionally to N, so the first term can be thought of as constant for all values of N.[1] On the other hand, the second terms in both criteria tend to zero as $N \rightarrow \infty$ (this will be important later). For intermediate values of N, BIC has a greater tendency to select simpler models.

Cross validation methods divide the N data into a *calibration* set and a *test* set (Browne 2000). The model is then fitted to the calibration set and then scored by its fit with the test set. There is no need to use simplicity. The cross-validation score is a direct test of predictive accuracy. This method is widely used in learning algorithms in neural networks and machine learning. However, statisticians generally worry about its inefficient use of the data and the arbitrary division of the data. So they tend to favor a version of cross validation known as the leave-one-out method. In this method, each of the N data takes its turn at being the test set, and the N scores are then averaged. The rule selects the model with the highest average score.

Stone (1977) proves that leave-one-out cross validation is asymptotically equivalent to AIC. The reference here to ‘asymptotic’ refers to the limit for large N, which is also a sufficient condition for Akaike’s theorem to be true (under some weak regularity conditions). Stone’s theorem therefore independently verifies that AIC estimates interpolative predictive accuracy, since leave-one-out cross validation is a direct estimate of this quantity. In contrast, ordinary cross validation, with a judicious choice of

[1] This is an approximate claim. Its truth depends on what is held fixed. Clearly, with smaller N, overfitting will increase the per datum log-likelihood, so the expected per datum log-likelihood will decrease slightly as N increases, other things being equal. But these details do not affect the point to be made.

the calibration and test sets, is a direct estimate of extrapolative predictive accuracy. I will return to this point later.

Classical Neyman-Pearson hypothesis testing is also connected to AIC in an interesting way, in at least some examples. Consider a test between a simple hypothesis $\theta = 0$ and a composite hypothesis $\theta \neq 0$, where the distribution of θ is normal (Gaussian). The null hypothesis is $\theta = 0$, and $\theta \neq 0$ is a family of alternative hypotheses. θ could be the (unknown) propensity of a coin to land heads and the test statistic (the measured quantity) could be the frequency of heads in N tosses. θ then determines a distribution for the test statistic (the sample mean), which is approximately normal for a large number of tosses. The composite hypothesis will fit the data at least as well as the simple hypothesis because the composite hypothesis contains hypotheses that are arbitrarily close to the null hypothesis. Now, note that a Neyman-Pearson test will reject the null hypothesis only if the best case of the composite hypothesis does 'significantly' better than the null hypothesis. *Therefore Neyman-Pearson tests trade off fit against simplicity,* even though they were not designed to do so.

In the same example, AIC is equivalent to a Neyman-Pearson test with a rejection area of 15.73% (Forster 2000, 212). The conventional Neyman-Pearson rejection area of 5% makes it harder to reject the simple hypothesis. *Therefore,* a conventional Neyman-Pearson test gives a *greater* bias toward simple hypotheses *in this example.* Neyman-Pearson testing with a 5% rejection area is between AIC and BIC in the weight it gives to simplicity.

The last few paragraphs show that other methods of model selection may be equivalent in the sense of selecting the same model within a large class of examples, even though they are derived from different premises. However, the equivalence is sufficient to show that they may function as means to the same goals in that class of examples. The *intention* of the designers of a method is irrelevant to the kind of questions asked within the Akaike framework. Certainly, we may *ask* whether a criterion serves the purpose for which it was designed. But we may also ask whether it is an effective means to a different end. The answers to these questions are non-trivial. It may turn out that in some circumstances, the BIC rule, or Neyman-Pearson testing, is a more effective means of maximizing interpolative predictive accuracy because they give a greater weight to simplicity. And in other circumstances, the opposite may be true (Forster 2000). Moreover, the outcome of such an investigation does not answer questions about how the criteria compare in achieving other goals. That is why the three steps in Akaike's methodology are so important.

4. The Alleged Inconsistency of AIC. It is a part of the folklore of statistics that AIC is not statistically consistent, where an estimator is *consistent* if and only if it converges to its goal as the number of data $N \to \infty$. This is a good example of the muddle-headedness that can result from paying insufficient attention to goals. There is a sense in which AIC is inconsistent, but a full treatment of the issue shows that it is not inconsistent in the sense that matters.[2] Here are specifics of the case.

