A Philosopher's Guide to Empirical Success

Malcolm R. Forster
Draft, March 16, 2006

ABSTRACT: The simple question—What is empirical success?—turns out to have a surprisingly intricate answer. The paper begins with the point that empirical success cannot be equated with goodness-of-fit without making some kind of distinction between meritorious fit and "fudged fit". The proposal that empirical success is adequately defined by Akaike's Information Criterion (AIC) is analyzed in this light. What is called cross-validated fit is proposed as a further improvement. But it still leaves something out. The final proposal is that empirical success has a hierarchical structure that commonly emerges from the agreement of independent measurements of theoretically postulated quantities.

1. Introduction: It would be a miracle if our best scientific theories were empirically successful without any of their postulated entities really existing or without the theories being approximately or partially true. This is commonly known as the miracle argument. An equally well known response claims that the mere empirical adequacy of these theories is sufficient to explain their empirical success.[1] Realists and antirealists have not fussed too much about what empirical success is. It is tacitly assumed to be something like the degree to which a theory fits the observed phenomena. In the ideal case, it

[1] A theory is empirically adequate if and only everything it says about the observed phenomena (past, present, and future) is true. See van Fraassen 1980, chapter 2, for an introduction to the realist debate, and the antirealist position mentioned here is, of course, van Fraassen's Constructive Empiricism.

consists in the truth of the observed consequences of a theory. In the less ideal case, some account of observational error is made; in which empirical success is defined in terms of a "least squares" measure of fit, or by some probabilistic measure of fit such as likelihood or the log-likelihood.[2]

But what is empirical success, exactly? The problem is surprisingly complicated. For instance, empirical success cannot be goodness-of-fit with the data, in any unqualified sense, because good fit can be "fudged", for instance, by introducing adjustable parameters. Yet it is standard practice in science to use adjustable parameters; so we need to distinguish between meritorious fit and fudged fit, especially when they occur together.

In section 2, the problem is motivated by simple sounding example—Why are Kepler's laws empirically more successful than Copernicus's theory of planetary motion? Hitchcock and Sober (2004) appeal to Akaike's information criterion (AIC) as a way of distinguishing fudged and meritorious fit, but this proposal has its limitations (Section 3). Section 4 improves upon the proposal, in terms of cross-validated fit, while section 5 explains why this improved answer is incomplete. The final suggestion is that empirical success recurs at successively higher levels of generality as science progresses.

2. Why Should Kepler's Laws Supersede Copernicus's Theory? Kepler's first law states that each planet moves around the sun on an ellipse with the sun at one focus. The law introduces a handful adjustable parameters for each planet—the mean radius, R, or the semi-major axis, the eccentricity and the orientation of the ellipse. Kepler's second

[2] Likelihood is a technical term, which refers to the probability of the observations given the hypothesis (not to be confused with the probability of the hypothesis given the observations, which is a distinctly Bayesian concept).

law tells us that the line drawn from the sun to the planet sweeps out equal areas in equal times. In the special case of a circle (an ellipse with zero eccentricity), the area law implies that the planet moves around the sun with uniform angular velocity. For an ellipse of non-zero eccentricity, a planet has to move with greater angular velocities when it is close to the sun, as Newton would later explain in terms of the inverse square law of gravitation. The second law introduces the period of revolution T as an adjustable parameter. Kepler's third law, also known as the harmonic law, introduces no additional adjustable parameters, but postulates a regularity amongst those already introduced. It says that ratios R^3/T^2, measured independently for each planet, are equal. Qualitatively speaking, the harmonic law says that the planets closer to the sun revolve around the sun with greater angular velocities. There is a sense in which the area law says the same thing about a single planet in different parts of its orbit.

Call a specific set of Keplerian trajectories, one for each planet, a *predictive hypothesis* (or hypothesis, if no confusion will result). It is "predictive" in the sense that it makes exact predictions about the position of any planet at any given time. Kepler's laws define a family of such hypotheses, which I shall call a *model*. According to this terminological convention, Kepler's laws define a model.

