The Positive Economics of Methodology

by

James A. Kahn

Steven E. Landsburg

and

Alan C. Stockman

Suppose that you pick up the latest issue of a scientific journal and find an article with three sections. The first section presents a theory that is consistent with all known facts. The second section presents an entirely new observation, discovered by the author of the article, that is also consistent with the theory. In the third section the author argues convincingly that he was not aware of the new observation at the time when he constructed the theory. Should the third section contribute to the degree of belief that you attach to the theory? Does an observation constitute stronger evidence for a theory if it was made *after* the theory was proposed rather than earlier, when it might have influenced the theory's formation?

To put the issue another way: When a researcher has a body of data at his disposal, he can follow either of two research strategies. The first is to examine only a portion of the data before formulating a theory, and then use the remainder of the data to test the theory. The second is to examine all the data and then construct a theory that fits. We refer to these as the "theorize first" strategy and the "look first" strategy. Under what circumstances, and in what sense, does it matter which strategy the researcher pursues? And how should society structure rewards to scientists to induce them to choose the right strategies?[1]

This paper examines three questions about research strategies which we shall argue are closely related:

We thank many colleagues at Rochester and elsewhere for informal and lively discussions of this topic, and a referee and associate editor at this journal for helping us improve the paper.

[1] Musgrave (1974, 1978) puts the question the following way. Suppose two theorists, X and Y, devise identical theories. X constructs his theory to account for known facts e_1 and e_2, but Y is unfamiliar with recent empirical results and so devises his theory to account for e_1 alone. Does e_2 then lend support to the theory as proposed by Y but not the identical theory as proposed by X? Could a sinister researcher anonymously provide his rivals with new observations so that those observations cannot be used to test their theories? Can researchers be made worse off as their background information grows?

1) Conditional on having a certain theory and a certain body of evidence consistent with that theory, how and why does the research strategy affect the conditional probability that the theory is true?

2) What is the socially optimal set of research strategies?

3) In a world of asymmetric information, what incentive structure can induce that social optimum?

Question 1) has been debated heavily in the philosophy of science literature for over 400 years. In current jargon it is known as the problem of "novel confirmation". Novel evidence for a theory is evidence that is obtained *after* the theory is proposed. In a survey of this literature, Campbell and Vinci (1983) write that

> Philosophers of science generally agree that when observational evidence supports a theory, the confirmation is much stronger when the evidence is 'novel'.

Nevertheless,

> The notion of novel confirmation is beset by a theoretical puzzle about how the degree of confirmation can change without any change in the evidence, hypotheses, or auxiliary assumptions...There have not yet appeared any obviously satisfactory solutions to these problems.

In the philosophical literature, the novel confirmation problem is frequently addressed in the following way. First one writes down an expression for the conditional probability that a theory is true, given certain observational evidence and other ambient conditions. Then one manipulates this expression via repeated applications of Bayes's Law, reducing it to a formula involving auxiliary expressions whose dependence on the novelty of the evidence is either asserted or denied. To an economist, it is striking that these conditional probabilities are manipulated with abandon despite the absence of any kind of well-defined sample space, random variable, or other mathematical paraphernalia that might be considered prerequisites for the probability calculations to be meaningful.

Therefore it is our position that question 1) cannot be satisfactorily addressed in the absence of an explicit model of the process by which theories are generated. In this paper we will provide such a model. Because our model contains scientists who are maximizing agents in possession of private information, the outcome depends heavily on the incentives that the scientists face. This leads us to the conclusion that questions 1), 2) and 3) are intricately intertwined, and that a satisfactory treatment of any must entail simultaneous investigations of the others.

The most rudimentary model that we can construct, presented in Section 1, implies that research strategy does not matter. This goes counter to the most common intuition. Therefore in sections 2 and 3 we embellish the model to provide a justification for the widespread belief that when researchers have theorized first, their theories are more likely to be true. In Section 4 we examine the model for relevance to the real world. The thrust of the model is that scientists can use their choice of research strategy as a way to signal private information about their abilities. We do not claim that this is the only possible model, or even the most plausible one. In fact, in Section 5 we outline several alternative approaches that we think could lead to the same conclusion. What we *do* claim is to have provided the first full-fledged model for addressing the novel confirmation problem in terms that an economist would consider intelligible. How close it is to the *right* model we do not know. We hope, however, that it is the starting point for bringing rigorous economic reasoning to bear on the problems of scientific method.[2]

1. A Simple Model.

We begin with a simple model of the scientific process. A theory is a set of statements that predicts that under certain circumstances, certain events will occur. We call a theory *true* if its predictions are always accurate.[3]

Imagine a single scientist working on a problem, with time to make one new observation and construct one new theory. The observation may result from an experiment. For now, we take the choice of experiment performed as exogenous; Section 3 discusses the relaxation of this assumption. Past experiments and previously constructed theories guide the scientist in ways to be discussed below.

Assume that theories can be divided into four mutually exclusive sets. Type A theories are true. Type B theories are false, but consistent with all past observations *and* the new observation. (However, type

[2] Our companion paper (Kahn, Landsburg and Stockman 1992) discusses the relations between our economic analysis the philosophy literature.

