# Unification and Evidence

Malcolm Forster

University of Wisconsin-Madison

March 7, 2005

**The Value of Good Illustrative Examples**:  In order to speak as generally as possible about science, philosophers of science have traditionally formulated their theses in terms of elementary logic and elementary probability theory.  They often point to real scientific examples without explaining them in detail and/or use artificial examples that fail to fit with intricacies of real examples. Sometimes their illustrative examples are chosen *to fit their framework*, rather than the science.  Frequently these are non-scientific examples, which distances the discussion from its intended target.  In the final analysis, philosophical discussions of explanation, confirmation, scientific realism, and the nature of theories are often too abstract, or too imprecise, or too disconnected with real science, to allow scientists to benefit from the discussion.  This is a great loss for both parties.  In my experience, working scientists are confronted with philosophical issues not only in their role as researchers, but also in their role as tertiary teachers of science.

| Table 1: Data | |
|---|---|
| *x* | *y* |
| 1 cm | 1 cm |
| 2 cm | 2 cm |
| 3 cm | 3 cm |

There are no institutionalized rewards to encourage scientists to develop answers to their own philosophical questions in ways that are well thought-out and publicly scrutinized.  And there *should* be an intellectual division of labor; indeed, that is exactly why NSF supports the philosophy of science and other STS disciplines.  On the other hand, it is intended that the *fruits* of the labor should be shared.

The solution is not to abandon illustrative examples.  Nor is the solution to use detailed examples that are necessarily too complicated to explain to non-specialists.  The solution is to construct examples that are simple in their rudimentary form, but can be extended *as the need arises*.  Once introduced, they can be enriched and expanded to illustrate more complicated ideas.  One such example is the beam balance example.  The paragraphs that follow run through some of the ways in which it raises various issues in the philosophy of science.

**The Beam Balance Example**: Suppose we hang an object, labeled *a*, on one side of a beam balance (Fig. 1), and find how far a second object *b*  has to hang on the other
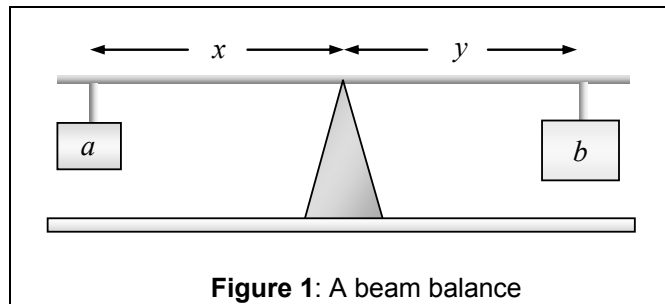


**Figure 1**: A beam balance

side to balance the first.  The distance that *a* is hung from the point at which the beam is supported (called the *fulcrum*) is labeled *x*, while the distance to the right of the fulcrum at which *b* balances *a* is labeled *y*.  If *b* is moved to the left of this point then the beam tips until the object *a*  rests on the ground and if *b* is moved to the right the beam tips the other way.  In the first instance, *x* is 1 centimeter (cm) and the beam balances when *y* is 1 cm.  This pair of numbers is a *datum*, or a *data point*.  We repeat this procedure 2 more times with different values of *x*, and tabulate the resulting data in Table 1.

*Prediction*:  We are now asked:  Given that *a* is hung at a distance of 4 cm to the left of the fulcrum, predict the distance at which *b* will balance; *viz.* predict *y* from *x*.  Once we notice that for all 3 data, *x* is equal to *y*, it appears obvious that we should *predict* that *y* will be 4 cm from the fulcrum when *x* is 4 cm.

*Curve Fitting*:  Curve fitting incorporates a statistical procedure that addresses the prediction problem in the following way:  Once a decision is made about what quantity we want to predict (represented by the dependent variable *y*) and what information we think is sufficient in the context to make the prediction (the independent variable *x*), one needs a formula, or model (an equation or equations with at least one adjustable parameter).  In our example, it is $y = \beta x$, where $\beta \geq 0$. The model is represented by a *family* of curves in the *x-y* plane, in this case, the family of straight lines with non-negative slope that pass through the origin.  A statistical procedure, such as the method of least squares, can now be used to estimate the value of the free parameter $\beta$.  Its estimated value is the slope of the line in the family that *best fits* the data in the sense defined by the method of least squares.  This picks out a unique curve from the family, sometimes called the *fitted model*, which can be used to make precise predictions.