First of all, how can we define the goodness-of-fit of Kepler's model? It doesn't matter exactly how fit is defined, so assume that it is the sum of the squared residues, where the residue is the spatial distance between the observed position of a planet and the position predicted by the hypothesis at a particular time. Most of the hypotheses in the model will fit the data very badly. So, how do we define the fit of a *family* of hypotheses?

A charitable definition is that *model fit* is the best fit achieved by any hypothesis in the model.

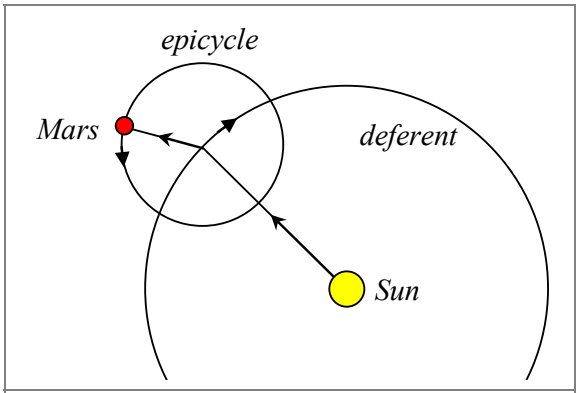

Figure 1: A two-circle Copernican model for the planet Mars. The motion of Mars relative to the sun is modeled as the sum of two vector motions; one represented by the arrow from the sun to the circumference of the main circle, called the deferent, and one from that point to Mars. Each vector has a fixed length and rotates with uniform motion. (The sun could be placed a short distance from the center of the deferent circle, although this would be mathematically equivalent to adding an epicycle.)

Contrast Kepler's model with Copernicus's use of a circle on circle construction (see the caption of Fig. 1 for details). Copernicus's theory allows for many models, each defined by fixing the number of circles assigned to each planet. The adjustable parameters include the radius of each circle, its period of revolution, and the initial position of each circle.

If empirical success were defined as model fit, then it is simply untrue that Kepler's model is empirically more successful than any Copernican model. Take any Copernican model, *C*, and consider another Copernican model, *C*+, that adds one or more epicycles to *C*. Then *C* is *nested* in *C*+ in the precise sense that all the predictive hypotheses in *C* are also in *C*+ (**Proof**: Consider the special cases in which the added epicycles have zero radii). The nested property is sufficient to prove that the more complex model can only improve the model fit, for any hypothesis in *C* is also available in *C*+. Anything that *C* can do, the more complex model can do better. At least in terms of fit. The argument rests solely on the nesting relationship between models—not on how fit is defined.

In fact, there is a theorem in mathematics, called the Fourier theorem, that implies that one can, in principle, approximate any planetary trajectory to an arbitrary degree of accuracy if we use a sufficient number of circles. This also proves that there is a Copernican model that can approximate any finite set of points sampled from the true planetary trajectories to an arbitrary degree of fit. Kepler's model fits only approximately (as we know from Newton's theory). Therefore, there is a Copernican model that exceeds the best fit achieved by Kepler's laws.

So, we can't define empirical success in terms of model fit if we want to maintain the view that Kepler's laws are empirically more successful than Copernicus's theory. The intuitive response is that empirical success must, somehow, take account of the fact that complex Copernican models "fudge" their fit by using a large number of circles. It is not easy to capture this idea precisely.

3. AIC as a Measure of Empirical Success: Hitchcock and Sober (2004) address a related problem—that of distinguishing prediction from accommodation. The idea is that mere accommodation is a kind of fudged fit, and what's left over is a meritorious kind of predictive fit. In what follows, I shall re-describe what they do as providing a definition of empirical success.

