Parsimony

Scientists and philosophers often claim that the parsimony (or simplicity) of a theory is relevant to deciding whether the theory is true, or approximately true, or would make accurate predictions. It is a central puzzle in the epistemology of science how this can be so. It is not puzzling that people find parsimonious theories aesthetically attractive and easy to understand and manipulate. Rather, it is the *epistemic value* of parsimony, not its *pragmatic value*, that requires elucidation.

Just as "parsimony" is predicated of people when they are abstemious in the money they spend, so theories are parsimonious when they are tight-fisted with respect to the entities, processes, or events they postulate. There is no cut-off separating theories that are parsimonious from theories that are not; rather, the difference is a matter of degree. The fundamental idea is comparative – one theory is *more parsimonious than* another. For example, if one theory postulates causes A and B to explain an observed effect E, while a second postulates A as causing E but makes no mention of B, it is the latter theory that is more parsimonious. One epistemically significance feature of this difference is not far to seek – if A and B are mutually independent, then the axioms of probability theory guarantee that the conjunction (A and B) will be less probable than A. Does this mean that parsimony and probability always coincide? We will see in what follows that a number of philosophers have strenuously denied this. But even in the case at hand, there is reason to be careful about this suggestion. The second theory is *agnostic* about the relevance of B. But now consider a third theory, which asserts that A causes E and *denies* that B does so. This third theory is "atheistic" about B, not agnostic. This third theory is more parsimonious than the first. However, it is not a theorem of probability theory that (A and -B) is more probable than (A and B). The hypothesis that there is *at least one* cause of E is more probable than the hypothesis that there are *at least two*, but there is no *a priori* reason to think that *exactly one* is more probable than *exactly two*. The principle of parsimony – a.k.a. Ockham's razor, named for William of Ockham, the medieval philosopher who said that plurality is not to be assumed without necessity and that what can be done with fewer assumptions is done in vain with more (Wood 1996, pp. 20-22) – has an obvious link with probability when a logically stronger hypothesis is compared with one that is simpler and logical weaker; however, when two theories are mutually incompatible, the connection is anything but obvious.

The giants of the Scientific Revolution frequently referred to the importance of parsimony and its cognates. In *De Revolutionibus Orbium Caelestium*, Copernicus emphasizes that his heliocentric theory differs from Ptolemy's geocentric theory in that Ptolemy requires an independent model for the motion of each planet, whereas he, Copernicus, *unifies* the models for the different planets by including a common earth-sun component in each. Copernicus remarks that his approach "follow[s] Nature, who producing nothing vain or superfluous often prefers to endow one cause with many effects (Kuhn, pp. 176-179)." Newton (1686, p. 3) echoes this sentiment in *Principia* when he states as his first Rule of Reasoning in Philosophy that "we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity and affects not the pomp of superfluous causes." Leibniz (1686, p. 11) defended the use of a parsimony criterion in scientific reasoning by appeal to his doctrine that God created the best of all possible

worlds; our world is perfect because it is "at the same time the simplest in hypotheses and the richest in phenomena." For all these thinkers, the methodological principle rests on an ontological foundation. We should use the principle of parsimony in our reasoning because nature is simple. And nature is simple because God made it so.

With the falling away of divine design as an acceptable justification of methodological principles, a vacuum appeared in the foundations of scientific inference. If the justification of the principle of parsimony is not to be traced back to a parsimonious Creator, what could that justification be? Does the justification of the principle require any substantive assumptions about the natural world? Or is the principle just part and parcel of what it means to be "rational," which we are required to be no matter what the world is like? If the theological account is the *thesis*, its *antithesis* is the idea that the principle of parsimony is purely methodological. In between these two extremes, there is much room for *synthesis*.

Local versus Global Accounts

Most attempts to explain the epistemic relevance of parsimony treat the problem *globally*. They assume that if parsimony is epistemically relevant across a range of inference problems, that it must have that relevance always for the same reason. However, it is worth pondering the possibility that the justification for using a principle of parsimony may vary from problem to problem. Perhaps parsimony needs to be understood *locally*, not globally (Sober 1990).

As an example, consider the longstanding use of parsimony as a criterion for inferring phylogenetic relationships in evolutionary biology (Sober 1988). When we observe the similarities and differences that characterize a set of species, how are we to use these data to figure out which species are closely related and which are related only more distantly? A standard procedure is to find the phylogenetic tree that requires the smallest number of changes in character state to explain the data. This methodology assumes that the species are genealogically related and proceeds to identify the most parsimonious hypothesis concerning what that pattern of relatedness is. However, there is a prior question about phylogeny – why think that the species we observe share any common ancestors? Perhaps each traces back to a separate origination event.