*Statistical Inference*:  If curve fitting is viewed as an inference, then the fitted model is picked out in the conclusion, while a statement of the model and the data are the premises.  Since the model can be viewed as an infinitely long disjunction of 'curves', the inference is deductive if the data are regarded as logically incompatible with all curves but one.  While this is the view of curve fitting often found in philosophy (e.g. Reichenbach 1938; Hempel 1966), it is not found in science.  For the data are never noise-free, and even if they were, we don't generally have a model that fits perfectly.  Gauss and Legendre introduced the method of least square to separate the 'signal' from the 'noise' in the calculation of planetary trajectories.  The procedure introduces an important non-deductive element into scientific practice.  Despite its ubiquitous use in science, and despite a huge industry in logic devoted to non-deductive inference, statistics is not studied widely by logicians or logic-trained philosophers.  There are notable

exceptions.[1]

*Laws*:  The fitted model in our case is $y = x$, which is represented by the curve $C$ in Fig. 2.  This may be thought of as the law (assuming it true) that governs the quantities $x$ and $y$.  Certainly, it seems to support counterfactuals in the sense that it implies that if $x$ were to be 4 cm, then $y$ would be 4 cm.  But is it the fitted model or the unfitted model that is the law?  The model does not entail such counterfactuals, but the fitted model
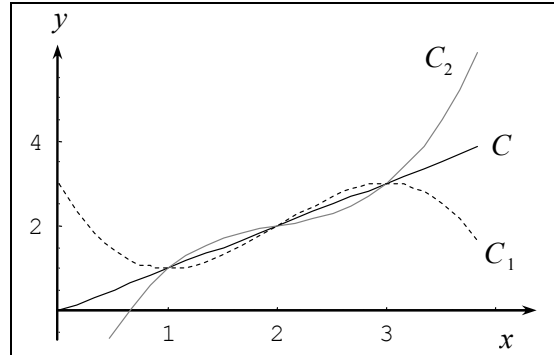


**Figure 2**: The plot of the data points in Table 1, showing three curves that fit the data perfectly.

lacks generality in that it does not 'cover' other beam balance experiments.   So, what is the law properly so-called?  Despite the huge amount of philosophical literature devoted to the nature of laws, this kind of question is obscured by the illustrative examples used.

*The Underdetermination of Theory by Facts*:  The data points represent the observed facts and the fitted model represents the 'theory'.  But there are infinitely many curves, such as $C_1$ and $C_2$ in Fig. 2 that fit the seen data equally well while making incompatible predictions.  Why choose $C$?  This traces back to the question:  Why choose the *model* $y = \beta x$?  The branch of statistics known as *model selection* aims to answer this question.

*Model Selection*:  The beam balance example can be used to illustrate the model selection problem in the following way.  Suppose we consider the possibility that the beam is not properly centered, or that it is heavier on one side than the other.  Then it's as if there is a third object hanging on the beam of unknown mass, and we may account for this by adding an additional adjustable parameter to the model:  $y = \alpha + \beta x$.  How do we decide whether the evidence favors the first model or the second, and what utility is maximized by any given decision rule?

*Simplicity*:  The role of simplicity in scientific inference has always puzzled scientists and non-scientists alike.  There are appeals to vague methodological principles such as Occam's razor or the metaphysical thesis that nature is simple.  None is particularly satisfying.  There is even disagreement about how to judge simplicity in the first place.  Judging the simplicity of *curves* does not help because different curves can be used to represent the same fitted model (Priest 1976).  And if simplicity is a factor in model selection, then it is the simplicity of *models*, not fitted models, that must be instrumental.  The model selection literature agrees almost universally that the simplicity (or

---

[1] Intellectual debts of gratitude are owed to Hacking (1965), Howson and Urbach (1989), Mayo (1996), Rosenkrantz (1977), Sober (1988), Turney (1990), and numerous publications by Teddy Seidenfeld, amongst others.

complexity) of a model is measured by the number of adjustable parameters.[2]

*Naïve Empiricism Doesn't Work!* A simple-minded solution to the model selection problem is that one should favor that model that is capable of fitting the data the best, where 'best' is measured, say, in terms of the least sum of squares. Naïve empiricism implies that no super-empirical features matter. An amazingly simple argument shows that this is wrong! Consider the best fitting curve in the simpler model. The same curve is in the second model, because the first model is nested in the second (when $\alpha = 0$, the second equation reduces to the first).[3] So, there is always a curve in the second model that fits the data as well as the best fitting curve in the first model. *This is true no matter how one measures fit*. The argument also extends to Bayesian hypotheses constructed as averages over the members of a model: Whatever can be constructed from the smaller model is available to the larger model because the first is nested in the second.

Naïve empiricism can *never* favor the simpler model in the case of nested models. Assuming that the simpler model *should* be favored in some cases, then we must take something like simplicity into account. But, how and why? Here is one reason.

*Overfitting*: Models with large numbers of free parameters tend to overfit data. For example, it is a well known theorem of mathematics that an *N*-degree polynomial can fit *N* data points exactly (provided none is exactly on top of any other). Yet we should not trust such a curve for the purposes of prediction. Sufficiently complex models can *accommodate* the data better than simpler models, so good accommodation does not always lead to good prediction. The statistician Akaike (1969, 1971, 1973, 1974, 1977, 1985, 1987, 1994) began with the premise that prediction is the goal in statistical inference, and proved that in a wide variety of conditions, a model's predictive accuracy is better estimated not by its degree of fit, but by its degree of fit plus a penalty for complexity that is proportional to the number of adjustable parameters. Forster and Sober (1994) brought Akaike's work to the attention of philosophers of science (and inadvertently to many scientists and some statisticians). Forster and Sober described Akaike's work in this area not because of any conviction that Akaike's model selection criterion (AIC) is the best, but because Akaike's *analysis* of the phenomenon of overfitting is insightful. Nor was there any claim that Akaike's analysis exhausted everything there was to be said about scientific inference, or even the problem of model selection. Akaike had an *explanation* of the *problem* of overfitting based on the precise assumption that predictive accuracy is the goal of statistical inference. Whether, or to

---

[2] There is more to be said here because the number of adjustable parameters is sometimes an artifact of how the model is described—see the reply to De Vito 1997 in Forster 1999.

