Instrumentalism, Parsimony, and the Akaike Framework

Elliott Sober[1]

Abstract: Akaike's *framework* for thinking about model selection in terms of the goal of predictive accuracy and his *criterion* for model selection have important philosophical implications. Scientists sometimes test models whose truth values they already know, and then sometimes choose models that they know full well are false. Instrumentalism helps explain this pervasive feature of scientific practice, and Akaike's framework provides instrumentalism with the epistemology it needs. Akaike's criterion for model selection also throws light on the role of parsimony considerations in hypothesis evaluation. I explain the basic ideas behind Akaike's framework and criterion; several biological examples, including the use of maximum likelihood methods in phylogenetic inference, are considered.

Philosophers of science usually agree that the point of testing theories – indeed, the point of doing science -- is to try to determine which theories are true. Of course, we all recognize that scientists never have access to all possible theories on a given subject; they are limited by the theories they have at hand. But given a set of competing theories, the point of theory assessment is to ascertain which of these competitors is one's best guess as to what the truth is. Bayesians tend to see things this way, so do scientists who use orthodox Neyman-Pearson methods, and likelihoodists tend to fall into this pattern as well. To be sure, there are deep differences among these outlooks. Bayesians assess which hypotheses are most probable, frequentists evaluate which hypotheses should be rejected, and likelihoodists say which hypotheses are best supported. But these assessments typically invoke the concept of truth; the question is which hypotheses are most probably *true*, or should be rejected as *false*, or are most likely to be *true*.

This obsession with truth also finds expression in the debate between realism and empiricism. Realism says that the goal of science is to discover which theories are true; empiricism maintains that the goal is to discover which theories are empirically adequate (Van Fraassen 1980). A theory is empirically adequate if what it says about observables is true. Realists think that theories should be assessed by considering the truth values of everything they say, while empiricists hold that theories should be assessed by considering the truth values of part of what they say. In both cases, truth is the property that matters.

An older tradition, now not much in evidence in these post-positivist times, holds that the point of science is to provide accurate predictions, not to tell us which theories are true. This is *instrumentalism*, stripped of the defective philosophy of language that led instrumentalists to deny that theories have truth values (Morgenbesser 1960). Ernest Nagel (1979) is often taken to have punctured the instrumentalist balloon with his suggestion that the difference between instrumentalism and realism is nonsubstantive; if true theories are the ones that maximize predictive accuracy, then the goal of seeking predictive accuracy and the goal of seeking truth come to the same thing.

With the demise of positivism and the ascendancy of realism, why even consider instrumentalism? The reason the case needs to be reopened has two parts. First, there are aspects of scientific practice that don't make sense on the model of science as the quest for truth. And second, there is an alternative framework for understanding scientific inference, one that is used increasingly by scientists themselves, which says that the goal of theory evaluation is to estimate predictive accuracy. It turns out in this framework that a true theory can be *less* predictively accurate than a false one. Nagel's suggestion that truth and predictive accuracy always coincide is not correct. In addition, this framework can handle a body of inference problems that no other inferential framework is able to address.

The simple but pervasive fact about scientific practice is that scientists often test hypotheses that they know full well are false and they often refuse to reject such hypotheses in the light of evidence. Consider, for example, the simple statistical problem of deciding whether two large populations of corn plants have the same mean heights. Where u_1 and u_2 are the two means, the two hypotheses to consider are

(Null) $u_1 = u_2$
(Diff) $u_1 \ldots u_2$.

Surely no scientist could or should believe that two such populations have exactly the same average heights. Yet, this and similar hypotheses are tested everyday and sometimes the conclusion is drawn that one should not reject the null hypothesis. Scientists must be crazy if this assessment concerns what is true. But if their goal is to assess which model is more predictively accurate, there may be method in this madness (Sober 1998, Forster 2000a).

My next example (for others, see Yoccuz 1991, Johnson 1995, and Burnham and Anderson 1998), also from biology, concerns hypotheses about a "molecular clock." Consider two lineages that stem from a common ancestor and connect to contemporary descendant species B and C. There are millions of nucleotides in the DNA of the organisms in these two lineages. Let b = the rate of nucleotide substitution in the lineage leading to B, and c = the rate of substitution in the lineage leading to C. The clock hypothesis is expressed by the first of the following two hypotheses (Felsenstein 1983):

(Constrained) b = c
(Unconstrained) b = c or b... c.