Following Forster and Sober (1994), they postulate that the goal of modeling is to maximize predictive accuracy of a model.[3] To define predictive accuracy, we return to a consideration of how well hypotheses fit the true trajectory. To simplify the exposition, let us call the hypothesis that best fits the data the *likeliest*. The negative of this quantity is called the *predictive accuracy of the likeliest hypothesis*. The predictive accuracy of a

[3] The term 'predictive accuracy' was introduced by Forster and Sober (1994).

model is now defined as the average, or the mean, predictive accuracy of the likeliest hypotheses over repeated re-samplings of the observed data. Although these re-samplings are imaginary, the concept is well defined as soon as the method of re-sampling is specified. Models do not wear their predictive accuracies on their sleeves—predictive accuracy is a truth-related utility that is being held up as a *goal* of scientific modeling. It is not proposed as a definition of empirical success.

The predictive accuracy of a model is a property of the whole model because the likeliest hypothesis changes from one data set to the next. It is also a property that depends on the number of observed data. As the number of data points increases, the sampling errors in the estimation of parameters decreases; so that in the limit, the predictive accuracy of a model is the same as the predictive accuracy of the very best hypothesis in the model (called the *model bias*).[4] This means that a sufficiently rich data set will increase the predictive accuracy of sufficiently complex Copernican models without bound. Busemeyer and Wang (2000) and Forster (2002) conclude that predictive accuracy is not the *only* goal of scientific modeling. But there is no argument against considering predictive accuracy as one goal, and an important one at that. So let's consider it.

The question is: How does one estimate predictive accuracy from the observed data? Hitchcock and Sober (2004) appeal to Akaike's theorem, which says that under certain conditions (Akaike 1973; see Forster and Sober for an simple exposition), there is a way of *correcting* the observed model fit so that it provides an *unbiased* estimate of the

[4] I am following Kruse (1997) in terming this *model* bias rather than bias, in the hope of making it harder to confuse the bias of a model with the bias of a statistical estimator.

model's predictive accuracy. The adjusted model fit is referred to as Akaike's Information Criterion (AIC).[5]

AIC provides an aesthetically pleasing definition of empirical success because it divides the model fit into two parts. The penalty term represents the "fudged" part of the fit because it is directly attributable to the use of adjustable parameters, and what's left over is the meritorious part of the model fit. As Hitchcock and Sober correctly emphasize, "fudging" is a normal part of scientific modeling. What matters is our ability to winnow the wheat from the chaff, to *distinguish* the part of the fit that provides a good estimate of predictive accuracy from the "fudged" part.

It is now plausible that the AIC score for Kepler's model is better than the AIC score of *any* Copernican model, which makes AIC score attractive as a definition of empirical success. It is capable of explaining why Kepler's model should have superseded Copernicus's theory.

The Hitchcock-Sober proposal has many other virtues as well. AIC is defined only in terms of the observed model fit, the number of data, and the number of adjustable parameters in the model. So it conforms to Sober's (1993) principle of Actualism, which says what counts as evidence should depend only on what is actually observed. Also, the solution does not rely on the existence of observational errors (even though observational error is, in fact, ubiquitous).

[5] The statistical notion of an "unbiased estimator" is defined in the following way. Imagine repeated re-samplings of the same number of data points from the same segment of the planet's trajectory, where each sample is randomly generated by selecting a set of points on the trajectory according to some fixed probability distribution. It could be, for example, a uniform distribution over the time interval under consideration. Then AIC is an *unbiased estimator* of the model's predictive accuracy if and only if its average value over repeated re-samplings is equal to the true predictive accuracy.

There are many competing model comparison criteria in the literature that also correct the model fit by adding a penalty term—and they are also defined only in terms of the number of adjustable parameters and the number of data; the only difference being in the magnitude of the penalty term. AIC has the special property of being *unbiased* (if the conditions of Akaike's theorem hold). But what's so special about an *unbiased* estimator? Surely, it is more important is to find an estimate that minimizes the expected squared error between the estimator and what's being estimated. And there is no general proof that unbiased estimators must minimize the estimation error in this sense; in fact Stein (1956) describes an example that proves that there is no such proof. However, within the restricted class of estimators that differ from an unbiased estimator by an additive *constant*, there is a proof that the unbiased estimator minimizes the estimation error better than any other estimator in the class.[6] Since the competing adjusted measures of fit, such as the Bayesian Information Criterion (BIC) (Schwarz 1978), differ from AIC by a constant term, AIC is provably better in the sense defined (provided that the conditions of Akaike's theorem hold).

