Ronald N. Giere

Testing Theoretical Hypotheses

1. Introduction

Philosophers of science concerned with theories and the nature of evidence tend currently to fall into several only partially overlapping groups. One group follows its logical empiricist ancestors at least to the extent of believing that there is a "logic" in the relation between theories and evidence. This logic is now most often embedded in the theory of a rational (scientific) agent. Bayesian agents are currently most popular, but there are notable dissenters from The Bayesian Way such as Henry Kyburg and Isaac Levi. Another group derives its inspiration from the historical criticisms of logical empiricism begun a generation ago by such writers as Gerd Buchdahl, Paul Feyerabend, N. R. Hanson, Thomas Kuhn, and Stephen Toulmin. Partly because their roots tend to be in intellectual history, and partly in reaction to logical empiricism, this group emphasizes the evolution of scientific ideas and downplays the role of empirical data in the development of science. For these thinkers, the rationality of science is to be found in the historical process of science rather than in the (idealized) minds of scientists. If there is something that can rightfully be called a middle group, it consists mainly of the followers of the late Imre Lakatos, who skillfully blended Popper's version of empiricism with elements of Kuhn's account of scientific development. Yet Lakatos's "methodology of scientific research programmes" also locates the ultimate rationality of science in a larger historical process rather than in relations between particular hypotheses and particular bits of data.

I shall be arguing for a theory of science in which the driving rational force of the scientific process is located in the testing of highly specific theoretical models against empirical data. This is not to deny that there are elements of rationality throughout the scientific enterprise. Indeed, it is only as part of an overall theory of science that one can fully comprehend what goes on in tests of individual hypotheses. Yet there is a "logic" in the

The author's research has been supported in part by a grant from the National Science Foundation.

parts as well as in the whole. Thus I agree with contemporary students of probability, induction, and the foundations of statistics that the individual hypothesis is a useful unit of analysis. On the other hand, I reject completely the idea that one can reduce the rationality of the scientific process to the rationality of individual agents. The rationality of science is to be found not so much in the heads of scientists as in objective features of its methods and institutions.

In this paper I shall not attempt even to outline an overall theory of science. Rather, I shall concentrate on clarifying the nature of tests of individual hypotheses, bringing in further elements of a broader theory of science only when necessary to advance this narrower objective. My account of how individual hypotheses are tested is not entirely new. Indeed, it is a version of the most ancient of scientific methods, the method of hypothesis, or, the hypothetico-deductive (H-D) method. But some elements of the account are new, and some have been borrowed from other contexts.

2. Models, Hypotheses, and Theories

Views about the nature of evidence and its role in science depend crucially on views about the nature of hypotheses and theories. The major divergences of current opinion in the philosophy of science are correlated with strong differences as to just what the highly honorific title "theory" should apply. For the moment I shall avoid the term "theory" and speak of "models" and "hypotheses" instead.

My use of the term "model" (or "theoretical model") is intended to capture current scientific usage—at least insofar as that usage is itself consistent. To this end, I would adopt a form of the "semantic" or definitional view of theories (hereafter, models). On this view, one creates a model by defining a type of system. For most purposes one can simply identify the model with the definition. But to avoid the consequence that rendering the definition in another language would create a different model, it is convenient to invent an abstract entity, the system defined, and call it the model. This move also preserves consistency with the logician's and mathematician's notion of a model as a set of objects that satisfies a given linguistic structure. For present purposes it will make no difference whether we focus on the definition itself or its nonlinguistic counterpart, so long as there is no presumption that in referring to "the model" one is thereby committed to there being any such thing in the empirical world.

Philosophers differ as to the appropriate form of these definitions. I much prefer the state-space approach of van Fraassen or Suppe to the set-theoretical approach of Suppes, Sneed, and Stegmüller, partly because the former seems better to correspond to scientific practice.[1] Here a system is defined by a set of state variables and system laws that specify the physically possible states of the system and perhaps also its possible evolutions. Thus, for example, classical thermodynamics may be understood as defining an ideal gas in terms of three variables, pressure, volume and temperature, and as specifying that these are related by the law PV = kT. Similarly, classical mechanics defines a Newtonian particle system in terms of a 6n-dimensional space (three components each of position and momentum for each particle) and Newton's laws of motion. A wide variety of models in population genetics (with states given by gene frequencies and development governed by Mendel's laws) are also easily expressed in this framework. So also are learning models in psychology and models of inventory and queuing in economics.