The role of parsimony in answering this question can be understood by examining Crick's (1968) argument that the near universality of a single genetic code among the organisms now on earth is evidence that they are all genealogically related. Crick says that the genetic code we share is arbitrary – it is one among a large number of viable mappings of nucleotide triplets onto amino acids. However, once an organism uses a given code, there are likely to be deleterious fitness consequences if it or its descendants modify the code already in place. Stabilizing selection then makes it highly probable that descendants will use the same genetic code as their ancestors. These biological assumptions (which Crick summarizes with the phrase "frozen accident") entail that the code's universality would be very surprising if the organisms now on earth were not genealogically related (e.g., were products of 27 separate start-ups), but is precisely what one should expect if all life traced back to a single progenitor. Because of this difference, Crick concludes that the observed universality *strongly favors* one hypothesis over the other. Notice that Crick's argument compares the *likelihoods* of two hypotheses:

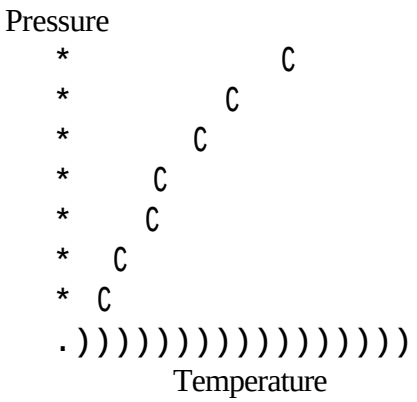
P(the code is now universal * all current life traces back to a single progenitor) >
P(the code is now universal * current life traces back to 27 original progenitors and no fewer).

Here *likelihood* is used in the technical sense introduced by R.A. Fisher (1925); the likelihood of a hypothesis is the probability it confers on observations, not the probability of the hypothesis, given the observations. H's likelihood is P(O * H); its posterior probability is P(H * O). The *law of likelihood* says that the observations differentially support the hypothesis of higher likelihood (Hacking 1965, Edwards 1971, Royall 1997).

The hypothesis that life traces back to a single progenitor is more parsimonious than the hypothesis that it traces back to 27 separate start-ups (since 1 < 27). Crick's argument thus provides an example in which the principle of parsimony has a likelihood justification. However, the connection of likelihood and parsimony in this instance depends on specifically biological assumptions about the genetic code – that it is arbitrary and that it is subject to stabilizing selection. If parsimony has a likelihood rationale in inference problems that arise in other sciences, different empirical assumptions will be required to show that this is so. But more importantly, there seem to be problems in which parsimony cannot be justified in terms of likelihood; in these problems, likelihood and parsimony are actually at odds with each other.

The inferential task of *curve-fitting* provides an example. Consider the following experiment. You put a sealed pot on the stove. The pot has a thermometer attached to it as well as a device that measures how much pressure the gas inside exerts on the walls of the pot. You heat the pot to various temperatures and observe the resulting pressure. Each temperature reading with its associated pressure reading can be represented as a point in the coordinate system depicted in the accompanying figure. The problem is to use these observations to decide what the general relationship is between temperature and pressure for this system. Each hypothesis about this general relationship takes the form of a curve. Which curve is most plausible, given the observations?

Figure



One factor that scientists attend to is goodness-of-fit. A curve that comes close to the data fits them better than a curve that is more distant. If goodness-of-fit were the only relevant consideration, scientists would always choose curves that pass exactly through the data points. They do not do this (and even if they did, the question would remain of how they choose among the infinity of curves that fit the data perfectly). Another consideration apparently influences their decisions, and this is simplicity. Extremely bumpy curves are often thought to be complex, whereas smoother curves are often thought to be simpler. Scientists sometimes reject an extremely bumpy curve that fits the data perfectly in favor of a smoother curve that fits the data a little less well. Scientists care about goodness-of-fit *and* simplicity; both considerations influence how they choose curves in the light of data. What is more, these two *desiderata* are in conflict -- a sufficiently complex curve will fit the data perfectly, whereas a simpler curve will often fail to do this. Increasing simplicity typically involves reducing goodness-of-fit.