Consider a hierarchy of nested models. For example, consider the hierarchy of all n-degree polynomials, beginning with straight lines (LIN) and parabolas (PAR) at the lower end and heading toward polynomials with terms containing x to the power of n, for high values of n, at the other end. The models are nested because a lower degree polynomial is a special case of a higher degree polynomial when the coefficients of higher degree terms are constrained to be 0. Let the true curve first appear in the model with k^* adjustable parameters. AIC is inconsistent in the sense that the model selected by AIC overshoots k^* as $N \to \infty$. BIC is consistent because it converges to k^* in the same limit. However, this property of AIC and BIC is only important if one of two conditions is met: (1) The estimation of k^* is an important goal in and of itself, or (2) a failure to converge on k^* is symptomatic of a failure to converge on the true hypothesis in the model k^* (since every model has a unique k, we denote the model by k). I shall argue that neither of these conditions is met.

The number k^* is the number of adjustable parameters, or dimension, of the model in which the true hypothesis *first* appears in the hierarchy. But the true hypothesis is also a member of all the models higher in the hierarchy, since the models are nested. To make sure that this fact is not lost in the notation, let h^* denote the true hypothesis. Then h^* is in k^* and in all models such that $k > k^*$. Therefore, there is no sense in which k^* is *the* true dimension of h^*. After all, the equation for h^* does not have adjust*able* parameters, only adjust*ed* parameters. To put the point another way, we could define a quite different hierarchy of nested models in which h^* first appears in a model of dimension different from k^*. Hence, k^* is an artifact of the representation, and not something that one should be interested in estimating.

Nor does condition (2) hold, since h^* also appears in all models in the hierarchy higher than k^*, the fact that AIC overshoots k^* does not exclude the possibility that AIC converges on h^* as $N \to \infty$. It only shows that it may converge on h^* *from a different direction.* Of course, this does not prove that AIC converges on h^*. I have only argued that the fact that AIC overshoots k^* does not prove that it does not.

To see that, in fact, AIC does converge on h^*, it is sufficient to note that even a maximum likelihood method (which gives no weight to sim-

[2] And there is also a question about whether consistency is necessary for an estimator to be a good one. See Sober (1988) for an argument that it is not necessary.

plicity, and immediately selects the most complex model in the hierarchy) will converge on h^*. This is proven by a well known theorem in statistics to the effect that the maximum likelihood estimation (MLE) of parameters will converge on their true value as $N \rightarrow \infty$ in a truncated hierarchy. So, truncate the hierarchy at a value of k higher than any reached by AIC. If MLE converges on h^*, then so will AIC. We have already observed that the simplicity term gets washed out by the maximum likelihood term in both AIC and BIC, which is to say that the criteria become indistinguishable from MLE in the limit. Of course, one cannot use this equivalence to prove that they select the same value of k in the large sample limit, because they do not. That is why the argument must also appeal to the convergence theorem about MLE. Finally, I refer to simulations in Forster 2000, which confirm that all of this is correct.

Therefore, the limited sense in which AIC is inconsistent, and BIC is not, is not an argument in favor of BIC *with respect to the goal of predictive accuracy.* My main purpose is not to defend AIC over BIC, but to illustrate how the clarity of thought demanded by the predictive accuracy framework plays an important role in resolving the issue.

5. Three Open Problems. To prove my sincerity, I shall mention three unresolved problems for AIC, although I hasten to add that they pose equal problems for BIC. The first problem arises from the phenomenon that Zucchini (2000) refers to as *selection bias:* If one begins with a very large collection of rival models, then we can be fairly sure that the winning model will have an accidentally high maximum likelihood term. It is a kind of second order overfitting effect. The selection bias can be expected to be less if we begin with a small set of rival models. But then we have the problem of how to select the few models that go into this set. Perhaps this is one of the reasons that theories are so prevalent in science, since a background theory will constrain the set of models under consideration. After selecting the best model from rival theories, the two winners may play off against each other. This is one idea, and there are others (Wasserman 2000 suggests that the Bayesian idea of model averaging may address this problem). However, it is a fairly recent area of research and I believe that the issue is unresolved.