[3] Obviously, we are abstracting from several things here. The reader who is squeamish about the metaphysical implications of the word *true* can substitute a notion of *usefulness*; our requirement that predictions are always verified can be replaced by a requirement that predictions are verified with high probability, or come close to being verified in some appropriate sense; theories might predict not actual events but probability distributions over events; and so forth.

B theories are liable to be falsified by future observations not yet made.) Type C theories are false and consistent with all past observations, but *not* with the new observation. Type D theories are false and inconsistent with past observations. Constructing a theory involves some fundamental randomness, so we model it as the selection of a ball from an urn, with the balls representing possible theories. We assume that researchers are able to avoid type D theories by "building in" established observations; thus the urn contains no type D balls. It contains balls of types A, B and C in some proportions p, q, and $1-p-q$.

We suppose that the scientist either *theorizes first* or *looks first*, and that his choice of strategy is exogenous. (Later sections will relax this exogeneity.)

Suppose first that the scientist *looks first*, beginning with a new observation and then choosing a theory that is consistent with it. We view this scientist as removing all type C balls from the urn prior to his drawing. So the probability that his theory is true is $p/(p+q)$.

Suppose instead that the scientist *theorizes first*, constructing a new theory and then making an observation *that subsequently proves to be consistent with his theory.* The unconditional probability that the theory is true (type A) is p. If it is subsequently found to be consistent with the new observation, then the theory is either type A or B. So the probability of truth *conditional* on consistency with the new observation is $p/(p+q)$.

Our first result, then, is a *neutrality result*. In this model, *the probability that a theory is true, conditional on its being consistent with all (old and new) observations, is independent of the research strategy.* This result is hardly surprising. Nothing in classical statistical inference allows the timing of an observation to matter. The data themselves are a sufficient statistic.

This neutrality result strikes many scientists and philosophers as plainly wrong. The question then is: What new elements can be added to the model to break this result and justify the widespread intuition that theorizing first is better than looking first? In Sections 2, 3 and 4 we will provide a detailed development of one answer to this question. In Section 5 we will outline some alternative approaches that we believe could be equally fruitful.

2. Heterogeneous Scientists.

Now assume that there are two types of scientists, type i and type j. No one knows a given scientist's type (except *possibly* the scientist himself), but everyone knows that the population of scientists consists of a proportion i who are type i and a proportion $j = 1 - i$ who are type j. Scientists differ only in the compositions of the urns from which they draw, as indicated in the following table:

Theory Type	Type i Scientist	Type j Scientist
A	p	r
B	q	s
C	$1-p-q$	$1-r-s$

We make the following assumptions throughout the paper:

$$p + q > r + s \tag{2.1}$$

$$\frac{p}{p+q} > \frac{r}{r+s} \tag{2.2}$$

Either equation (2.1) or (2.2) by itself is essentially vacuous, since types i and j can always be redefined to make it true. However, in tandem they state that type i researchers are *both* more likely to have their results confirmed when they theorize first, *and* more likely to construct true theories when they look first, than are type j researchers. Thus the equations assert roughly that there is a positive correlation between two different types of scientific ability.

Consider a scientist who *looks first* and then constructs a theory consistent with his observation. If he is type i, he constructs a theory of type A with probability $p/(p+q)$, and if he is type j, he constructs a theory of type A with probability $r/(r+s)$. Any observer (including the scientist) who does not know the scientist's type calculates that the theory is true with probability

$$\gamma = i \cdot \frac{p}{p+q} + j \cdot \frac{r}{r+s}. \tag{2.3}$$

Suppose instead that the scientist *theorizes first* and produces a theory that survives testing (i.e. is consistent with the new observation). This conveys information about the scientist's type. The updated probability that the scientist is type i is given by the expression

$$i' = i \cdot \frac{p+q}{i \cdot (p+q) + j \cdot (r+s)}. \tag{2.4}$$

Condition (2.1) implies that $i' > i$. Let $j' = 1 - i'$ be the updated probability that the researcher is type j. Then the theory is true with probability

$$\gamma' = i' \cdot \frac{p}{p+q} + j' \cdot \frac{r}{r+s}. \tag{2.5}$$

Since $i' > i$, condition (2.2) guarantees that $\gamma' > \gamma$. Thus the neutrality result of Section 1 is overturned:

Assume that the scientist's type is unknown. Then the probability that the theory is true, conditional on its being consistent with all (old and new) observations, is higher if the scientist theorized first than if he looked first.

The difference between expressions (2.3) and (2.5) measures the extent to which the reader should discount evidence that was examined by the scientist prior to theorizing. It is worth remarking that this "pretest discount" depends on the parameters i, j, p, q, r and s, which describe not just the circumstances of a particular research project, but the characteristics of the scientific community as a whole.

3. A Social Planner's Problem.

The preliminary models of Sections 1 and 2 take the scientist's choice of research strategy to be exogenous. Although this is consistent with the standard treatments by philosophers, we do not believe that the exogeneity assumption is a plausible basis for a complete model. Indeed, we view the tacit maintenance of this assumption as a major failing of the philosophy literature.

As a first step toward endogenizing the research strategy decision, we consider the problem faced by a social planner who can mandate rewards for scientists as a function of their research strategies and the outcomes of their research. Assuming that the planner maximizes welfare, we will determine the mix of research strategies that is adopted and the information conveyed by those research strategies. It is natural to ask whether the planner's optimum can be supported as a decentralized competitive equilibrium. The answer is no; we will elaborate on this point in Section 4.