[3] Even if $\alpha = 0$ is excluded by the second model (as Bayesians must insist in order to allow that the first model may sometimes have a higher probability than the second), any curve in the first model is still arbitrarily close to some curve in the second model.

what extent, it is a problem for other goals of inference is an open problem. The *strength* of the Akaike framework is that it is precise about the means and the end and the connection between them, and it is therefore precise about the limitations of AIC.

*Predictive Accuracy*: Suppose we have two sets of data; call them the training data and the test data. Define the degree to which a model succeeds in predicting the test data from the training data by how well the curve that best fits the training set fits the test data. This is a prediction score. Now consider repetitions of this procedure with new training data and new test data, and define predictive accuracy as the average prediction score over all these instances generated by whatever mechanism generated the original data sets. Choosing a model with high predictive accuracy is a *goal* of statistical inference. A single prediction score is an unbiased estimate of the model's predictive accuracy. When a model overfits the training data, it will tend to show up in the predictions of the test data. Thus, prediction scores provide an empiricist solution to the problem of overfitting. The prediction score is not equal to the accommodation score, which is obtained by pooling the training and test data. The accommodation score is biased as an estimate of predictive accuracy because it reuses the same data—once for training, then for testing. The argument against naïve empiricism does not apply to the use of prediction scores; they can favor simpler models.

Thus, the argument against naïve empiricism does not prove that we *must* take super-empirical virtues into account. For example, leave-one-out cross validation is a prediction score that is well known in statistics. It averages all the prediction scores obtained by leaving out a single datum a the test set and using the remaining data as the training set. This is an unbiased estimate of the model's ability to predict new data from sets of $N–1$ data, where $N$ is the number of data is the total data. AIC and leave-one-out cross validation are different means to the same end (or roughly the same end).

*Instrumentalist and Realist Goals*: There are two main goals of scientific inference that philosophers have discussed. One is based on the instrumentalist view that the goal of science is to maximize predictive accuracy. Another is based on the realist view that the goal of model selection is to represent the reality underlying the observed phenomena (the data). In my view, both of these goals are legitimate and important in their own right. They are different goals, even though they seem to be related. For surely any model that represents the underlying reality correctly is true, and therefore all its logical consequences are true, and so all its predictions are true. This argument is too quick. If we think of the true hypothesis (with the correct numerical assignments to all free parameters), then there is no hypothesis that is more predictively accurate. But if we think of a true model (true because it contains the true hypothesis) then its predictive accuracy will be less than optimal (because the estimation of the parameter introduces error). In fact, a simple false model can have greater predictive accuracy than a complex

true model.  Less obvious is the fact that a complex false model can have greater predictive accuracy than a simple false model, where the simpler model is better at satisfying realist goals.  The last point is dramatically illustrated in versions of the beam balance example (see below).

*Inference to the Best Explanation*:  Consider the realist goal.  How should one compare rival models in order to optimize this goal?  One initially attractive idea is to appeal to *inference the best explanation*.  A fitted model solves the prediction problem, but something more is needed for explanation.  The idea is that the something extra may provide a criterion for judging when realist goals are met.  But what counts as the best explanation and why?  For that matter, what is an explanation?[4] Van Fraassen (1980) attacks inference to the best explanation on two fronts:  First, even if it can be made precise, it could be reinterpreted as inference to the most empirically adequate hypothesis.  Second, he doubts that it can be made precise.  The next paragraph exposes some basic intuitions about explanation, which end up supporting van Fraassen's view.  That provides a reason for shifting one's focus from explanation towards unification.

*What is a Good  Explanation?*  Suppose that the fitted model $y = x$ solves the prediction problem, at least within the narrow predictive context pertaining to repeated trials of the same experiment (with different values of $x$).   The equation merely states a correlation between two quantities, just as low barometer readings are correlated with stormy weather, or the phases of the moon are correlated with the height of low tides.  There is no intuitive sense in which the model explains the facts, even if it does express a law.  There are two responses to this point.  One is the classical response (Hempel 1965) that the model is not a *fundamental* law of nature.  The second (Salmon 1984) is that the model is not a *causal* law.  The first response makes more sense in the context of the beam balance example, although neither response is entirely satisfactory.