The second issue is: What happens when the assumptions of Akaike's theorem are false? Akaike's theorem assumes that the maximum likelihood hypothesis, when represented as a point in parameter space, behaves as a multivariate normal random variable under data re-sampling. This is a condition that holds asymptotically under weak regularity conditions, and entails all the nice theorems about maximum likelihood estimation (Cramér 1946a, chaps. 32 and 33, and Cramér 1946b). As an empirical law, the normality condition is also related to what became known in the eighteenth century as the *law of errors* (which states that the error of parameter estimation is Gaussian—see Porter 1986; also Stewart 1989, chap. 3).

The question about when these assumptions hold, or otherwise, is unresolved in many important cases. For example, Kieseppä (1997) describes an artificial example, which he proves violates Akaike's theorem, and then concludes that AIC is not a good estimate of predictive accuracy in this case. He then argues by analogy that, because his example and the historically real example of predicting planetary motions both involve similar functional forms, that the normality condition does not apply to planetary astronomy. But arguments by analogy are notoriously unreliable, unless they are independently verified.

As Cramér says in the introductory preamble to his treatment of theorems (1946, 146, his emphasis), "Certain propositions of a mathematical theory may, however, be *tested by experience.*" It is therefore relevant to quote Cramér at length on the history of the law of errors:

> Gauss and Laplace were both led to the normal function in connection with their work on the theory of errors of observation. Laplace gave, moreover, the first (incomplete) statement of the general theorem studied under the name of the Central Limit Theorem . . . Under the influence of the great works of Gauss and Laplace, it was for a long time more or less regarded as an axiom that statistical distributions of all kinds would approach the normal distribution as an ideal limiting form, if only we could dispose of a sufficiently large number of sufficiently accurate observations. The deviation of any random variable from its mean was regarded as an "error", subject to the "law of errors" expressed by the normal distribution. Even if this view was definitely exaggerated and has had to be considerably modified, it is undeniable that, in a large number of important applications, we meet distributions which are at least approximately normal. Such is the case, e. g., with the distributions of errors of physical and *astronomical measurements,* a great number of demographical and biological distributions, etc. (Cramér 1946a, 231, my emphasis)

It was Gauss who analyzed the method of least squares and he introduced it as a method of inferring planetary motions. So, the law of errors has empirical support in planetary astronomy. This does not prove that Akaike's theorem applies to planetary astronomy, but it does suggest that one needs something stronger than an analogy with a contrived example to prove that the theorem does not apply at least approximately in this example.

Planetary astronomy is also an interesting example because it is clearly *extrapolative* predictive accuracy that is at stake. Astronomers fit their models to past data and then use the fitted models to predict future events.

It is also an example in which there is a rich supply of observational data, which has the following significance: The standard criteria, like AIC and BIC, have the property that the simplicity factor *k* makes a negligible difference when *N* is large. In that case, the criteria select the best fitting hypothesis. Therefore, complex models will be favored when *N* is large. As far as interpolation is concerned, I have already argued that this is a good consequence. However, for the purpose of *extrapolation,* there is reason to believe that it is a bad consequence (see also Muliak 2001). If this is right, then it follows that the standard criteria, like AIC and BIC, will fail to be an effective means to extrapolative predictive accuracy when the number of data is large. The argument (Busemeyer and Wang 2000) is summarized as:

1. All the standard criteria, like AIC and BIC, have the property that simplicity makes a negligible difference when *N* is large.
2. Complex models will be favored when *N* is large.
3. Complex models are good for interpolation but bad for extrapolation when *N* is large.

---

4. None of the standard criteria is a good means to extrapolative predictive accuracy when *N* is large.

Simulations performed by Forster (2000) confirm that premises 2 and 3 are correct in at least some examples.

Upon reflection, the B-W argument makes sensible recommendations. It tell us, for example, that we should not re-introduce epicycles into planetary astronomy now that we have copious data and fast computers, just because Fourier's theorem tells us that we could fit planetary trajectories better by this method than by using Einsteinian theory. The reason is that the predictive accuracy of our extrapolative predictions will not improve.