So, what's the problem? The problem with the Hitchcock-Sober proposal is three-fold:

(1) Akaike's theorem itself is technically difficult and fairly opaque, at least compared to the alternative definition of empirical success provided in the next section.

(2) The conditions of Akaike's theorem still may not hold (as Hitchcock and Sober are well aware). In fact, Kieseppä (1997) has questioned whether the conditions

[6] Here's the proof: Let x be the unbiased estimator, and let x^* be the quantity being estimated. By definition of unbiased, $E(x) = x^*$. Now consider the biased estimator $x+b$, where b is a constant. Then $E[(x+c-x^*)^2] = E[(x-x^*)^2] + c^2 > E[(x-x^*)^2]$.

apply to Copernican models. There is no problem with the definition of predictive accuracy in the Copernicus-Kepler example; rather Kieseppä's claim is that AIC is not *justifiably* taken to be an unbiased estimate of the predictive accuracy. His argument rests on the fact that there is no known proof that the conditions of the theorem hold. Of course, this does not settle the issue conclusively—one way of resolving it might be to conduct computer simulations. But in the absence of such studies, Kieseppä's point stands.

(3) The theorem does not apply to the extreme case in which there are more adjustable parameters than data points. This is the paradigmatic case of overfitting; for example, when an $(n-1)$-degree polynomial (with n adjustable parameters) is fitted to n data points, the fit will be perfect. We know that the fit is fudged. AIC is not justifiably applied to this extreme case.

The question is whether it is possible to improve upon AIC. The following section puts forward a proposal.

4. Cross-Validated Fit as a Measure of Empirical Success: Let me begin with a description of the least squares measure of fit to see how it might be modified to provide a more adequate definition of empirical success. Consider a generic curve fitting example in which the model is $y = \beta x$, where β is an adjustable parameter. Now look at the two data points in Fig. 2. The 'distance' of an arbitrary curve in the model, say C,

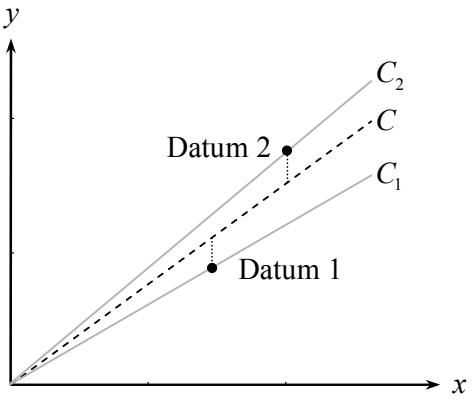

Figure 2: Empirical success in a simple curve fitting example based on a least squares criterion.

from the data may be measured by the *sum of squared residues* (SSR), where the residues are defined as the *y*-distances between the curve and the data points. The residues are the lengths of the vertical lines in Fig. 2. If the vertical line is below the curve, then the residue is negative; otherwise it is positive. Squaring the residues ensures that the SSR score is always greater than or equal to zero, and equal to zero if and only if the curve passes through all the data points exactly. So, the SSR is an intuitively good measure of the discrepancy between a curve and the data.

Now define the curve that *best* fits the data as the curve that has the *least* SSR. Recall that any assignment of numbers to the adjustable parameters determines a unique curve, and vice versa. So, in particular, the best fitting curve automatically assigns numerical values to all the adjustable parameters. These values are the least squares estimates the parameters, and this method of parameter estimation is called the method of least squares.

By fitting a model to the data, we obtain a unique best fitting curve.[7] The values of the parameters determined by this curve are often denoted by a hat. Thus, the best fitting hypothesis in the model would be $y = \hat{\beta} x$. The hypotheses represented by the curves C_1 and C_2 are also in the model, but they have a higher SSR score with respect to the data, even though each fits *one* of the data points perfectly.