A curve represents a deterministic relationship between temperature and pressure; it maps x-values onto unique y-values. However, a curve plus an error distribution represents a probabilistic relationship – each x-value is associated with a distribution of possible y-values, each with its own probability (density). In our experiment, the latter conception is more plausible, since the data are the joint product of the true underlying relationship of temperature and pressure and the measurement errors introduced by the imperfections of the thermometer and the pressure gauge. A standard model of error effects a connection between goodness-of-fit and likelihood – if one curve fits the data better than another, then the former confers a higher probability on the data. A straight line will have a lower likelihood, given the data set depicted in the figure, than a sufficiently complex curve that passes exactly through each data point. Thus, even if parsimony has a likelihood rationale in Crick's argument, simplicity and likelihood are apparently in conflict in the context of curve-fitting.

Simplicity and Parsimony

It seems natural to say that curves differ in their simplicity. But what would it mean to say that they differ in parsimony? Parsimony involves paucity of postulation,. but how does the idea of abstemiousness apply in the context of curve-fitting? Curves are visual representations of equations. For example, a straight line is a representation of an equation that has the form

(LIN) $y = a + bx$

and a parabola is a representation of an equation that has the form

(PAR) $y = a + bx + cx^2$,

where x and y are the independent and dependent *variables*, and a, b, and c are *adjustable parameters.* In such equations, the adjustable parameters represent existential quantifiers – for example, (LIN) says that *there exist* values for a and b such that $y = a + bx$. It may seem, therefore, that (LIN) is more parsimonious than (PAR) because the former makes two existence claims, whereas the latter makes three. This point pertains to (LIN) and (PAR), not to a specific straight line and a

specific parabola (an important distinction, which will come up again). Do simplicity and parsimony in their vernacular meanings always come to the same thing? The accounts described in what follows draw no distinction between them.

Bayesianism

Bayesianism is not the same as Bayes's theorem. The theorem says that the conditional probability P(H * O) -- the probability of H, given O – is a function of three other quantities:

$$P(H * O) = P(O * H)P(H)/P(O).$$

The theorem is a consequence of the standard definition of conditional probability – P(H * O) = P(H and O)/P(O). Bayesianism is a philosophical position, not a mathematical truth; in its strongest form it asserts that the epistemic notion of plausibility can be understood in terms of the mathematical concept of probability and, furthermore, that all the epistemic concepts that bear on empirical inquiry can be understood in terms of the probabilistic relationships described by Bayes's theorem. A double application of this theorem yields the following comparative principle:

$$P(H_1 * O) > P(H_2 * O) \text{ if and only if } P(O * H_1)P(H_1) > P(O * H_2)P(H_2).$$

This biconditional makes it clear that there are exactly two ingredients that Bayesianism gets to use in explaining how parsimony can make one hypothesis more plausible than another in the light of a set of observations. If parsimony influences plausibility, it must do so via the prior probabilities or via the likelihoods (or both). If the relevance of simplicity cannot be accommodated in one of these two ways, then either simplicity is epistemically irrelevant or (strong) Bayesianism is mistaken. As noted previously in connection with the curve-fitting problem, likelihood can be maximized by making one's hypothesis sufficiently complex; this seems to leave the Bayesian only one alternative -- if simplicity in such cases influences a hypothesis' plausibility, it must do so because simpler theories have higher prior probabilities. This led Jeffreys (1957) to introduce a *simplicity postulate*, according to which the complexity of an equation is measured by summing the number of variables, exponents, and parameters it contains. This simplicity ordering is then said to provide an ordering of the hypotheses' prior probabilities.

Popper (1959) pointed out that this postulate is incompatible with the axioms of probability. It assigns (LIN) a higher prior probability than (PAR), but this is impossible, since (LIN) entails (PAR). Howson (1988) replied that this problem can be evaded by stipulating that the parameters in a model have nonzero values. Instead of comparing (LIN) and (PAR), we are to compare (LIN*) and (PAR*) (which stipulate that a,b,c ... 0); these models are disjoint, not nested, so assigning the former a higher prior probability is consistent with the axioms of probability. Two new questions now arise. The first concerns why we should ignore the original problem of comparing (LIN) and (PAR). Should we say that these two models are not in competition because they are compatible? If so, scientific practice needs to change, since scientists often compare nested models. The second question concerns why (LIN*) should be assigned a higher prior probability than (PAR*). Why think that c=0 is more

probable than c...0? If probabilities are merely subjective degrees of belief, it is not to be denied that someone *might* have greater confidence in the hypothesis that c=0. But it is puzzling why, in the absence of evidence, one should feel this way. If a sharp pin is dropped on a line a mile long, would you bet that the pin will land exactly at the beginning of the line, or that it will land somewhere else? In the absence of information concerning how the pin is dropped, it is hard to see why you should bet on the former option. Yet, this is precisely what Jeffreys' simplicity postulate recommends.