Viewed as definitions, theoretical models have by themselves no empirical content—they make no claims about the world. But they may be *used* to make claims about the world. This is done by identifying elements of the model with elements of real systems and then claiming that the real system exhibits the structure of the model. Such a claim I shall call a *theoretical hypothesis*. These are either true or false. From a logical point of view, the definition of a model amounts to the definition of a predicate. A theoretical hypothesis, then, has the logical form of predication: This is an X, where the corresponding definition tells what it is to be an X.

Our primary concern here is with the testing of theoretical hypotheses and the role of such tests in the overall practice of science. For such purposes, the *logical* differences between statements and definitions are not very important. More important are the implications of this difference for what we take to be the form of the major claims of science.

Since Aristotle it has been assumed that the overall goal of science is the discovery of *true universal generalizations* of the form: All A's are B. Moreover, it has often been supposed that the wider the scope of the antecedent the better. Thus Newton's Law of Universal Gravitation, interpreted as a generalization covering "all bodies," is seen as the epitome of a scientific conclusion. Philosophers, beginning with Hume, have reduced the concept of physical necessity to that of universality, and scientific explanation has been analyzed in terms of derivation from a

generalization. Within this framework it is easy to regard a theory as simply a conjunction of universal generalizations. This would mean that testing a theory is just testing universal generalizations.

The distinction between models and hypotheses permits a view of the goals of science that is more particularized, or at least more restricted—and therefore, I think, more applicable to the contemporary practice of science. The simplest form of a theoretical hypothesis is the claim that a particular, identifiable real system fits a given model. Though extremely limited in scope, such claims may be very complex in detail and wide-ranging in space and time. The claim that the solar system is a Newtonian particle system (together with a suitable set of initial conditions) contains the whole mechanical history of this system—so long as it has been or will be a system of the designated type. Moreover (although this is more controversial), the same hypothesis contains all the different *possible histories* of this system that could result from different, but physically possible, initial conditions. Thus even a very particular theoretical hypothesis may contain a tremendous amount of empirical content.

Contrary to what some philosophers have claimed, one can have a science that studies but a single real system. Current geological models of the earth are not less than scientific, or scientifically uninteresting, simply because the only hypotheses employing these models refer to a single entity limited in time and space. Nor would models of natural selection be in any way scientifically suspect if there were no life anywhere else in the universe. Geology, however, is atypical. The models of a typical science are intended to apply to one or more *kinds* of systems, of which there are numerous, if only finitely many, instances.

What then is a "theory"? It is tempting to identify a theory with a generalized model; for example, the theory of particle mechanics with a generalized Newtonian model (i.e., one in which the number of particles is left unspecified). But most physicists would immediately reject the suggestion that "Newton's theory" is just a definition. And most scientists would react similarly concerning the theories in their fields. They think that "theories" have empirical content. This is a good reason to use the term "theory" to refer to a more or less generalized theoretical hypothesis asserting that one or more specified kinds of systems fit a given type of model.[2] This seems broad enough to encompass all the sciences, including geology and physics.

Testing a theory, then, means testing a theoretical hypothesis of more or

less restricted scope. This is an important qualification because the scope of a hypothesis is crucial in any judgment of the bearing of given evidence on that hypothesis. Knowing what kind of thing we are testing, we can now turn to an analysis of empirical tests. Here I shall not be challenging, but defending, a time-honored tradition.

3. The Hypothetico-Deductive Tradition

To put things in proper perspective, it helps to recall that the hypothetico-deductive method had its origins in Greek science and philosophy. Its most successful employment, of course, was in astronomy. Recast in the above terminology, the goal of astronomy was to construct a model of the heavens that one could use to deduce the motions of the various heavenly bodies as they appear from the earth. "Saving the phenomena" was thus a *necessary* requirement for an acceptable hypothesis. The methodological issue, then as well as now, was whether it is also *sufficient*.