The constrained model entails the unconstrained model, but not conversely. I think we know, before any data are gathered, that the constrained model is almost certainly false and that the unconstrained model must be true; the latter is, after all, a tautology. Furthermore, we know that if we gather data, our observations will not dislodge this two-part verdict. Yet, scientists go to the trouble of gathering data, and when they run statistical tests, they sometimes decline to reject the clock hypothesis. The

puzzlement is why scientists bother to run these tests in the first place, if the goal is to discover which models are true. Does science, like poetry, demand the willing suspension of disbelief?

These peculiar practices start to make sense in the light of an inferential framework developed by the Japanese statistician H. Akaike (1973) and his school (Sakamoto *et al.* 1986) for thinking about how models are used to make predictions.[2] Models, first of all, are statements that contain adjustable parameters. They are disjunctions, often infinite disjunctions, over all the different parameter values that are consistent with the constraints the model specifies. If models are disjunctive in this way, how can they be used to make predictions? The answer is that one estimates the values of parameters by consulting data; one finds the parameter values that maximize the probability of the data -- i.e., the values that have *maximum likelihood*. In the case of (Null) and (Diff), suppose one samples from each population and finds that the mean height in the first sample is 62 inches and the mean height in the second is 64 inches. If so, the likeliest members of (Null) and (Diff) are:

L(Null) $u_1 = u_2 = 63$ inches
L(Diff) $u_1 = 62$ inches; $u_2 = 64$ inches.

Notice that L(Diff) fits the data better than L(Null) does. However, the goal is not to fit old data, but to predict new data. The question is how well L(Null) and L(Diff) will do in predicting new data drawn from the same two populations. It is perfectly possible that L(Null) will predict new data better than L(Diff), even though (Null) is false and (Diff) is true.

The concept of predictive accuracy describes a two-step process. One uses old data to find the most likely member of each model; then one uses those likeliest members to predict new data:

(Null)

! L(Null)

ú

Old Data New Data

ü

! L(Diff)

(Diff)

Imagine repeating this process again and again. The predictive accuracy of a model is its *average* performance in this iterated task. Predictive accuracy is a mathematical expectation.

Akaike not only articulated a *framework* in which predictive accuracy is the goal of inference; in addition, he provided a *methodology* for estimating a model's predictive accuracy.
Given the data at hand, how is one to estimate how well a model will do in predicting new data – data

that one does not yet have? Akaike's criterion for model selection is expressed by a theorem he proved. He was able to show that[3]

An unbiased estimate of the predictive accuracy of model M . log-Pr[Data * L(M)] - k.

The probability that L(M) confers on the data is relevant to assessing M's predictive accuracy, but it is not the only consideration. The other factor that matters is k, the number of adjustable parameters the model contains. Akaike's theorem imposes a penalty for complexity. In the examples we have considered, it is inevitable that L(Null) will have a lower likelihood than L(Diff) and that L(Constrained) will be less likely than L(Unconstrained). However, it also is true that (Null) is simpler than (Diff) and that (Constrained) is simpler than (Unconstrained). Likelihood and simplicity are in conflict; Akaike's theorem shows how each contributes to estimating a model's predictive accuracy. If two models fit the existing data about equally well, then the simpler model can be expected to do a better job predicting new data. For the more complex model to receive the higher AIC (Akaike information criterion) score, it must fit the data a lot better, not just modestly better. Akaike's theorem quantifies this trade-off – it describes how much of a gain in likelihood there must be to off-set a given loss in simplicity. Just as Akaike's framework breathes new life into instrumentalism, his theorem provides powerful insights into the relevance of parsimony considerations in many inference problems (Forster and Sober 1994).

Akaike's theorem is a theorem, so we should note the assumptions that go into its proof. First, in his definition of predictive accuracy, Akaike defines the distance between a fitted model and the truth by using the Kullback-Leibler distance. Second, he assumes that the new data will be drawn from the same underlying reality that generated the old (this has two parts – that the true function that connects independent to dependent variables is the same across data sets, and that the distribution that determines how the values of independent variables are selected is also the same); this might be termed a Humean "uniformity of nature" assumption (Forster and Sober 1994). And third, Akaike makes a normality assumption; roughly, this is the idea that repeated estimates of each parameter are normally distributed. In the model selection literature, there is discussion of other distance measures, such as mean-squared error and Kolmogorov's absolute difference measure (McQuarrie and Tsai 1998). It also turns out to matter whether one's data set is small or large. These are two reasons why model selection criteria other than Akaike's have attracted attention (see also Burnham and Anderson 1998). Proposed criteria differ in terms of the penalty imposed for complexity, but it is not controversial that high likelihood of the fitted model is good news and that complexity is bad.[4]