Another problem with the simplicity postulate – one that has to do with its completeness, not its correctness – is that it imposes an ordering of prior probabilities without providing specific values. This is important in inference problems where more complex hypotheses have higher likelihoods. If H_1 has the higher likelihood and H_2 has the higher prior, which of them has the higher posterior probability? The question of how simplicity *trades off* against likelihood requires more than a simplicity ordering.

Although Jeffreys held out no hope of getting likelihood and parsimony to coincide, later Bayesians saw a way to reopen the question. To grasp their idea, it is important to attend to the difference between *models* (which contain at least one adjustable parameter) and *specific hypotheses* (which contain none). (LIN) is a model, but "y= 2 + 3x" is not; it is a specific linear hypothesis. In effect, a model is a disjunction of specific hypotheses. When it was noted earlier that a sufficiently complex equation will fit the data better than a simpler equation, the point pertains to specific hypotheses. But what would it mean to talk about the likelihoods of models? It is clear how "y = 2 + 3x" probabilifies the data (once an error distribution is specified). But what probability does (LIN) confer on the data? The answer is that the likelihood of (LIN) is the *average* likelihood of the set of straight lines (i = 1,2, ...):

$$P(\text{Data} * \text{LIN}] = \sum_i P(\text{Data} * \text{straight line i})P(\text{straight line i} * \text{LIN}).$$

The first term in this summation makes sense, but what are we to make of the second? If the relation between temperature and pressure in our example is linear, what probabilities do the different specific linear hypotheses have? Schwarz (1978) approached this problem by thinking about the *ratio* of the average likelihoods of two models, using the assumption that there is a flat, uniform distribution over parameter values in each model. He derived the following result,
which came to be known as the *Bayesian information criterion* (BIC):

$$\text{Log}[P(\text{Data} * \text{Model M})] \approx \text{Log}\{P[\text{Data} * L(M)]\} - (k/2)\text{Log}(N).$$

Here L(M) is the likeliest member of model M, N is the number of data, and k is the number of parameters in M. Notice that BIC includes a penalty term for complexity. If the best-fitting straight line and the best-fitting parabola fit the data in the figure about equally well, (PAR) will have the lower estimated average likelihood because it is more complex. Complexity is relevant to estimating the average likelihoods of models, so Jeffreys' recourse to priors in his simplicity postulate is not, it turns out, the only Bayesian approach to the problem.

One virtue of Schwarz's analysis is that it avoids the criticism already noted that it seems arbitrary and implausible, if not contradictory, to assign simpler models higher prior probabilities

(nonetheless, questions can be raised about the assumed flat prior distribution of the values a parameter might have in a model). Another virtue is that BIC specifies an exact quantitative rule for trading off simplicity and the likelihood of L(M); it describes how much of a gain in one is required for a given loss in the other, if there is to be a net improvement in the model's estimated average likelihood. However, there is a fly in the ointment. Schwarz's derivation uses improper priors (i.e., priors that do not sum to unity) in such a way that his derivation is not invariant under reparameterization (Forster and Sober 1994). Subsequent Bayesian work derives BIC so as to avoid this defect; the strategy is to use some of the data to transform the initial, improper, priors into proper posteriors, after which the rest of the data are taken into account to compute the final, average, likelihood. For further discussion see Wasserman (2000).

Popper and Falsifiability

Popper (1959) proposed a demarcation criterion that separates scientific from nonscientific statements – the former are falsifiable. A falsifiable statement is one that is incompatible with a finite conjunction of observation statements. Falsifiable statements don't have to be false; rather, they have the nice property that observation can disprove them if, in fact, they are untrue.

Just as falsifiability separates science from nonscience, so degree of falsifiability distinguishes some scientific statements from others. The (LIN) model can be falsified by three data points, but not by any smaller number. A single data point, or any pair of them, can be supplied with a straight line that passes through them exactly. (PAR), on the other hand requires at least four data points to be falsified. This means that (LIN) is more falsifiable than (PAR). Popper saw this as the key to understanding simplicity in science. Simpler theories are easier to falsify; they take less data to show that they are false, if indeed they are. Popper turns Jeffreys' simplicity postulate on its head; whereas Jeffreys thinks that simpler theories are more probable, Popper thinks that simplicity goes with greater content – simpler theories say more, and hence are more *im*probable.