On the other hand, *prediction tests,* or *cross validation* with a directionality built in, as well as Whewell's *consilience of inductions* (1858!), do not use simplicity explicitly, so they are *immune* to the B-W argument. Indeed, the same computer simulations show that these criteria perform better at extrapolation. However, the effectiveness of such criteria in more general situations is an open research problem.

6. Conclusion. While AIC is a good means to achieving the goal of *interpolative* predictive accuracy, Akaike's real legacy is his general *framework,* for it requires us to make careful distinctions. It is only by carefully defining predictive accuracy that one notes the distinction between interpolative and extrapolative predictive accuracy, and only then does one think of asking whether they are achieved by the same methods or by different methods. The limitation of AIC to interpolation is a small setback for AIC and a major leap forward for the Akaike framework.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", in B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory.* Budapest: Akademiai Kiado, 267–81.

——— (1994), "Implications of the Informational Point of View on the Development of Statistical Science", in H. Bozdogan (ed.) *Engineering and Scientific Applications,* Vol. 3, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach.* Dordrecht: Kluwer, 27–38.

Browne, Michael (2000), "Cross-validation Methods", *Journal of Mathematical Psychology* 44: 108–132.

Busemeyer, J. R. and Yi-Min Wang (2000), "Model Comparisons and Model Selections Based on Generalization Test Methodology", *Journal of Mathematical Psychology* 44: 177–189.

Cramér H. (1946a), *Mathematical Methods of Statistics.* Princeton, N.J.: Princeton University Press.

——— (1946b), "A Contribution to the Theory of Statistical Estimation", *Skandinavisk Aktuarietidskrift* 29: 85–94.

Forster, Malcolm R. (2000), "Key Concepts in Model Selection: Performance and Generalizability", *Journal of Mathematical Psychology* 44: 205–231.

Forster, Malcolm R. and Elliott Sober (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**:** 1–35.

Kieseppä, I. A. (1997), "Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity", *British Journal for the Philosophy of Science* 48**:** 21–48.

Kruse, Michael (1997), "Variation and the Accuracy of Predictions", *British Journal for the Philosophy of Science* 48: 181–193.

——— (1999), "Beyond Bayes: Comments on Hellman", *Philosophy of Science* 66**:** 165–174.

——— (2000), "Invariance, Symmetry and Rationality", *Synthese* 122: 337–357.

Kuhn, Thomas (1970), *The Structure of Scientific Revolutions,* Second Edition. Chicago: University of Chicago Press.

Muliak, Stanley A. (2001), "The Curve-Fitting Problem: An Objectivist View", *Philosophy of Science* 68: 218–241.

Niiniluoto, Ilkka (1998), "Verisimilitude: The Third Period", *British Journal for the Philosophy of Science* 49: 1–29.

Popper, Karl (1968), *Conjectures and Refutations* : *The Growth of Scientific Knowledge.* New York: Basic Books.

Porter, Theodore (1986), *The Rise of Statistical Thinking 1820–1900.* Princeton, N.J.: Princeton University Press.

Rosenkrantz, Roger D. (1977), *Inference, Method, and Decision.* Dordrecht: D. Reidel.

Shamos, Morris H. (ed.) (1959), *Great Experiments in Physics: Firsthand Accounts from Galileo to Einstein.* New York: Dover Publications.

Sober, Elliott (1988), "Likelihood and Convergence", *Philosophy of Science* 55**:** 228–237.

——— (2002), "Instrumentalism, Parsimony, and the Akaike Framework" *Philosophy of Science* (forthcoming).

Stewart, Ian (1989), *Does God Play Dice? The Mathematics of Chaos.* Oxford: Basil Blackwell.

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: 44–47.

van Fraassen, Bas (1980), *The Scientific Image.* Oxford: Oxford University Press.

Wasserman, Larry (2000), "Bayesian Model Selection and Model Averaging", *Journal of Mathematical Psychology* 44: 92–107.

Whewell, William ([1858] 1967), *Novum Organon Renovatum.* Originally published as Part II of the 3rd edition of *The Philosophy of the Inductive Sciences,* London: Cass.

Zucchini, Walter (2000) "An Introduction to Model Selection", *Journal of Mathematical Psychology* 44: 41–61.