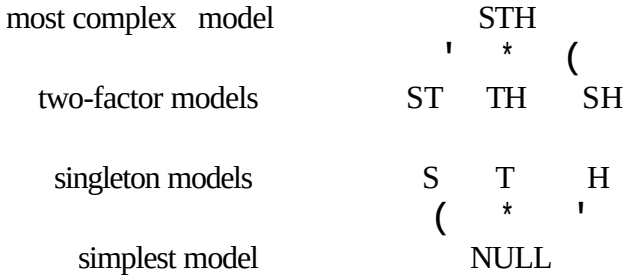
The uniformity of nature and normality assumptions that go into the proof of Akaike's theorem are empirical claims about the prediction problem at hand. This has important implications for the question of whether simplicity is a "super-empirical virtue." Although it is clear that simplicity is a separate consideration in model selection from fit to data, the justification provided by Akaike's theorem for using simplicity depends on empirical assumptions. Simplicity is therefore an empirical consideration. This is good news for empiricism, since empiricists have had a hard time reconciling their epistemology with the role that simplicity evidently plays in scientific inference, and sometimes have

gone so far as to claim that simplicity considerations are merely pragmatic (e.g., Van Fraassen 1980).

With this sketch of Akaike's criterion added to the previous sketch of the Akaike framework, we can fine-tune our claim concerning instrumentalism and realism. These philosophies are usually understood *globally*; the goal of inference is *always* to find theories that make accurate predictions, or the goal is *always* to find theories that are true. Akaike's framework and theorem show that each must be reformulated *locally*. The assessment of models containing *adjustable* parameters conforms to instrumentalism. But the assessment of fitted models, all of whose parameters have been *adjusted*, can be construed realistically. The data may lead us to judge that (Null) is a better predictor than (Diff), even though we know that (Null) is false and (Diff) is true,[5] but, if so, we also will judge that L(Null) is closer to the truth (in the sense of Kullback-Leibler distance) than L(Diff). The operative slogan is: *instrumentalism for models, realism for fitted models*.[6] Notice that the realism that pertains to fitted models does not mean that one regards them as true (surely one knows that they are not).

Akaike's criterion allows one to compare both nested and non-nested models. To explain what this means, I'll describe some of the models that Burnham and Anderson (1998, pp. 110-114) consider as explanations of data gathered by Schoener (1970) on resource utilization in two species of *Anolis* lizard in Jamaica. Schoener repeatedly inspected the sites occupied by different lizards in an area that had been cleared of trees and shrubs. Each time he observed a lizard perching, he noted which of two species (S) it belonged to, the height (H) of the perch (< or $ five feet), the perch's diameter (D) (< or $ 2 inches), the site's insolation (I) (whether it was sunny or shady), and the time of day (T) (early morning, midday, or late afternoon).

The most complex model that Burnham and Anderson consider says that an individual's probability of perching on a site may be influenced by S, H, D, I, and T, its being left to the data to say how much of a difference, if any, each makes. The simplest model says that an individual's probability of perching is not affected by any of these factors; (NULL) is the nihilistic model that nothing matters. In between are singleton models, two-factor models, three-factor models, and so on; each says that the variables cited may matter, and that the ones that go unmentioned do not. Let's consider just the following:



These models form a partial ordering, depicted by the lines connecting models at different levels. NULL is the logically strongest model; it entails all the others. That is, NULL is nested inside of S, S is

nested inside of ST, and so on. Simpler models can be obtained from the more complex models in which they are nested by setting parameters equal to zero.

Akaike's theorem allows us to compare all of these models. Nested models can be compared for their estimated predictive accuracy, and so can disjoint models, with the answer always depending on the data at hand. This isn't so for either Bayesianism or for Neyman-Pearson procedures. When Bayesians look at nested models, the verdict is pre-ordained; for example, since NULL entails T, NULL cannot be more probable than T, no matter what the data say. This means that Bayesians cannot represent the fact that scientists often evaluate logically stronger models as "better" than logically weaker models. They can't be better in Bayesian terms, because they can't be more probable (Popper 1959, Forster and Sober 1994). The standard Bayesian response is to "change the subject." Instead of comparing NULL with T,
they'll choose to compare NULL with T*; where T asserts that time of day *may* make a difference, T* says that time *does* make a difference. T* is not nested in NULL: they are disjoint. With the problem redefined in this way, there now is no logical prohibition against claiming that NULL is more probable than T*. Of course, two problems remain – how are the likelihoods of composite hypotheses to be assessed, and how is one to justify an assignment of prior probabilities (especially one that says that NULL is more probable *a priori* than T*)?