Therefore our model can be interpreted in either of two ways, depending on the reader's predilections. Those who believe that social institutions generally evolve to implement optimal outcomes can interpret it as a positive model of how the scientific establishment sets rewards, and how scientists respond. Those who believe otherwise can interpret it as a normative model to be used as a benchmark for measuring the failure

of the scientific establishment, and as a guide to future policy.

We want our model to incorporate the value of believing a true theory and of disbelieving a false one. Our device for doing so is to assume that the planner uses the theories to guide the construction of bridges. Each bridge requires a separate theory. True theories produce bridges that stand and false theories produce bridges that fall. If the bridge stands, the planner receives utility $G > 0$; if it falls, he receives utility $L < 0$. If he chooses not to build the bridge, the planner receives zero utility.

It turns out that our analysis is somewhat dependent on the sign of $rG + sL$, the expected value of a bridge built by a type-j researcher. In our earlier working paper (Kahn, Landsburg and Stockman, 1991) we worked out both cases; here for brevity we will simply assume that $rG + sL < 0$.

We want also to incorporate the notion that scientists can have private information about their ability to produce true theories. Among other things, this information can take the form of confidence in one's intuitive grasp of a particular problem. For this purpose, we adopt the model of Section 2, with two types of scientists. Scientists know their own types, but the planner knows only the fraction of each type of scientist in the population.

The planner is permitted to reward scientists for their research, with separate rewards for three possible outcomes:[4,5]

a) a reward y^L for a theory produced under a "look-first" strategy (hence known to be of type A or B).

b) a reward y^T_{AB} for a theory produced under a "theorize-first" strategy and confirmed by observation

[4] An alternative mechanism would be to make payments contingent on whether the bridge stands or falls. By ruling this out, we capture the notion that the social benefits of a true theory often accrue too late to affect the compensation of the original scientist. In our earlier working paper (Kahn, Landsburg and Stockman, 1991) we did allow rewards contingent on whether the bridge stands or falls. Coupled with an auxiliary assumption that the planner cannot make negative payments to scientists, this yielded essentially the same conclusions that we will arrive at here.

[5] Our mechanism deviates from the standard revelation procedure in which scientists announce their types, and rewards are contingent on those announcements as well. It turns out that incorporating this additional instrument complicates matters without materially changing the outcome. Therefore we rule it out for the time being, for purely expositional reasons. In Section 3.6 below we summarize the effects of including it.

(hence known to be of type A or B).

c) a reward y_C^T for a theory produced under a "theorize- first" strategy and rejected by observation (hence known to be of type C).

Finally, we assume that scientists (or potential scientists) are risk-neutral, that they have alternative uses of their time, and that labor-supply schedules to the scientific research industry reflect the social opportunity costs of research. For concreteness, we assume that the supply curve of type-i scientists is linear with slope $\alpha > 0$, so that an expected wage of αN calls forth N type-i scientists. Similarly, the supply of type-j scientists is linear with slope $\beta > 0$.

To maximize welfare, the planner can proceed as follows. Consider four possible scenarios. First, all scientists could theorize first. Second, type-i scientists could theorize first while type-j scientists look first. Third, type-i scientists could look first while type-j scientists theorize first. Fourth, all scientists could look first. Call these scenarios TT, TL, LT and LL respectively. For each of the four scenarios, one can write down an expression for social welfare as a function of the rewards y^L, y_{AB}^T and y_C^T. Now maximize the TT welfare expression over the set of triples (y^L, y_{AB}^T, y_C^T) that actually elicit TT-behavior. Do the same for the other three scenarios. Finally, maximize over the four resulting welfare expressions.

We will provide a detailed analysis for two of the four scenarios, and state the (similarly derived) results for the others.

3.1. Scenario LL. Under this scenario, the values y_{AB}^T and y_C^T might as well be set to zero (they need only be sufficiently small so that nobody will choose to theorize). The important number is y^L, which is what all researchers earn. Figure 1 illustrates this wage and the number of researchers of each type that it draws forth.

Because all researchers look first, the expected social value of a type-i theory is