Let's look at how the beam balance data are actually explained.  Newton's theory talks about forces and how they produce motion.  In our example, Newton's theory tells us that the beam will remain motionless (*i.e.*, it will balance) when the forces at every point balance.  What are these forces?  The idea of leverage is familiar to everyone.  For example, everyone knows that it is possible to "magnify" a force using a rigid body such as a crowbar.  A long rigid beam may lever a large boulder if the point at which we push down is a long distance from the fulcrum compared with the distance from the fulcrum to the end of the beam applying the force to the boulder (Fig. 3).  Of course, you have to

---

[4] 'The best explanation' is ambiguous.  On the one hand, 'best explanation' could refer to the *true* explanation, for what could be better than that?  Call this the metaphysical notion of explanation.  If inference to the best explanation refers to an inferential method, then it needs to refer to some means of comparing rival explanations.   Here explanation is used in an epistemological sense.  Sober 2002 compares the use of unification in explicating both senses of 'explanation'.

apply the downward force through a longer distance than the distance that the boulder moves upwards, so the work you do is equal to the work done on the boulder. This is required by the conservation of energy.



**Figure 3**: The principle of leverage is about the magification of forces.

The same principle applies to beam balances. The forces applied to the beam arise from the gravitational forces on the two objects. If $m(a)$ is the mass of $a$, $m(b)$ is the mass of $b$, and $g$ is the gravitational field strength, then $a$ exerts a gravitational force of $m(a).g$ on the beam at the point at which it is hung, and $b$ exerts a force of $m(b).g$ at the point at which it is hung. Now focus on the object $b$. If the beam is to balance then the forces acting on $b$ must balance. That is, the upward leverage of $a$ on $b$ must balance the downward gravitational force $m(b).g$. By the principle of leverage, $a$ is exerting an upward force on $b$ equal its downward force magnified by the ratio of the distance $x$ to $y$. The background theory, viz. Newton's second law, tells us that these two forces must be equal:

$$m(b)g = m(a)g\frac{x}{y}.$$

If we multiply both sides of this equation by $y$ and divide both sides by $m(b).g$, and simplify, we derive the model:

$$y = \frac{m(a)}{m(b)}x.$$

The same equation ensures that the forces acting of $a$ also balance.

The derivation makes tacit appeal to various *auxiliary assumptions*. One is that the beam is uniform in mass and balanced at its center. If we were to abandon this assumption, then we would derive the more complex model mentioned earlier.

Notice that the parameter $\beta$ is now interpreted by the theory as a mass ratio. Strictly speaking, there are now two adjustable parameters, $m(a)$ and



**Figure 4**: The beam balances if and only if the leverage acting upward balances gravitational force acting downward.

$m(b)$, but they cannot be independently estimated, so the effective number of parameters is still one (the dimension of the space of functions). The introduction of mass ratios will be important in understanding the unification of otherwise disparate phenomena.

There is an un-explicated sense in which Newton's laws do provide an explanation of the beam balance facts. The operative word is 'un-explicated'. For we know that it is not merely the deduction that matters, for then X&Y would explain Y, even when X is
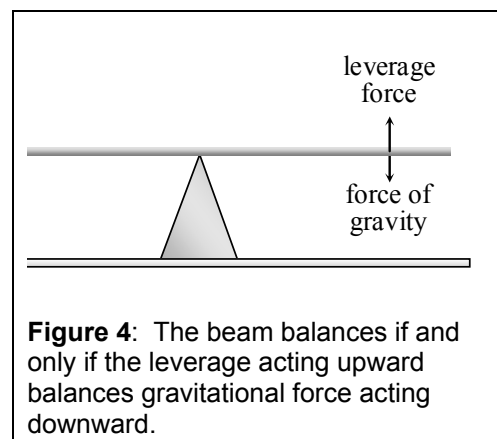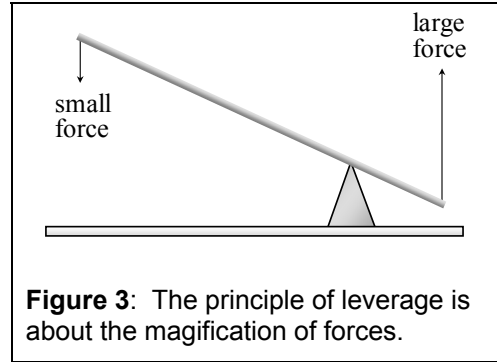
explanatorily irrelevant.  Nor is it enough that the fundamental laws cover more phenomena in a purely deductive sense, or that condition could be satisfied by the irrelevant conjunction as well.

Causal theories of explanation say that the "something extra" that is needed for genuine explanation is some kind of causal story.  But the beam balance example provides a prima facie counterexample to that proposal.  It's not that there is no causal story to be told.  It's that the story actually told is not causal (it appealed only to equilibrium conditions), and nothing more seems to be required.  More specifically, there are at least two distinct causal stories that could be told, and it doesn't seem to matter which one is true.  The first is that the scientist places $b$ at a position $y$ (so the scientist's intentions become part of the explanation of why $b$ is at $y$).  Then she moves $b$ to the right if that sides rises and to the left if it drops, until finally it balances. So, if physicists are interested in causal explanation, why aren't they interested in psychology?   The second story is that that a computer randomly generates a pair of numbers $x$ and $y$ and a robot places the objects at those positions.  The numbers are recorded if and only if the beam balances.  This time the causal story is entirely different, and the difference does not seem to matter.