The limitation imposed by Neyman-Pearson procedures is different. The likelihood ratio test used in frequentist statistics allows nested models to be compared, but not models that are disjoint. The members of the set consisting of NULL, S, ST, and STH can all be compared, but S cannot be compared with T, nor with TH, for example. In this situation, frequentists often compare each singleton model with the null hypothesis, and then construct a multi-factor model that includes all and only the singleton factors that were able to "beat" the null hypothesis. This is an expedient procedure that has no mathematical rationale within frequentist philosophy. If STH beats each of S, T, and H, and each of these singleton models beats the null hypothesis, then it makes sense to embrace the STH model and reject the simpler alternatives. However, it is perfectly possible that NULL is rejected each time it is compared with the singleton models S, T, and H, but that one of these singleton models does *not* get rejected when it is compared with the three-factor model STH. In this case, what is one to do? Neyman-Pearson statistics provides no answer. This is one reason why scientists have embraced Akaike-style model selection procedures. Most scientists follow Neyman-Pearson methods when they can; however, when they want to compare non-nested models, they have had to find a different approach.

I now want to describe an area of research in evolutionary biology in which Akaike's ideas are just starting to be used. In the 1960's biologists began developing maximum likelihood methods for inferring phylogenetic relationships. Consider the simplest case, in which the inference problem involves three species (humans, chimpanzees, and gorillas, for example). Assuming that there is a common ancestor that unites all three and that the phylogeny is bifurcating, there are three possible trees. Biologists try to use the observed characteristics of these species to assess which tree is most

likely; the task is to say whether

Pr[Data * (HC)G] > Pr[Data * (H(CG)] , Pr[Data * (HG)C].

The problem, however, is that phylogenetic hypotheses are composite. The probability that a tree topology confers on the data depends on a model of the evolutionary process *and* on the values of the parameters in that model; that is, the likelihood of the hypothesis is an average:

Pr(Data * (HC)G) = $\Sigma_i \Sigma_v$ Pr[Data * (HC)G & Model-i & parameters in Model-i have values v]Pr[Model-i & parameters in Model-i have values v * (HC)G].

This poses a problem for the likelihood approach, since no one has the slightest idea how to evaluate the second product term on the right side of the equality.

If the goal of one's inference is to say which tree topology is best supported, then the process model and the parameters in that model are "nuisance parameters;" they affect the likelihood, but they are not what one wishes to infer. One way to deal with nuisance parameters is to "change the subject" – instead of trying to determine the *average* likelihood of a topology, given a model, one assesses its likelihood under the assumption that the parameters in the model have their *maximum* likelihood values. This expedient solution is not entirely satisfactory (Edwards 1972, Sober 1988, Royall 1997; Forster 1986, 1988 disagrees). But even granting this reformulation, two new problems arise. First, the evaluation of tree topologies depends on the process model used:

(*) Pr[Data * (HC)G & L(Model-1)] > Pr[Data * H(CG) & L(Model-1)]
Pr[Data * (HC)G & L(Model-2)] < Pr[Data * H(CG) & L(Model-2)].

Unfortunately, biologists who don't already know which phylogeny is correct are unlikely to know which process model is correct for the taxa and traits at hand (Felsenstein 1978, Sober 1988). Furthermore, the ability to discriminate among tree topologies tends to decline as more complex and realistic process models are employed (Lewis 1998, p. 139):

Pr[Data * (HC)G & L(Model-n)] . Pr[Data * H(CG) & L(Model-n)].

Although the likelihood of a topology goes up as more complex process models are employed, the likelihoods of different topologies come closer together. How depressing that greater realism about the evolutionary process should impair, rather than enhance, our ability to reconstruct phylogenies!

The solution to both these problems has been to use the frequentist methodology of likelihood ratio tests. Instead of passing automatically from simpler to more complex models, one asks whether the shift represents a *significant* improvement in fit. However, we now need to attend to the fact, noted earlier, that likelihood ratio tests are meaningful only for nested models. To explore the import of

this point, I should correct a bit of misleading notation. In (*), L(Model-1) appears on both sides of the inequality. The point that needs to be recognized is that the same process model, when conjoined with different tree topologies, often yields different maximum likelihood estimates of parameter values. It would be better to write L[(HC)G & Model-1] and L[H(CG) & Model-1] to make this point clear.