$$\frac{p}{p+q} \cdot G + \frac{q}{p+q} \cdot L$$

while the social value of a type-j theory is

$$\frac{r}{r+s} \cdot G + \frac{s}{r+s} \cdot L$$

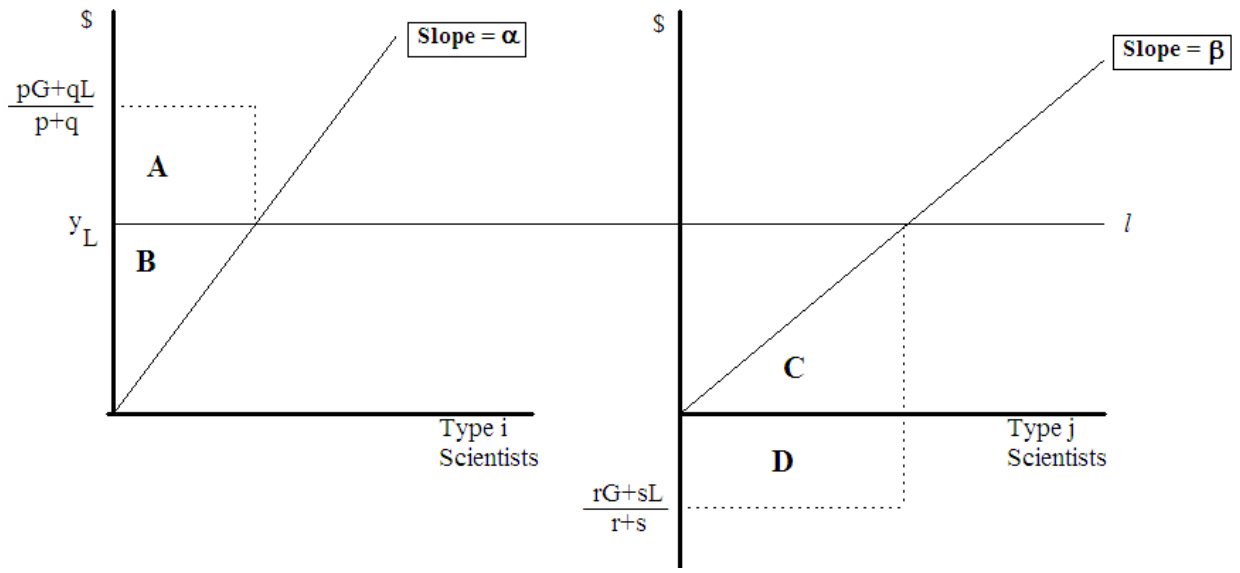

Figure 1

(which we have assumed to be negative). In Figure 1, this means that type- i research creates a net social gain of areas $A + B$ while type-j research creates a net social loss of areas $C + D$. Thus the planner chooses y_L to maximize the difference $(A + B) - (C + D)$.[6]

Note that the social loss in the type j market could be reduced from $C + D$ to just C if there were a way to identify the type-j scientists so that their theories could be discarded before bridges were built.

3.2. Scenario TL. Here the rewards must satisfy certain constraints. A type-j scientist who theorized would earn an expected reward of $(r+s)y_{AB}^T + (1-r-s)y_C^T$. In order to ensure that type-j scientists choose to look first, we must have

$$(r + s) \cdot y_{AB}^T + (1 - r - s) \cdot y_C^T \leq y^L. \tag{3.1}$$

At the same time, to insure that type-i scientists theorize first, we must have

$$(p + q) \cdot y_{AB}^T + (1 - p - q) \cdot y_C^T \geq y^L. \tag{3.2}$$

[6] The picture is drawn with the assumption that y_L is both positive and strictly less than the value of a type-i theory; it is easy to verify that these conditions hold at the optimum.

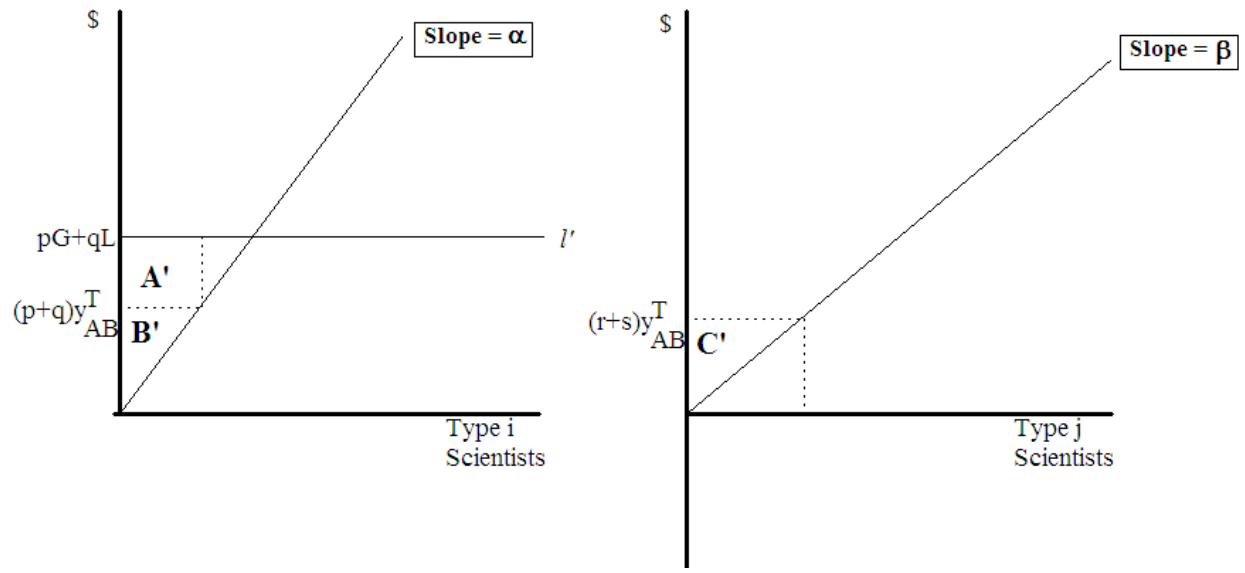

Figure 1

The only values that matter are the value of y^L and the value of the left side of (3.2). The latter value can be maintained, without upsetting (3.1), by setting $y_C^T = 0$ and adjusting y_{AB}^T accordingly. Thus we might as well assume that $y_C^T = 0$, and the constraints become

$$(r + s) \cdot y_{AB}^T \leq y^L. \tag{3.1′}$$

$$(p + q) \cdot y_{AB}^T \geq y^L. \tag{3.2′}$$

It is easy to verify that the optimal rewards are achieved when (3.1′) is taken as binding and the planner maximizes with respect to y^L.