More exactly, the model fit is calculated in the following way:

Step 1: Find the hypothesis that best fits the data. Denote this hypothesis by *h*.

Step 2: Consider a single datum. Square the residue of this datum determined by *h*.

Step 3: Go back to Step 2 and repeat this procedure for all *n* data.

[7] There are exceptions to this, for example when the model contains more adjustable parameters than data.

Step 4: Sum the SR scores and divide by n.

This number actually measures the badness-of-fit of the model. The model fit is defined as minus this score.

The reason that we take the *average* SSR in step 4 is that we want to use the goodness-of-fit score to estimate how well the model will predict a "typical" data point. The goal is the same as in simple enumerative induction—to judge how well the "induced" hypothesis predicts a "next instance", where we assume that the seen instances are representative of the parent population.

If the goal is to measure the *predictive accuracy* of the model, then we can see why the SSR score is biased. For each datum has been used twice; once in the construction of the "induced" hypothesis (Step 1), and then to calculate how well the "constructed" hypothesis predicts a typical data point (Step 2). The problem is not the seen data are unrepresentative of the parent population. The problem is that best fitting hypothesis, which is used to represent the model, has been selected, in part, to minimize the "predictive" error. That is why the SSR score is adversely affected when a model is good at *accommodating* data.[8] The problem has nothing to do with the psychological bias of the practitioner; it is a logical problem. And it has a logical solution.

The solution is to measure empirical success in terms of its leave-one-out cross validation score (CV score), which turns out to be surprising similar to the SSR score.

Step 1: Choose a data point i, and find the hypothesis that best fits the remaining $n-1$ data points. Denote this hypothesis by h_i.

Step 2: Square the residue of this datum with respect to h_i.

[8] This does not undermine the least squares method of parameter estimation. There is no bone to pick with statisticians here.

Step 3: Go to Step 1, and repeat this procedure for all *N* data (in all experiments).

Step 4: Sum the scores and divide by *n*.

The difference is the left-out datum *i* is no longer used to "construct" the hypothesis h_i in Step 1. It is therefore an unbiased measure of *prediction*, not accommodation. The comparison of CV scores places simple and complex models on an even playing field; there is no need to factor in non-empirical virtues such as simplicity or unification. The CV score provides a measure of empirical success that is acceptable to realists and antirealists alike.

(1) Not only does the CV score more perspicuously measure the predictive abilities of a model, but it also gives finer-grained information about the nature of its evidence. To show this, let *C* be the curve that best fits the total data, and C_i the curve that best fits the data with datum *i* left out. If SR_i is the squared residue of datum *i* relative to *C*, and PE_i is the squared predictive error of datum 1 relative to C_i, then, by definition, $\mathrm{CV} = \frac{1}{n}\sum PE_i$ and $\mathrm{SSR} = \frac{1}{n}\sum SR_i$. Trivially, $\mathrm{CV} = \mathrm{SSR} + \frac{1}{n}\sum(PE_i - SR_i)$. So CV is equal to the SRR plus a term that corrects the model fit for "fudging". What is not so trivial is that $(PE_i - SR_i)$ is greater than or equal to zero *for each datum*.[9] What this means is that the degree of fudging is estimated for each datum, so that the comparison

[9] **Proof**: Let F be the SSR of the remaining data relative to C, while F_1 is the SSR of the remaining data relative to C_1. Both F and F_1 are the sum of $n-1$ squared residues. By definition, C fits the *total* data at least as well as C_1. Moreover, the SSR for *C relative to the total data* is just $SR_1 + F$ while the SSR of C_1 relative to the total data is $PE_1 + F_1$. Therefore, $PE_1 + F_1 \geq SR_1 + F$. On the other hand, C_1 fits the $n-1$ data at least as well as *C*, again by definition of "best fitting". This implies that $F \geq F_1$. Putting the two inequalities together: $PE_1 + F_1 \geq SR_1 + F \geq SR_1 + F_1$ implies that $PE_1 \geq SR_1$, which is what we set out to prove.

between CV and SSR is heuristically more valuable than the comparison between AIC and SSR. Cross-validated fit can point to the particular data that are not predicted well by the model and pose specific questions about the reliability of those data or how the model might be modified to improve its predictions.