Greek astronomers were well aware that the phenomena could be equally well saved by more than one hypothesis. This methodological fact was exemplified by the construction of both heliocentric and geocentric models. But it was also evident on general logical grounds. Every student of Aristotle's logic knew that it is possible to construct more than one valid syllogism yielding the same true conclusion, and that this could be done as easily with false premises as with true. Truth of the conclusion provides no logical ground for truth of the premises. This obvious logical principle generated a methodological controversy that continues to this day. If two different hypotheses both saved the phenomena, there could be no *logical* reason to prefer one to the other. Some thinkers seemed content to regard any empirically adequate hypothesis acceptable and did not attempt to argue that one was fundamentally better. Others, however, wished to regard one model as representing the *actual* structure of the heavens, and this requires some way of picking out the correct hypothesis from among those that merely save the phenomena.

In the ensuing centuries of debate, the antirealists clearly had logic on their side. The realists, however, did offer several suggestions as to what, in addition to saving the phenomena, justified regarding a hypothesis as uniquely correct. Some appealed to the internal simplicity, or harmony, of the model itself. But this suggestion met the same objections it meets today. There is no objective criterion of simplicity. And there is no way to justify thinking that the simpler of two models, by whatever criterion, is

more likely to provide a true picture of reality. Of course one may prefer a model regarded as simpler for *other* reasons having nothing to do with truth, but we shall not be concerned with such reasons here.

Another suggestion, and one I shall explore further in this paper, is that true hypotheses are revealed by their ability to *predict* phenomena before they are known. This suggestion appears explicitly in the late sixteenth century in the writings of Christopher Clavius, although it must certainly have been advanced earlier.[3] In any case, it became a standard part of the methodology of continental philosophers in the seventeenth century. Thus Descartes, in *Principles of Philosophy* (that is, *natural philosophy*) writes: "We shall know that we have determined these causes correctly only when we see that we can explain in terms of them, not merely the effects we had originally in mind, but also all other phenomena of which we did not previously think."[4]

Leibniz agrees as follows: "Those hypotheses deserve the highest praise (next to truth) . . . by whose aid predictions can be made, even about phenomena or observations which have not been tested before . . . ."[5] The best statement I know, however, occurs in the preface to Huygens's *Treatise on Light* (1690). Having carefully distinguished the deduction of theorems from "certain and incontestable principles" (as in geometry) from the testing of principles by verifying their consequences (the method of science), he continues:

> It is possible in this way to attain a degree of probability which very often is scarcely less than complete certainty. This happens when the things which have been demonstrated by means of the assumed principles agree perfectly with the phenomena which experiment brings to light; especially when there are a great number of them, and, furthermore, principally, when one conceives of and foresees new phenomena which must follow from the hypotheses one employs, and which are found to agree with our expectations."[6]

Huygens refers to the three conditions (agreement, number, and anticipation of new phenomena) as "proofs of probability"—meaning that their satisfaction confers probability on the assumed hypotheses. This is noteworthy because Huygens was one of the first students of probability in its modern form. Yet it would be at least a century, and arguably two, before there were any serious attempts to develop and justify the hypothetical method using ideas from the theory of probability.

In the eighteenth century, the success of Newton's physics sanctified Newton's methodology, including his professed abhorrence of "hypothe-

ses." The method of hypothesis was apparently thought to be almost as discredited as Cartesian physics and Ptolemaic astronomy. Inference to general laws "by induction" from the phenomena was the methodological rule of the day. Interest in the hypothetical method did not revive until the triumphs of wave theories of optics in the nineteenth century—association with scientific success being the apparent standard against which methodological principles are in fact judged. Thus by the third quarter of the nineteenth century we find such eminent methodologists as Whewell and Jevons expounding the virtues of hypotheses with explicit reference to the remarkable predictions that had been based on the wave theory of light. Whewell, for example, writes: "If we can predict new facts which we have not seen, as well as explain those which we have seen, it must be because our explanation is not a mere formula of observed facts, but a truth of a deeper kind."[7] This passage is typical of many of Whewell's writings.