Figure 2 shows the number of scientists of each type called forth by the optimal rewards. The *ex ante* value of a type-i theory is now $pG + qL$ (because type C theories are detected by observation and not used to build bridges). Type-j theories, which if implemented would yield bridges with negative expected value, are not used to build bridges and therefore have social value zero. In Figure 2, the net social gain is the difference of areas $(A' + B') - C'$.

3.3. Looking versus theorizing. Experiments with parameter values reveal that either the LL-optimum or the TL-optimum could dominate the other. Before exploring the planner's other options, we

pause to comment on the nature of the trade-offs.[7]

"Theorize-first" research is wasteful in the sense that it sometimes produces type C theories that are rejected by the evidence. In terms of the Figures, this means that line ℓ' in Figure 2 is situated lower than line ℓ in Figure 1. In the absence of any informational problems regarding scientists' abilities, it would always be optimal to have everybody look first.

The offsetting advantage of the TL regime is that it allows the planner to separate scientists by type. This information is valuable for two quite separate reasons:

1) The planner is able to avoid building those bridges (designed by type-j scientists) whose expected value is negative. That is, under TL there is no loss analogous to area D in Figure 1.

2) Under TL, the two types of scientists receive different expected compensations. The planner can then come closer to paying each scientist his marginal value. Consequently area $A' + B'$ in Figure 2 can be larger than area $A + B$ in Figure 1, and area C' in Figure 2 can be smaller than area C in Figure 1. These inequalities certainly hold, for example, when $p + q = 1$.

Either of these considerations could be sufficient to justify the cost of separation. For example, suppose that instead of assuming $rG + sL < 0$, we had assumed $rG + sL > 0$. In that case the planner would build all bridges even in the TL case and advantage (1) would no longer apply. Nevertheless, for appropriate parameter values TL can still dominate LL, entirely in consequence of advantage (2).

On the other hand, suppose we modify the model so that scientists of each type are supplied perfectly inelastically. In this case, advantage (2) disappears entirely, but the theorize-first regime can still dominate the look-first regime if $rG + sL$ is sufficiently negative, which increases the size of advantage (1).

The bottom line then is that either regime might be preferred. LL is preferable when the costs of

[7] Area $(A + B) - (C + D)$ is equal to

$$\frac{\big(\beta(pG + qL)(r + s) + \alpha(rG + sL)(p + q)\big)^2}{2\alpha\beta(\alpha + \beta)(p + q)^2(r + s)^2}$$

and area $(A' + B' - C')$ is equal to

$$\frac{\beta(p + q)^2(pG + qL)^2}{2\alpha\big(\beta(p + q)^2 + \alpha(r + s)^2\big).}$$

It is these two expressions that must be compared to determine the planner's optimum.

theorizing first are high (as when $p+q$ is small); TL is preferable when the benefits of separation are high (as detailed under advantages (1) and (2) above).

3.4. Other regimes. Although we can make no unambiguous statement about the choice between LL and TL, we can say something about the other two alternatives. First, it is easy to prove that TL unambiguously dominates TT. Intuitively, the TL regime fully separates type i and type j researchers, while TT does not. TT does achieve a partial separation in the sense that more type-j than type-i theories are rejected. (And this advantage is sufficient so that for appropriate parameter values TT can dominate LL.) However, the partial separation is inferior to the full separation that occurs under TL. In addition, TT completely sacrifices the second advantage of separation, namely the ability to compensate the two types of scientists at different levels. Thus TT has exactly the same costs as TL, but only some of the benefits.

The LT scenario cannot be supported. The incentive- compatibility conditions for this scenario are

$$y^L \geq p \cdot y^T_{AB} + q \cdot y^T_C$$

$$r \cdot y^T_{AB} + s \cdot y^T_C > y_L$$

Together with our other assumptions, these imply that $y^T_C > y^T_{AB}$. But this creates an incentive for scientists to falsely report that their theories have been rejected, which we take as a fatal flaw.

3.5. Conclusions. The social optimum is either of type LL or type TL. This means that type-j scientists always look first, whereas type-i scientists might either look or theorize first, depending on parameter values.

3.6. Additional revelation mechanisms. If it is advantageous to do so, the planner might want to make rewards contingent on information other than research strategy and confirmation by observation. The revelation principle tells us, however, that without loss of generality we can assume that he simply requires scientists to announce their types (although he cannot verify directly that the announcements are truthful). Based on their announcements, scientists are directed to perform either "look-first" or "theorize-first" research. Rewards are then contingent on both the scientist's announced type and (in the case of those who theorize first) on whether the theory is confirmed by observation. This would be the standard approach

in theliterature on mechanism design, and we have deferred it to the end of this section only to facilitate exposition. However, the results are essentially the same.