The argument is even stronger in the case of quantum mechanical phenomena, where Bell (1964) famously proves that some sets of correlations have no plausible causal explanation (that is, no local hidden variable interpretation).  Van Fraassen (1980) views this not merely as an argument against believing in hidden variables, but as an argument against any realist interpretation of quantum mechanics.  This presents a challenge to the realist, which I think is best met by shifting the focus away from intuitions about explanation, and towards unification as a way of making the realist position more precise.  An extension of the beam balance example illustrates a way of developing a realist position along the lines of Earman (1978), Friedman (1981), and Forster (1986, 1988).

*Unification*:  In an application of the beam balance model to a single pair of objects $\{a, b\}$, the following two equations differ only in how the adjustable parameter is represented:

$$y = \frac{m(a)}{m(b)} x \,, \; y = \alpha x \,.$$

The reason that the equations are empirically equivalent in the given context is that they parameterize exactly the same family of curves; namely all the straight lines with positive slope that pass through the origin.  The difference between are merely linguistic.  But if we think of the model more broadly, for example, as applying to three pairs of objects, $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$, then this changes.  Now there are three equations in each model.

UNIFIED: $y_1 = \dfrac{m(a)}{m(b)} x_1$, $y_2 = \dfrac{m(b)}{m(c)} x_2$, and $y_3 = \dfrac{m(a)}{m(c)} x_3$.

COMPLEX: $y_1 = \alpha x_1$, $y_1 = \beta x_2$, and $y_3 = \gamma x_3$.

Each single equation is a model that is also a sub-model of a broader model. The surprising fact is that the broader models are no longer empirically equivalent, even though their sub-models are empirically equivalent. So, in the case of UNIFIED, there is a sense in which the model is greater than the sum of its parts.

This is not because UNIFIED is derivable from Newton's theory, while COMPLEX is not. In fact, there is a clear sense in which UNIFIED entails COMPLEX because UNIFIED is nested in COMPLEX; UNIFIED is a special case of COMPLEX in which the constraint $\gamma = \alpha\beta$ holds necessarily. This is because the third mass ratio, $m(a)/m(c)$ is equal to the product of the other two mass ratios according to the mathematical identity:

CONSTRAINT: $\dfrac{m(a)}{m(c)} = \dfrac{m(a)}{m(b)} \dfrac{m(b)}{m(c)}$.

Since UNIFIED entails COMPLEX and Newton's theory plus certain auxiliary assumptions entail UNIFIED, then by the transitivity of entailment, Newton's theory plus the same auxiliary assumptions entail COMPLEX. Both model are consequences of Newton's theory.

UNIFIED is logically stronger, but it's not merely in the sense in which X&Y is stronger than Y, where X some irrelevant metaphysical assertion that has no empirical consequences. UNIFIED is *empirically* stronger because it entails a relational fact about the data that COMPLEX does not. To draw this out, suppose that both models are true, and there is very little noise in the data or that the data set is large enough that the noise tends to cancel out, so that the statistical estimation of the parameters $\alpha$, $\beta$, and $\gamma$ is fairly accurate. Let $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ denote these estimated values. Then UNIFIED predicts that $\hat{\gamma} - \hat{\alpha}\hat{\beta} = 0$, within the bounds of approximation predicted by the statistical assumptions. COMPLEX does not make this prediction.

UNIFIED provides a better "explanation" of the observed facts not merely because it is entailed by more fundamental laws (for both models are), but because UNIFIED predicts the relational fact $\hat{\gamma} - \hat{\alpha}\hat{\beta} = 0$, whereas COMPLEX merely *accommodates* that fact. It is analogous to Copernicus's argument that the heliocentric view of the solar system *entails* the observed fact that the retrograde motion of the outer planets occurs if and only if the sun is in opposition, whereas the Ptolemaic theory merely accommodates this fact. Copernicus's appeal to the unifying power (harmony) of his system is

accompanied by an *empirical* advantage.  Harmony is not merely an aesthetic virtue of Copernicus's theory, as Kuhn (1957) claimed.

Glymour (1980) provides numerous examples in which empirical regularities are represented as mathematical identities in more advanced theories.  But what is the virtue of the ensuing unification?  Earman (1978) and Friedman (1981) should be credited with seeing clearly that there is an *empirical* virtue involved, while Glymour (1980) saw that there is a conceptual element as well.  Whewell (1858) saw both sides of the issue (see his essay on the fundamental antithesis of philosophy in Butts 1989), but failed to make the point precise (it was missed by Mill, 1874).[5]