Process models, when separated from tree topologies, are partially ordered (Swofford *et al.* 1996, p. 434). For example, in the case of inferring phylogenies from DNA sequence data, the first process model to be explored was also the simplest – that of Jukes and Cantor (1969). This model assumes that all changes at a site have the same probability, that all sites in a lineage evolve independently and according to the same rules, and that a site in one lineage obeys the same rules as the same site in any other. The model therefore assumes that selection does not favor one nucleotide at a site over any other; this is a pure drift model in which the effective population sizes in different lineages are the same. Subsequent models have relaxed different assumptions in the Jukes-Cantor model in different ways. It isn't that the newer models assume the *opposite* of what the Jukes and Cantor model stipulates. Rather, these models leave this or that matter open, and let the data decide what the best settings of the parameters are (Lewis 1998).

Consider the following four conjunctions; each includes a tree topology and a process model fitted to that topology:

L[(HC)G & Model-2] L[H(CG) & Model-2]

L[(HC)G & Model-1] L[H(CG) & Model-1]

Likelihood ratio tests permit *vertical* comparisons, if (as I assume) Model-1 is nested in Model-2. However, it isn't so easy to determine, within frequentist statistics, how one should make *horizontal* comparisons (Swofford *et al.* 1996, p. 506). Given a single process model, how much difference in likelihood between two fitted topologies is needed for there to be a *significant* difference? But even more puzzling are *diagonal* comparisons. Likelihood ratio tests do not permit us to test L[(HC)G & Model-2] against L[H(CG) & Model-1]. However, this comparison and the others as well make perfect sense in the Akaike framework. The likelihood of each conjunction is relevant, but so too is the number of adjustable parameters. If it turns out that L[H(CG) & Model-1] has the highest AIC score among the conjunctions considered, there is no need to apologize for the fact that Model-1 is a process model that one knows is false.

In closing, I want to discuss one more example, just for fun. Some years ago, cognitive psychologists discussed the phenomenon of "hot hands" in sports. Everyone with even the most superficial familiarity with professional basketball believes that players occasionally have "hot hands." When players are hot, their chance of scoring improves, and team-mates try to feed the ball to them. However, a statistical analysis of scoring patterns in the NBA yielded the result that one cannot reject the null hypothesis that each player has a constant probability of scoring throughout the season (Gilovich

et al. 1985). Scientists who believe in statistics took this result seriously, but no one else did. The scientists concluded that belief in hot hands is a "cognitive illusion," while basketball mavens reacted to this statistical pronouncement with total incredulity. Placing this dispute in the Akaike framework allows it to make more sense. Scientists should not feel shy about admitting that the null hypothesis is false. The idea that players never waiver in their probabilities of scoring *is* preposterous. The point of doing statistics is not to see whether this silly hypothesis is true, but to see how good it is at predicting new data. Presumably, the truth about basketball players is very complex. Their scoring probabilities change as subtle responses to a large number of interacting causes. Given this complexity, players and coaches may make better predictions by relying on simplified models. Hot hands may be a reality, but trying to predict when players have hot hands may be a fool's errand.

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. 267-281.

Burnham, K. and Anderson, D. (1998): *Model Selection and Inference – a Practical Information-Theoretic Approach*. New York: Springer.

Edwards, A. (1972): *Likelihood*. Cambridge: Cambridge University Press.

Felsenstein, J. (1978): "Cases in which Parsimony and Compatibility Methods can be Positively Misleading." *Systematic Zoology* **27**: 401-410.

Felsenstein, J. (1983): "Statistical Inference of Phylogenies." *Journal of the Royal Statistical Society* A **146**: 246-272.

Forster, M. (1986): 'Statistical Covariance as a Measure of Phylogenetic Relationship.' *Cladistics* **2**: 297-317.

Forster, M. (1988): "Sober's Principle of Common Cause and the Problem of Comparing Incomplete Hypotheses." *Philosophy of Science* **55**: 538-559.

Forster, M. (2000a): "Hard Problems in the Philosophy of Science – Idealisation and Commensurability." In R. Nola and H. Sankey (eds.), *After Popper, Kuhn, and Feyerabend.* London: Kluwer, pp. 231-250.

Forster, M. (2000b): "Key Concepts in Model Selection – Performance and Generality." *Journal of Mathematical Psychology* **44**: 205-231.