(2) When the conditions of Akaike's theorem hold, the AIC score is approximately equal to the CV score (Stone 1977). So, the CV score can do everything that AIC can do. And it has broader appeal because it does not depend on the assumptions of Akaike's theorem. Even if AIC not an unbiased estimate of predictive accuracy in the case of planetary models, the CV score is still providing a measure of empirical success that it is unbiased by fudging factors.

(3) Now consider a generic curve fitting example in which there are just two data points as in Fig. 2, except that the model under consideration is LIN: $y = a + b\,x$, where a and b are adjustable parameters. The model achieves perfect fit with the data, which is entirely fudged! But AIC cannot sanction this conclusion because Akaike's theorem does not apply when there are as many adjustable parameters as there are data. So, what is the CV score? Well, leave one datum out and try fitting a straight line to a single datum. There is an infinite number of curves that pass through a single point, and they all have different *PE* scores with respect to the left-out datum. Do we say that the empirical success of the model is undefined, or do we somehow average the *PE* scores over a set of curves that best fit the remaining data? In either case, it is fair to say that the model has no empirical success.

Theories such as "God willed X" or "God designed X" are in the same boat. They fit the facts perfectly, but they do not achieve any kind of empirical success. There

is a sense in which they *explain* X, but that goes to show that explaining observed phenomena is not a defining feature of empirical science. They "explain" everything, but there is no explanation of their *empirical success* because they have none.

5. The Hierarchical Structure of Cross-Validated Fit: The final task is to argue that the CV score is an *incomplete* characterization of empirical success. There are two arguments for this. The first is a negative argument against criteria, including CV, AIC or BIC, that have certain asymptotic properties in the large data limit. For it is intuitively obvious that in that leaving one datum out will make no difference to the curve that best fits the remaining data in that limit; that is, C_i is asymptotically the same as C. As data accumulates, the CV score of a complex Copernican models may increase, even to the extent that it surpasses Kepler's CV score. The conclusion is not that these indicators are a bad indicators of empirical success. It's that they are incomplete.

So, what's left out? In the case of the Kepler-Copernicus example, the answer is very simple: Kepler's third law. Kepler's harmonic law does not enter into the calculation of the CV score. Imagine that we leave out a single observation of Mars. Then we find the ellipse that fits the remaining data best, and adjust the period of motion so as to minimize the SSR. We have only used the first two laws because they are the only ones that introduce adjustable parameters. But the presence or absence of Kepler's third law—that is, how well the independent measurements of the ratio R^3/T^2 agree or disagree—is surely part of what determines the empirical success or failure of Kepler's model. Such an agreement is *empirical* because the parameters R and T are empirically determined. Moreover, Newton put great weight on such evidence in his argument for

universal gravitation.[10] It is therefore philosophically significant that standard statistical indices such as CV, AIC or BIC take no account of higher level regularities

Myrvold and Harper (2002) make such a similar complaint about AIC, and I have attempted to bolster their analysis, for instance, by showing that many standard model selection criteria, including cross-validated measures of fit, fall prey to the same objection. Their conclusion is that scientific inference includes something that lies beyond the realm of statistical reasoning. While this issue cannot be resolved here, I would like to be more optimistic about the relevance of statistical notions. For one could introduce a "higher-level" CV index by leaving one planet out, and asking how well Kepler's third law predicts the ratio for the left-out planet. Whether the magnitude of this CV score is significantly greater than zero can be answered by standard statistical tests. On the view advocated here, cross-validated fit has a distinctly hierarchical structure.