It is clear that hypotheses that are more falsifiable have a pragmatic virtue – it is easier for us to prove them false if indeed they are. The principal hesitation that philosophers have had with Popper's analysis is that it fails to account for parsimony's epistemic significance. Why should we base our predictions on simpler models, rather than on more complicated models that fit the data equally well? It is here that Popper aligns himself with the skeptic and in opposition to the Bayesian. We have no assurance that our best hypotheses are true, or even probably true. All we can say is that they so far have evaded our best attempts to disprove them. Simplicity provides no guarantee of truth or of probable truth for the simple reason that nothing does.

There are further problems with Popper's account of simplicity. First, although it entails that (LIN) is simpler than (PAR), it does not have this consequence when we compare a specific straight line and a specific parabola. Each can be falsified by a single data point, so they are equally falsifiable; this means that Popper must say that they are equally simple. In addition, Popper's notion of degrees of falsifiability is restricted to hypotheses that have deductive consequences (perhaps in conjunction with auxiliary assumptions) about observations. If the hypotheses in question only confer probabilities on the data, they are not falsifiable. Since observation is virtually always subject to error, this is a large gap in Popper's theory.

Akaike and Model Selection

The Bayesian approach to model selection is not the only game in town. Before Schwarz (1978) proved his result, Akaike (1973) provided an alternative treatment (see also Sakamoto *et al.* 1986 and Burnham and Anderson 1998). In fact, Akaike's contribution was two-fold: he described a goal for model selection, *predictive accuracy*, and he proved a theorem concerning how the predictive accuracy of a model can be estimated (Forster and Sober 1994).

How might a model like (LIN) be used to make a prediction about the pressure in our pot, if we bring the pot to a certain temperature? A specific linear hypothesis, such as "y = 2 +3x," makes a prediction about the y-values that will be associated with newly observed x-values, but what does (LIN) tell us to expect? The answer is that (LIN) makes predictions via a two step-process. First, one uses old data to find the maximum likelihood estimates of the parameters in (LIN); then one uses this fitted model to predict new data. Thus, from the old data and (LIN), one obtains L(LIN), the likeliest member of (LIN), and it is L(LIN) that makes a definite prediction about new data.

How well will L(LIN) do in predicting new data? That depends, of course, on the true underlying relation of temperature and pressure. In addition, since different data sets drawn from the same underlying distribution may differ; L(LIN) may make fairly accurate predictions about some and rather inaccurate predictions about others. Because data sets may vary, it makes sense to define the predictive accuracy of a model as its average performance across multiple data sets.

If maximizing predictive accuracy is the goal, how is this goal to be achieved? How are we to tell whether a model will make accurate predictions about new data, given just the single data set we have at hand? If we simply find the model that best fits the data, we will usually opt for a fairly complex model. Working scientists know from practical experience that a complex model fitted to old data often does a poor job predicting new data; in such cases, the model is said to *overfit* the data. Sometimes a simpler model, though it fits the old data less well, will do a better job predicting new data. Akaike's (1973) theorem provides a mathematical explanation of this familiar fact. It says:

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

One obtains the log-likelihood of the best-fitting member of the model, and then subtracts k, where k is the number of adjustable parameters in the model; k is a penalty for complexity. This estimate is termed the model's AIC (Akaike information criterion) score. Forster and Sober (1994) recommend that the estimate be represented *per datum* – i.e., that the right-hand side be multiplied by 1/N, where N is the number of data; this helps diffuse the criticism that AIC is statistically inconsistent (Forster 2002). Although it is intuitive to think of Akaike's framework in the context of curve-fitting, it and other model selection criteria apply to a far larger range of inference problems, including ones that arise in causal modeling (Forster and Sober 1994).

Akaike's theorem is a theorem, so it is important to attend to the assumptions that go into its proof. First, there is a "uniformity of nature" assumption, which has two parts. It says that the old and new data sets described in the definition of predictive accuracy are drawn from the same underlying distribution. It also assumes that the x-values sampled in different data sets are drawn from a single distribution; for this reason, Forster (2000) describes AIC as addressing the problem of *interpolation*;

the model selection criterion appropriate for *extrapolation* is not addressed by Akaike's theorem. The proof of Akaike's theorem also requires a normality assumption; roughly, this says that repeated estimates of a parameter in a model form a normal distribution.