One new issue that arises is how type-j scientists behave when they are indifferent about revealing their true types.[8] If we assume that they do *not* reveal their types when indifferent, then the analysis is exactly as in Sections 3.1–3.5. If we assume that they *do* reveal their types then there are a few new observations:

1) The desirability of LL is enhanced, because theories of type-j researchers are no longer used to build bridges. This means that in Figure 1, the social gain is $A+B-C$ rather than $A+B-(C+D)$. (Also of course, the areas themselves change because the optimal reward values change.) Nevertheless, it remains the case that either the TL or LL regime can dominate the other.[9]

2) TT is no longer inferior to TL; it is now equivalent. The basic reason is that in either scenario, theories of type-j researchers are not used to build bridges, so their research strategies do not matter.[10]

3) LT can now be supported and is equivalent to LL. The intuition is as in (2): theories of type-j researchers are not used for bridges, so their research strategies do not matter.

3.7. Novel confirmation. We can now return to the question raised in the first paragraph of this paper. Is a theory more likely to be true by virtue of having been produced under a theorize-first regime?

Given full knowledge of the scientist's abilities, the model of Section 1 implies that the answer is no. The modification in Section 2 implies that in the presence of uncertainty about scientist's abilities, the answer becomes yes; however that model takes research strategy to be exogenous and we do not consider it fully

[8] In the LL case, small increases in rewards to announced type-j scientists designed to elicit truth from them are ineffective because type-i scientists would then claim to be type-j.

[9] In this situation the value of $A + B - C$ is

$$\frac{\beta(pG + qL)^2}{2\alpha(\alpha + \beta)(p + q)^2}$$

which is to be compared with the value of $A' + B' - C'$ from an earlier footnote.

[10] The preceding argument is not entirely conclusive since incentive compatibility constraints can change across regimes, but it is not difficult to show that the conclusion holds.

satisfactory.

In the context of our complete model, we can restrict ourselves to the case where the planner implements the TL outcome, since in the LL outcome nobody theorizes first and the question is vacuous. In TL, the planner learns everybody's type. This renders the question ambiguous. We must specify whether or not we are allowed to condition on that revealed information.

Conditioned on knowledge of the scientist's type, the answer remains no, as in Section 1. But if we interpret the question to mean "can we infer anything about the researcher's type from knowledge of his research strategy?" then the answer becomes yes, and in a much stronger sense than when research strategies are exogenous. In Section 2, successful novel confirmation raised the probability that a scientist is type i. In Section 3, the scientist's decision to theorize first raises the probability that he is type i (in fact raises it to 1) regardless of whether his theory is confirmed.

This is worth emphasizing: There are *two independent reasons* why a theorize-first theory has an enhanced probability of truth. One is that we can infer something about scientific ability from a successful *ex post* confirmation. The other, in the presence of a satisfactory theory of incentives, is that we can infer something about scientific ability from the fact that the scientist was willing to risk theorizing first at the outset.

4. The Model Versus Reality.

Our model is highly stylized and dependent on a number of arbitrary assumptions. It is intended as an *example* of an economic approach to the problem of methodology, not as a definitive solution. Nevertheless, the model calls attention to a number of issues that are worth examining for their importance in real life. We think the following aspects of the model bear consideration:

4.1. Sorting by type. The model predicts that in some circumstances scientists are sorted by type, with type-i scientists (those with the best chance of producing true theories) theorizing first and receiving payments contingent on whether their theories are confirmed, while type-j scientists look first and receive flat salaries for their efforts.

Perhaps the easiest way to envision this sorting mechanism is in the choice of academic jobs. Some people choose jobs at research institutions where successful research is heavily rewarded, while others choose

jobs at institutions where the rewards are relatively independent of the quality or success of their research. The model implies that those in the latter group produce low-quality research; moreover it implies that on average they have lower opportunity costs than those in the former group.

4.2. Support for wasteful research. In the TL outcome, there is a group of scientists who are paid to produce theories that are known *in advance* to have no social value; that is, type-j scientists are identified at the outset and everybody knows that their theories will never be used. Nevertheless, they must receive salaries as an inducement to confess to their limited abilities so that their theories can be discarded. Otherwise they would claim to be type i and too many bridges would fall down.

Because the type-j scientists have opportunity costs, social welfare could be improved by paying them to reveal their types and then freeing them from the obligation to actually perform research. Unfortunately, in this circumstance *everyone* would claim to be a potential researcher in order to claim the reward, necessitating massive lump-sum transfers that we are prepared to assume are infeasible. So the planner must accept the social loss inherent in requiring all declared "researchers" to actually produce theories.

There is one partial solution to this dilemma: Type-j agents might be able to perform some socially valuable function *at a research institution* (for example, teaching undergraduates). The planner can require presence at the research institution as a requisite for the rewards while still assigning type-j scientists to do something valuable; this improves the welfare computations in Section 3. The planner must still be able to verify that declared type-js do not simultaneously engage in other privately rewarding activities; thus he might want to limit consulting income, for example.

4.3. Competitive equilibrium. The TL optimum requires that identifiably unproductive researchers be compensated; this outcome cannot be supported in competitive equilibrium. We have worked on the problem of competitive equilibrium in a model like the present one and plan to report our results in a subsequent paper.

4.4. Suboptimality. In either the TT or the TL outcome, there are too many type-j (unproductive) researchers and too few type-i (productive) researchers relative to the full-information first-best. Regarding conformity to real-world conditions, we feel that no further comment is required.