*Falsificationism*: This story is subtly different from Popper's thesis that one should favor the most falsifiable model of those that are unfalsified.  In our story both models are unfalsified, and UNIFIED *is* more falsifiable.  So, while Popper's story gives the same answer in this example, it does so for the wrong reasons.  In a case in which X is a purely metaphysical assertion, X&Y is unfalsified if and only if Y is unfalsified, and X&Y is more falsifiable in Popper's logical sense.  But in this case, there is no *empirical* reason to favor X&Y over Y.  There are no relational features of the evidence that support X&Y in favor of Y.  By dismissing the notion of confirmation, Popper's methodology fails to make the vital distinction between strength with confirmation and strength without confirmation.[6]

*The Variety of Evidence*:  In order to show that UNIFIED is better supported by the evidence than COMPLEX, we need values of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$.  This requires that the evidence is sufficiently *varied* or *diverse*.  There may well be other reasons why the variety of evidence is important in science, but this particular reason has not received much emphasize in the philosophical literature, despite the fact that the illustrative example used by Hempel (1966) involves three applications of Snell's law, which have the same logical structure as the beam balance example used here.

*The Composition of Causes*:  Cartwright (1983) presents a skeptical argument against the existence of component causes.  The intuitive argument can be presented in the following way (Forster 1988).  The coffee cup sitting on the table is motionless because it is subject to two component forces, which cancel each other out.  One is the weight of the cup acting downwards, and the other is an equal and opposite force of the table on the cup acting upwards.  But what we see is consistent with the magnitude of the component

---

[5] An explication of Whewell's ideas as they apply to Newton's argument for universal gravitation is found in Forster 1988, and a historically more detailed treatment applied to explicitly to the model selection problem is found in Myrvold and Harper (2002).

[6] Popper (1959) *tries* to define 'falsifiability' in terms of the minimum numbers of data points needed to falsify a model.  But this still fails to emphasize the role of *relational* features of the evidence, and it is not clear that it is coherent in any case—see Howson and Urbach (1989) and Turney (1991) for critiques.

forces being of any magnitude, provide that they cancel each other out. The principle that a cause is measured by its effect therefore dictates that only the *resultant* force can be measured by its observed effect. There is no empirical justification for believing in the existence of component causes.

A full reply to this argument is found in Forster (1988), but a version of the beam balance example illustrates the essential idea. Suppose that a unit mass *u* is hung on the right side of the beam balance at a distance *y*, while *a* and *b* are hung *together* on the right at a distance *y*, such that the beam balances. The simplest Newtonian model is: $y = (m(a) + m(b))x$, where $m(u) = 1$. Within this narrow context, Cartwright's point is correct: There is no way of measuring the values of $m(a)$ and $m(b)$ separately. We can only estimate their sum. But if we repeat the same experiment with *b* and *c*, and then with *a* and *c*, the measurements of the resultants provide an estimate of the component masses, for it provides three equations in three unknowns. So, if the context is widened, so that we have, in effect, *independent measurements* of the same quantity, then the argument falls apart. Cartwright appears to believe that such replies depend on the additivity of masses. But the only requirement is that there is a sufficient number of equations to solve for the unknowns. This point has already been illustrated in the case of mass ratios.

*Inference to the Best Explanation Again*: Consider the model COMP+, which is by definition the model COMPLEX conjoined with the constraint $\gamma = \alpha\beta$. COMP+ is empirically equivalent to UNIFIED within the context of the three experiments, but does not seem to explain the phenomena because it appears *ad hoc*. The difference between them has an important psychological impact, and this influences our judgment of which model best explains the data. This encourages the idea that the problem is solved by viewing scientific inference in terms of inference to the best explanation. The fact that UNIFIED is a better explanation than COMP+ is uninteresting if it reduces to a psychological claim.

*The Realist's Explanandum*: Whatever the outcome of that debate, it is clear that the background theory plays an important role in binding submodels together, at least in the case of mature sciences. For example, Newton's theory coupled with the same auxiliary assumptions as before (that all the beams are centered, and so on) makes predictions about the regularities that holds in other experiments, such as when all three objects are hung together on the same beam. Or it can help predict how they behave on springs (Forster 1986). Even in this extended context, COMP+ is also entailed by the theory. So, while the theory plays an essential role in "covering" disparate phenomena, it does not adjudicate between the two kinds of representations; one in which the constraints are tacked on, and the one in which the constraints are mathematical necessities. In my view, this is inextricably tied to the realist's concern about how one should explain the

empirical success of a theory.  UNIFIED and COMP+ are constructed so that they have the same empirical success, so the realist is committed to the view that both successes should be explained in the same way.   The syntactic form of the theory and its models does not make any difference to what the realist wants to explain.  Nor does it make any difference which model best explains *the phenomena*, assuming this makes objective sense, for this is not what a realist seeks to explain.