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**: 1-36

Forster, M. and Sober, E. (2000): "Why Likelihood?" In M. Taper and S. Lee (eds.), *The Evidence Project*, Chicago: University of Chicago Press.

Gilovich, T., Valone, R., and Tversky, A. (1985): "The Hot Hand in Basketball – On the Misperception of Random Sequences." *Cognitive Psychology* **17**: 295-314.

Johnson, D. (1995): "Statistical Sirens -- the Allure of Nonparametrics." *Ecology* **76**: 1998-2000.

Jukes, T., and Cantor, C. (1969): "Evolution of Protein Molecules." In H. Munro (ed.), *Mammalian Protein Metabolism*. New York: Academic Press, pp. 21-132.

Lewis, P. (1998): "Maximum Likelihood as an Alternative to Parsimony for Inferring Phylogeny Using Nucleotide Sequence Data." In D. Soltis, P. Soltis, and J. Doyle (eds.), *Molecular Systematics of Plants II*. Boston: Kluwer, pp. 132-163.

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

Morgenbesser, S. (1960): "The Realist-Instrumentalist Controversy." In. S. Morgenbesser, P. Suppes, and M. White (eds.), *Philosophy, Science, and Method*. New York: Harcourt, Brace, and World, pp. 106-122.

Nagel, E. (1979): *The Structure of Science*. Indianapolis: Hackett.

Popper, K. (1959): *Logic of Scientific Discovery.* London: Hutchinson.

Royall, R. (1997): *Statistical Evidence – a Likelihood Paradigm*. Boca Raton: Chapman and Hall.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986): *Akaike Information Criterion Statistics*. New York: Springer.

Schoener, T. (1970): "Nonsynchronous Spatial Overlap of Lizards in Patchy Habitats." *Ecology* **51**: 408-418.

Schwarz, G. (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6**: 461-465.

Sober, E. (1988): *Reconstructing the Past – Parsimony, Evolution, and Inference.* Cambridge:

MIT Press.

Sober, E. (1998): "Instrumentalism Revisited." *Critica* **31**: 3-38.

Swofford, D., Olsen, G., Waddell, P. and Hillis, D. (1996): "Phylogenetic Inference." In D. Hillis, C. Moritz, and B. Marble (eds.), *Molecular Systematics*. Sunderland, MA: Sinauer, 2nd edition, pp. 407-514.

Yoccoz, N. (1991): "Use, Overuse, and Misuse of Significance Tests in Evolutionary Biology and Ecology." *Bulletin of the Ecological Society of America* **32**: 106-111.

Van Fraassen, B. (1980): *The Scientific Image*. New York: Oxford University Press.

Notes

1. My thanks to Marty Barrett, Kevin DeQuieroz, Michael Donoghue, Branden Fitelson, Malcolm Forster, and Dan Hausman for useful discussion.

2. Perhaps the simple explanation of why scientists behave in the peculiar way I have described is that they accept frequentist statistics. This raises two questions: Is the frequentist approach sound? Does frequentism have instrumentalist commitments?

3. Forster and Sober (1994) describe Akaike's estimated predictive accuracy as a quantity *per datum*, and so divided the right side of this equation by N, the number of data.

4. Not only can AIC be compared with other criteria that have been defended as methods for maximizing predictive accuracy; in addition, one can assess methods that have been developed for quite other reasons by seeing how well they do in prediction problems. Two examples are BIC, the Bayesian information criterion of Schwarz (1978), and Neyman-Pearson likelihood ratio tests. BIC was developed as a method for assessing the average likelihood of a model; it proposes a criterion that gives more weight to simplicity than AIC does. Similarly, when the likelihood ratio test is applied to nested models (with 1 degree of freedom), it too embodies a policy that gives more weight to simplicity than AIC does (Forster 2000b).

5. If (Diff) is true, then there exists a member of (Diff) – call it T(Diff) -- that is true. If one only knew the identity of T(Diff), one could use that hypothesis to predict new data, and no other assignment of parameter values, to either (Null) or (Diff), will yield more accurate predictions. In this sense, Nagel was right to suggest that the truth is the best predictor. But what is true of the members of (Null) and (Diff) is not true of those models themselves.

6. Can a realist adopt the ecumenical view that *one* of the things that science aims for is predictive accuracy and, therefore, Akaike's framework and criterion do not conflict with realism? That depends on whether realism is the innocuous claim that *one of the goals* of theorizing is to discover which theories are true, or the more substantive claim that *the unique ultimate goal* of theorizing is the discovery of truth (Sober 1998).