What does it mean to say that AIC is unbiased? If your bathroom scale is unbiased, it may give different readings of what you weigh, but the average of these must be your true weight. If the scale is unbiased, so is the procedure of adding or subtracting 50% of what it says, depending on the result of a fair coin toss. This second estimation procedure also is centered on the true value, but it has higher variance than the one that just takes the scale's reading at face value. Similarly, the fact that AIC provides an unbiased estimate of a model's predictive accuracy
leaves open whether its estimates have minimum variance. Furthermore, it is not clear that lack of bias should be regarded as a necessary condition on an acceptable estimator. Suppose your scale has very low variance, but is slightly biased; on average, it reads a little too high or a little too low (you don't know which). Would you decline to use this scale, if the alternative is to use a scale
that is unbiased but has enormous variance?

AIC and BIC are often treated as competitors in the model selection literature. This is odd, since the two criteria were derived as solutions for different problems. BIC estimates average likelihood; AIC estimates predictive accuracy. This does not mean that they cannot be considered as possible solutions to the same problem; however, to do so involves wrenching one of them from its natural conceptual home. Forster (2002) describes a set of simulations in which AIC does a better job estimating predictive accuracy in some circumstances, while BIC
does better in others. If one knew in advance where the problem one wishes to solve is located in parameter space, such simulations may indicate which model selection criterion to use. However, the sad fact of the matter is that one often does not know enough about a problem's factual setting for this to be possible.

The Akaike framework and criterion have important implications for the debate concerning realism, empiricism, and instrumentalism. It often turns out that a model known to be false has a higher AIC score than a model known to be true. This means that the goal of finding models that are predictively accurate differs from the goal of finding models that are true. If realism maintains that the goal of science is to find theories that are true, and empiricism maintains that the goal of science is to find theories that are empirically adequate (Van Fraassen 1980), then Akaike's framework and theorem open the door to a third possibility. Instrumentalism, shorn of the faulty philosophy of language that led it to deny that theories have truth values, becomes an option worth exploring (Sober 2002).

References

Akaike, H. "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, 1973, pp. 267-281.

Burnham, K. and Anderson, D. *Model Selection and Inference – a Practical Information-*

Theoretic Approach. New York: Springer, 1998.

Crick, F. "The Origin of the Genetic Code." *Journal of Molecular Biology* 38 (1968): 367-379.

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

Fisher, R. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

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

Forster, M.R. "The New Science of Simplicity." In A. Zellner, H. Keuzenkamp, and M. McAleer (eds.) *Simplicity, Inference, and Modeling*. Cambridge: Cambridge University Press, 2002, pp. 83-119.

Forster, M. and Sober, E. "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 (1994): 1-36.

Hacking, I. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press, 1965.

Howson, C. "On the Consistency of Jeffreys's Simplicity Postulate and its Role in Bayesian Inference." *Philosophical Quarterly* 38 (1988): 68-83.

Jeffreys, H. *Scientific Inference*. Cambridge: Cambridge University Press, 2nd ed, 1957.

Kuhn, T. *The Copernican Revolution*, Cambridge: Harvard University Press, 1957.

Leibniz, G. *Discourse on Metaphysics*, written in 1686, first published in 1840. La Salle, Illinois: Open Court Publishing Co., 1973.

Newton, I. "Rules of Reasoning in Philosophy," *Philosophiae Naturalis Principia Mathematica*, 1686. Reprinted in H. Thayer (ed.), *Newton's Philosophy of Nature*. New York: Hafner, 1953.

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

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

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

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

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

Sober, E. "Let's Razor Ockham's Razor." In D. Knowles (ed.), *Explanation and Its Limits*, Cambridge: Cambridge University Press, 1990, pp. 73-94. Reprinted in E. Sober, *From a Biological Point of View*. Cambridge: Cambridge University Press, 1994.

Sober, E. "Instrumentalism, Parsimony, and the Akaike Framework." *Philosophy of Science* // (2002): ///-////.

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

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

Wood, R. *Ockham on the Virtues*, West Lafayette, IN: Purdue University Press, 1996.

Acknowledgments

My thanks to Malcolm Forster and Steven Nadler for helpful discussion.