Interestingly, Copernicus's advance over Ptolemy's geocentric theory can be viewed in the same way—heliocentric models allow for the overdetermination of the relative motion of the earth and the sun from the motions of many planets (as seen from earth), and the agreement of these independent measurements could also be subjected to standard statistical tests. This higher-level empirical success, which speaks in favor of Copernican theories (including Kepler's and Newton's), is quite independent of whether Copernican models supersede Ptolemaic models with respect to the kind of "next instance" prediction, such as predicting when Easter will fall in coming years. Most historians agree that Copernicus failed to surpass Ptolemy in this regard, but then hastily conclude that there was no empirical evidence in favor of Copernicus's theory.

[10] The importance of the agreement of independent measurements has been recognized in Newton's work, most notably by Harper 2002.

Witness Kuhn, who claims that the "harmony" of Copernicus' system appeals to an "aesthetic sense, *and that alone*." As he puts it:

> The sum of evidence drawn from harmony is nothing if not impressive. But it may well be nothing. "Harmony" seems a strange basis on which to argue for the Earth's motion, particularly since the harmony is so obscured by the complex multitude of circles that make up the full Copernican system. Copernicus' arguments are not pragmatic. They appeal, if at all, not to the utilitarian sense of the practicing astronomer but to his aesthetic sense and to that alone. (Kuhn 1957, 181.)

Contra Kuhn, there is a way of describing the empirical consequences of Copernican theory that makes heliocentric harmony an essential component of its *empirical* success.[11]

It was no accident that Kepler referred to this third law as the harmonic law, and this play on words was not lost on Newton. There is at least one historian of science who understood this well: It was William Whewell (1958; Butts (ed.) 1989) who coined a word for the aspect of evidence most famously overlooked. He called it the *consilience* of inductions.[12]

References:

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

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.

Butts, Robert E. (ed.) (1989). *William Whewell: Theory of Scientific Method.* Hackett Publishing Company, Indianapolis/Cambridge.

Forster, Malcolm R. (1988), "Unification, Explanation, and the Composition of Causes

[11] It is an important caveat that not all of Copernicus's arguments for "harmony" can be explicated in this manner. As Tycho Brahe pointed out, all the empirical consequences of Copernican models are unchanged if one retains the same system of heliocentric circles and makes the earth stand still while the sun moves. That is why Copernican revolution was only *ended* after Newton correctly identified the empirical consequences of the earth's *motion*.

[12] See Forster (1988) for a detailed discussion of Whewell's analysis of Newton's argument for universal gravitation, and how it can be extended to reply to skeptical arguments about the existence of forces.

in Newtonian Mechanics." *Studies in the History and Philosophy of Science* **19:** 55 - 101.

Forster, Malcolm R. (2002), "Predictive Accuracy as an Achievable Goal of Science," *Philosophy of Science* **69**: S124-S134.

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.

Harper, William L. (2002), "Howard Stein on Isaac Newton: Beyond Hypotheses." In David B. Malament (ed.) *Reading Natural Philosophy*: *Essays in the History and Philosophy of Science and Mathematics*. Chicago and La Salle, Illinois: Open Court. 71-112.

Hitchcock, Christopher R. and Elliott Sober (2004): "Prediction versus Accommodation and the Risk of Overfitting," *British Journal for the Philosophy of Science* **55**: 1-34.

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.

Kuhn, Thomas (1957): *The Copernican Revolution*. Cambridge, Mass.: Harvard University Press.

Myrvold, Wayne and William L. Harper (2002), "Model Selection, Simplicity, and Scientific Inference", *Philosophy of Science* **69**: S135-S149.

Schwarz, Gideon (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6:** 461-5.

Sober, Elliott (1993): "Epistemology for Empiricists." In H. Wettstein (ed.), *Midwest Studies in Philosophy*. Notre Dame: University of Notre Dame Press; pp. 39-61.

Stein, C. M. (1956): 'Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution', *Proceedings of the Third Berkley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkley: University of California Press, pp. 197-206.

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.

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