*The Theoreticians' Dilemma*:  In an article of this title reprinted in Hempel (1965), Hempel argues that there is, in at least some instances, a third model that is empirically equivalent to UNIFIED and COMP+.  This is a model that eschews the use of theoretical parameters altogether.   COMP+ does not do this because there is a sense in which $\alpha$, $\beta$, and $\gamma$ could be posited by the theory.  In my view, all adjustable parameters are theoretical terms because they are not explicitly defined in terms of observational quantities.  But given the statistical method of estimation that fixes their values, there is a sense in which these parameters could be replaced by their statistics, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$, which are functions of observational quantities (the $x$'s and the $y$'s).  They are complicated functions, but it is possible in principle.  Let us call the resulting model CRAIG.  The replacement does not preserve the meaning of the original model, but it would replicate its empirical success. At first sight, the existence of CRAIG is an embarrassment to a realist because CRAIG cannot represent the theoretical world behind the observed phenomena—for it contains no theoretical terms.   But it should not be an embarrassment, for the realist is committed only to explaining the empirical success of CRAIG, and will use exactly the same explanation used to explain the empirical success of UNIFIED.  CRAIG may bolster the anti-realists' confidence in their position, but it does not weaken the realists' position.

This makes it clear that a realist is not committed to simply "reading off" a realist interpretation from the best scientific explanation.  He is free to read in an interpretation.  So, if inference to the best explanation refers to the best *extant* explanation of the phenomena provided by science, then this ties the realist's hands unnecessarily.  Nor is the realist obliged to interpret every part of a theory representing something in the mind-independent world (Friedman 1981, 1983; Forster 1988).

*Realism in Mathematics*:  Once we see clearly that realism about theoretical entities in science is not premised on their indispensability, then it cuts both ways.  Indispensability is neither a necessary nor sufficient condition for realism.  Thus, the indispensability of abstract mathematical entities in science, even if it is true, is not a good argument for Platonism (Sober 1993).

*Realist Goals versus Predictive Accuracy*:  A concrete example shows that these realist and predictive goals are not only different, but also that optimizing one need not

optimize the other. Imagine that we expand our collection of objects to include 10 different objects, and we do beam balance experiments with the 45 possible pairs of objects drawn from this set. Our beam balance model consists of 45 equations. The mass ratio representation binds them together in a way that is too complicated to display. If we compare UNIFIED and COMPLEX, then the first has 10 adjustable parameters, while the latter has 45. Now suppose that in each of the 45 experiments, the beam is off-center in some random way. This introduces a non-zero value of $\alpha$ into the true equations for the experiment, but the models under comparison do not include this factor. Note that the same value of $\alpha$ applies to all trials of the experiment with a given pair of objects. Also assume that the number of data in each experiment is reasonably large. Which model is better? The answer is that COMPLEX is better for the purposes of prediction if that means predicting new instances *with the beam off-center in exactly the same way*. For the constraint implicit in UNIFIED has the effect of estimating the mass values by averaging over a number of independent measurements, and errors will tend to cancel out. This is good for the purpose of representing the true masses, but bad for the purpose of predicting new instances of data of the same kind (where the "same kind" means with the beam in the same off-center position). Standard model selection criteria, such as AIC, will favor COMPLEX in such a situation, and *correctly so* if the goal is maximize predictive accuracy in this sense. But for other purposes, such as predicting the results of experiments in which the beam is off-center in a *random* way (unrelated to the situation in which the seen data were collected), UNIFIED is better. In addition, UNIFIED is better on realist grounds. What this shows is that the connection between realist goals and predictive accuracy is rather subtle.

*Bayesianism and Other Monolithic Philosophies of Science*: Standard Bayesian philosophy of science is monolithic in the sense that it tries to understand science in terms of a single goal—maximizing the probability of hypotheses. A monolithic philosophy of science will fail to understand the connections between different goals for the simple reason that it defines only a single goal. This limitation applies to the kind of Bayesianism examined by Earman (1992). It does not apply to decision-theoretic versions of Bayesian, such as those defended by Levi (1973) and Maher (1993).

*The Problem of Verisimilitude*: Popper (1968) invented the problem of verisimilitude as a solution to the realist problem of progress: How can we make sense of the fact that (1) the goal of science is truth, (2) science has made progress with respect to this goal, and (3) science consists of a sequence of false theories—Ptolemy, to Copernicus, to Newton. Popper's solution was to define what it means for one false theory to be closer to the truth than another false theory, so that a realist can say that one false theory has made realist progress relative to another. Unfortunately, Tichý (1974) and Miller (1974) proved that Popper's definition does not work. Tichý (1974) introduced a new definition,

which Miller (1974, 1975) criticized as language variant.  Others, such as Oddie (1986) and Niiniluoto (1987) have developed definitions along the lines of Tichý's suggestion, but these theories have not gained universal acceptance (see Niiniluoto (1998) for a recent survey).

One of the reasons that the problem has not received the attention it deserves is that examples tend to be taken from outside of science.  One notable exception is Miller (1975).  In this paper, Miller considers an example of predicting two continuous quantities, and shows how a naïve verisimilitude ordering can be reversed by a coordinate transformation.  Miller the a naïve definition of the accuracy of predictions is language variant.  Miller's problem can be made less abstract and more relevant when it is illustrated in terms of the beam balance example.  Space prevents me from doing that here.  The interesting point is that predictive accuracy (Forster and Sober 1994), which is defined in terms of the Kullback-Leibler discrepancy (Kullback and Leibler 1951), yields a language-invariant definition of verisimilitude and therefore solves the problem presented in Miller (1975) (along the lines suggested by Good 1975).  Alternative concepts, such as van Fraassen's (1980) notion of empirical adequacy do not solve this problem.