4.5. Verifying research strategies. We have assumed that scientists cannot make false claims about

their research strategies. If instead research strategies are unverifiable, then the planner is essentially forced to the LL outcome because researchers will not risk theorizing first (and the chance of a type-C theory) if unrejected theories are worth more to them than rejected theories. In those cases where the TL outcome would dominate, the difference between the welfare calculations in Figures 1 and 2 is a measure of how much society should be willing to pay to develop mechanisms to monitor research strategies.

We believe that some such mechanisms are already in place. In forums from the seminar room to the lunch table, scientists quiz each other about what they are up to. It would be very difficult completely to delude one's colleagues about the nature of one's research strategy.[11]

5. Other Considerations.

Our model is intended to illustrate that it is possible to approach the problem of novel confirmation in an economically rigorous way. This opens the door for the development of alternative models that will call attention to other reasons why we care whether scientists theorize before or after examining their data. In this section, we list some of the phenomena that we think should be incorporated in future models. In some cases, we have provided the rudiments of those models.

5.1. Theories suggest experiments. Our model assumes that there is only one observation that a scientist can make, and that he makes the same observation regardless of whether the observation precedes or follows the theory. More realistically, researchers have a choice of several experiments to perform, and the theory can help to suggest which one is most useful.

Consider a single researcher drawing from an urn with two types of theories, all consistent with past observations. Type T theories are true and type F theories are false. Type T theories suggest an experiment E_T, which tests an implication of type T theories, while type F theories suggest an experiment E_F, which

[11] In practice, too, scientists might have only a limited ability to choose their strategies. Scientific training in many fields (economics included) encourages individuals to commit early in their careers to becoming either a theorist or an applied scientist. Although one occasionally hears complaints about theorists' distance from the real world (or empiricists' distance from the theoretical one), the research market does not seem to discourage this sort of specialization. While standard arguments about the gains from specialization can account for this, it bears mentioning that our analysis suggests an additional reason. When theorizing first is a signal of ability, it can be useful to separate theorizers from lookers at the outset. Type-i people become theorizers, prohibited from learning how to do empirical work. This eliminates the problem of verifying that certain theories were arrived at without previous analysis of the data.

tests an implication of type F theories. Because type T theories are true, they are never rejected by experiment. Type F theories are rejected by experiment with probability ρ. A social planner who maximizes expected utility gains G from a true theory, $L < 0$ from a false theory, and zero from no theory at all.

A researcher who theorizes first constructs a type T theory with probability p and a type F theory with probability $q = 1 - p$. So his theory is true with probability p, not true but not rejected with probability $q \cdot (1 - \rho)$, and rejected with probability $q\rho$. The planner's expected utility is then

$$\max\{p \cdot G + q \cdot (1 - \rho) \cdot L, 0\} \tag{5.1}$$

A researcher who looks first chooses randomly between the two feasible experiments. Let $1 - \pi$ and π denote the probabilities of choosing experiment E_T and experiment E_F. If he chooses E_T, his observation is consistent with both theories; he chooses a type T theory with probability p. If he chooses E_F, then with probability ρ the observation rules out type F in which case he constructs a true theory with probability 1. With probability $1 - \rho$ the observation from E_F fails to rule out any theory, in which case the researcher constructs a true theory with probability p. So the *a priori* probability that he constructs a true theory is

$$(1 - \pi) \cdot p + \pi \cdot \rho + \pi \cdot (1 - \rho) \cdot p = p + q \cdot \pi \cdot \rho. \tag{5.2}$$

The probability that he constructs a false theory is $1 - (p + q\pi\rho) = q(1 - \pi\rho)$. So the planner's expected utility if the researcher looks first is

$$\max\{(p + q \cdot \pi \cdot \rho) \cdot G + q \cdot (1 - \pi \cdot \rho) \cdot L, 0\}. \tag{5.3}$$

Several points are worth noting. First, as in all our previous models, the probability of constructing a true theory is greater if the scientist looks first. Looking provides information that can help to select among theories. On the other hand, the probability of constructing an unrejected false theory is *also* greater under looking first because the chosen experiment is not expected to be as informative.

Interestingly, it is possible for non-novel evidence to raise the probability that a theory is true by more than the same evidence would if it were novel. The conditional probability that a theory is true if it is consistent with all observations and was produced by the theorize-first strategy is

$$\frac{p}{p + q \cdot (1 - \rho).} \tag{5.4}$$

The conditional probability that a theory is true if it is consistent with all observations and was produced by the look-first strategy is

$$p + q \cdot \pi \cdot \rho \tag{5.5}$$

which can be larger or smaller than (5.4) depending on parameter values. So when theories suggest experiments, the research strategy does affect the conditional probability that a theory is true, but not always in the direction of favoring the theorize-first strategy.