While the appeal to predictive accuracy as defined in Forster and Sober (1994) solves Miller's problem, it also leaves out the realist dimension, which is what motivated Popper's definition in the first place.  Nevertheless, it is important to understand that Miller's problem is solved, and to clearly separate the problems that are solved from the problems that are not.

# References:

Akaike, H. (1969): "Fitting Autoregressive Models for Prediction." *Ann. Inst. Statist. Math* **21**: 243-247.

Akaike, H. (1971)  "Autoregressive Model Fitting for Control." *Ann. Inst. Statist. Math* **23**: 163-180.

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.

Akaike, H. (1974): "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control,* vol. AC-19: 716-23.

Akaike, H. (1977): "On the Entropy Maximization Principle." P. R. Krishniah (ed.), *Applications of Statistics:* 27-41.  Amsterdam: North-Holland.

Akaike, H. (1985): "Prediction and Entropy."  In A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics.* New York: Springer. 1-24.

Akaike, H. (1987):  "Factor Analysis and AIC." *Psychometrika* **52:** 317-332.

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

Bell, John S. (1964). "On the Einstein-Podolsky-Rosen Paradox", *Physics* **1:** 195-200.

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

Cartwright, Nancy (1983): *How the Laws of Physics Lie.*  Oxford: Oxford University Press.

Earman, John (1978). "Fairy Tales vs. an Ongoing Story: Ramsey's Neglected Argument for Scientific Realism." *Philosophical Studies 33:* 195-202.

Earman, John (1992):  *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, The MIT Press, Cambridge.

Forster, M. R. (1986):  "Unification and Scientific Realism Revisited."  In Arthur Fine and Peter Machamer (eds.), *PSA 1986.*  E. Lansing, Michigan:  Philosophy of Science Association. Volume **1:** 394-405.

Forster, M. R. (1988):  "Unification, Explanation, and the Composition of Causes in Newtonian Mechanics."  *Studies in the History and Philosophy of Science* **19:** 55 - 101.

Forster, M. 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:** pp. 1 - 35.

Friedman, Michael (1981). "Theoretical Explanation," in *Time, Reduction and Reality.* Edited by R. A. Healey. Cambridge: Cambridge University Press. Pages 1-16.

Friedman, Michael (1983):  *Foundations of Space-Time Theories.* Princeton, NJ: Princeton University Press.

Glymour, Clark (1980). "Explanations, Tests, Unity and Necessity." *Noûs* **14:** 31-50.

Good, I. J. (1975): "Comments on David Miller." *Synthese* **30:** 205-206.

Hacking, Ian (1965).  *Logic of Statistical Inference.* Cambridge: Cambridge University

Press.

Hempel, C.G. (1965), "The Theoreticians' Dilemma", in *Aspects of Scientific Explanation*, N.Y. Free Press.

Hempel, Carl G. (1965): *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.

Hempel, Carl G. (1966): *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Howson, Colin and Peter Urbach (1989): *Scientific Reasoning: The Bayesian Approach*. La Salle, Illinois: Open Court.

Kullback, S. and R. A. Leibler (1951): "On Information and Sufficiency." *Annals of Mathematical Statistics* **22:** 79-86.

Levi, Isaac (1973): *Gambling with Truth; An Essay on Induction and the Aims of Science*. Cambridge, MIT Press .

Lipton, Peter (2004): *Inference to the Best Explanation*. Second Edition. London and New York: Routledge.

Maher, Patrick (1993): *Betting on Theories*. Cambridge: Cambridge University Press.

Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago and London, The University of Chicago Press.

Mill, John Stuart (1874), *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* (New York: Harper & Row).

Miller, David (1974): "Popper's Qualitative Theory of Verisimilitude." *The British Journal for Philosophy of Science* **25:** 166-77.

Miller, David (1975): "The Accuracy of Predictions." *Synthese* **30:** 159-191.

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

Niiniluoto, Ilkka (1987): *Truthlikeness*. Dordrecht: Kluwer Academic Publishing.

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

Oddie, Graham (1986): *Likeness to the Truth*. The University of Western Ontario Series in Philosophy of Science. Dordrecht: Kluwer Academic Publishing.

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

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

Priest, Graham (1976): "Gruesome Simplicity." *Philosophy of Science* **43:** 432 - 437.

Reichenbach, Hans (1938): *Experience and Prediction*. Chicago: University of Chicago Press.

Salmon, Wesley (1984): *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Sober, Elliott (1993): "Mathematics and Indispensability." *The Philosophical Review*. **102:** 35 - 56.

Sober, Elliott (2002): "Two Uses of Unification," in F. Stadler, ed., *Institute Vienna Circle Yearbook 2002*.

Tichý, Pavel (1974): "On Popper's Definitions of Verisimilitude." *The British Journal for the Philosophy of Science*. **25:** 155-160.

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

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