It is remarkable that neither p nor ρ plays any role in determining which strategy is socially optimal. In fact theorizing dominates looking if and only if $G/|L| < (1-\pi)/\pi$. A smaller (absolute) value of L relative to G makes looking first more attractive, whereas theorizing first will be preferred when failure is very costly. This accords with the intuition that if the primary goal is to come up with a true theory (*i.e.* $G/|L|$ is big), then one should first look at all the data, whereas if one is more concerned about not believing a false theory ($G|L|$ is small), one should theorize first. Finally, a lower value of π makes theorizing first more attractive because there is a smaller value to choosing an experiment without the guidance of the new theory. The basic point is that in a setting where the costs of proceeding on the basis of a bad theory are high relative to the benefits from successful research, theorizing first should be encouraged, whereas if the benefits from success are high relative to the costs from a mistake, looking first is preferred.[12]

5.2. Experiments suggest theories. Our models assume that the act of making an observation (discarding type C balls from the urn) does not affect the proportion of true theories to those that are false but consistent with the data (*i.e.* the ratio p/q or r/s.) Yet it seems quite reasonable to argue that the act of making an observation could hone a scientist's intuition in a way that causes these proportions to change. We do not know how to incorporate this effect into a model without introducing assumptions that are substantially more *ad hoc* than those we have resorted to already. But we believe this is an issue that

[12] An application of this principle might be the debates over activist discretionary macroeconomic policy versus fixed rules. Proponents of rules argue that discretionary policies aimed at "fine-tuning" have small potential benefits and large potential costs, so our theory predicts that they advocate theorizing first. Proponents of activist policy believe the cost of doing "nothing" is significant, while the risks are small, so our theory predicts that they advocate looking first. Perhaps this explains the contrast between the macroeconomic research programs of researchers such as Robert E. Lucas and Christopher Sims.

bears much further thought.

5.3. Sampling error. We have assumed that observations lead to facts, and that facts are consistent with true theories and inconsistent with some false theories. But experimental errors or sampling errors in statistical inference may create observations that are inconsistent with true theories and consistent with a different subset of false theories. So (in our model of theory-construction), sampling or experimental error implies that looking first may remove some type A balls from the urn and does not necessarily remove all the type C balls from the urn. This creates a cost and removes a benefit of looking first.

Suppose, for concreteness, that we subdivide the set of type C theories into subclasses $C_1, ..., C_N$ where N is large and assume that a researcher who constructs a type C theory is equally likely to choose any of them. Also suppose there are N different ways to misinterpret the experiment, with the n^{th} misinterpretation consistent *only* with theories of type C_n. Each misinterpretation is equally likely, so each has probability $\frac{(1-p-q)}{N}$. Finally, suppose there is only one type of researcher, who does the experiment correctly with probability i and misreads it with probability $1-i$.

Now suppose a researcher theorizes first. With probability $i \cdot (p+q)$ he constructs a theory of type A or B and correctly interprets his experiment as confirming it. With probability $\frac{(1-i)\cdot(1-p-q)}{N^2}$ he constructs a theory of type C_n and misinterprets his experiment so as to support the theory. There are N ways to do this, so the probability that the theory and experiment are compatible is then $i \cdot (p+q) + \frac{(1-i)\cdot(1-p-q)}{N}$. Therefore, if a researcher theorizes first and the new observation confirms the theory, the updated probability that he has interpreted the experiment correctly is

$$i' = \frac{i \cdot (p+q)}{i \cdot (p+q) + \frac{(1-i)\cdot(1-p-q)}{N}}.$$

The probability that the theory is true is $\frac{i' \cdot p}{(p+q)}$. Notice that i' approaches one as N becomes large. This is because it becomes increasingly unlikely that a researcher who theorizes first and constructs a false theory would misinterpret his experiment in precisely the right way as to support his theory. So for large N, the probability that a theory is true if the researcher theorizes first and the new observation is consistent with the theory exceeds $\frac{i \cdot p}{(p+q)}$, which is the probability that a theory is true if the researcher looks first. That is, even if all scientists are alike, the possibility of sampling error or experimental error can overturn the

neutrality result of Section 1.

5.4. A priori criteria. Theories are judged not just by their consistency with the data but by their inherent plausibility. We believe this feature would be easy to incorporate into our existing model. Consistency with an *a priori* criterion of believability can be treated formally just like consistency with another observation.

5.5. Multiperiod models. In a multiperiod model, information about a scientist's type that is gained in one period can be useful in subsequent periods. This raises the value of theorizing for young scientists and lowers it for old scientists. Casual empiricism suggests that scientists may do more theorize-first type research in their early years. Can this consideration provide an explanation?

5.6. Paths to truth. We have introduced an important abstraction by distinguishing between true and false theories, and assuming that a significant proportion of theories are true. Perhaps a more realistic setup would be to suppose that true theories are infinitely rare, but some false theories are closer to true than others. "Truth" would be a bliss point in some space of theories and the scientific process could be viewed as tracing out a path in theory-space. Research strategies would guide the progress of that path. We do not know whether this approach would yield significant insights beyond the models we have already provided.

References

Richmond Campbell and Thomas Vinci, "Novel Confirmation", *The British Journal for the Philosophy of Science* 34 (1983), 315-341.

James A. Kahn, Steven E. Landsburg and Alan C. Stockman, "The Positive Economics of Methodology: A Theory of Incentives and Methods in Scientific Research", Working paper available from the authors (1991). See also earlier version, "The Positive Economics of Methodology", Technical Working Paper No. 82, The National Bureau of Research (1989).

James A. Kahn, Steven E. Landsburg and Alan C. Stockman, "On Novel Confirmation", *British Journal for the Philosophy of Science*, forthcoming 1992.

A. Musgrave, "Logical Versus Historical Theories of Confirmation", *British Journal for the Philosophy of Science* 22 (1974).

A. Musgrave, "Evidential Support, Falsification, Heuristics and Anarchism" in *Progress and Rationality in Science* (Radnitzky and Andersson, eds.), Dordrecht: Reidel (1978).