Whewell's homage to the methodological virtues of successful prediction did not go unchallenged. Mill, in particular, denigrated the celebrated predictions of the wave theory as "well calculated to impress the uninformed," but found it "strange that any considerable stress should be laid upon such coincidences by persons of scientific attainments." Moreover, Mill goes on to explain why "coincidences" between "prophecies" and "what comes to pass" should not count for a hypothesis any more than simple agreement with the predicted occurrence. I shall pass over the details of his argument here.[8] Of more interest for this brief survey is that the essentials of the exchange between Whewell and Mill were repeated more than a half-century later in a similar exchange between Peirce and John Maynard Keynes.

In many of his scattered writings, Peirce advocated versions of the following "rule of prediction": "A hypothesis can only be received upon the ground of its having been *verified* by successful *prediction*."[9] Unlike his many predecessors who either lacked the necessary concepts, did not think to apply them, or did not know how, Peirce attempted to justify his rule by explicit appeal to considerations of *probability*. But even this appeal was not decisive. Keynes, whose own view of scientific reasoning incorporated a theory of probability, examined Peirce's arguments for the rule of prediction and concluded that "the peculiar virtue of prediction" was "altogether imaginary." Addressing the details of Keynes's argument would again take us too far afield.[10] I shall only pause to suggest that there must be methodological principles beyond a commitment to concepts of

probability that separate the tradition of Huygens, Whewell, and Peirce from that of Bacon, Mill, and Keynes.

Among contemporary methodolgists, the main defenders of the hypothetico-deductive method seem to be Popper and his intellectual descendents. Elie Zahar and Alan Musgrave have even advocated a special role for successful "novel predictions" in a Lakatosian research programme.[11] Yet these writers seem to me not to be the legitimate heirs of Huygens, Whewell, or Peirce. For the main stream of the hypothetico-deductive tradition, confirmation of a hypothesis through the verification of its consequences, particularly its *predicted* consequences, provides a reason to believe or accept that the hypothesis is *true*. Popper explicitly denies that there can be any such reasons. No matter how "severely tested" and "well-corroborated" a hypothesis might be, it remains a "conjecture" whose truth we have no more reason to believe than we did on the day it was first proposed. Similarly, for Lakatos or Zahar the success of a novel prediction is merely one sign of a "progressive" research program—not a sign of the truth of any particular theory or hypothesis. Only if one accepts the "problem shift" that replaces "reasons to regard as true" with the very different notions of "corroboration" or "progress" can one place these methodological suggestions firmly within the hypothetico-deductive tradition.

Similar remarks apply to those who take their methodological cues from Quine. Insofar as Quine belongs in the hypothetico-deductive tradition, it is that of the antirealists among the classical astronomers. Saving the phenomena is the main thing. Simplicity in one's hypotheses is desirable, but not because of any supposed link between simplicity and truth. Simplicity is desirable in itself or because it contributes to some pragmatic end such as economy of thought. Similarly with prediction. Hypotheses that are useful in making reliable predictions are desirable, but not because this makes them any more likely to be true. Rather, there is pragmatic value in being able to foresee the future, and we value hypotheses with this virtue without thereby ascribing to them any "truth of a deeper kind."

In championing the hypothetico-deductive method of testing scientific hypotheses, I am adopting only the "realist" tradition of Huygens, Whewell, Peirce, and, in part, Popper. I am not defending the more pragmatic or conventionalist versions represented by Quine.[12] Nevertheless, most of the following account is compatible with the subtle antirealism of van Fraassen's *The Scientific Image*.[13] Just how I would differ from van Fraassen will be explained later.

4. Tests of Theoretical Hypotheses

The secret to understanding the rationale of the H-D method is to focus not on the meager logical relations between hypothesis and data but on the notion of a *test* of the hypothesis. What kind of a thing is such a test? What is the purpose of testing hypotheses? What are the possible results of a test? How one answers these fundamental questions will to a large extent determine one's view of the legitimacy of the H-D method.

The ancient idea of a test as an *ordeal* is suggestive because it implies that a test is a process, a procedure to which a hypothesis is subjected. Furthermore, the idea that the purpose of the ordeal is to determine a person's guilt or innocence suggests that the purpose of testing a hypothesis is to determine its truth or falsity. Finally, to complete the analogy, those subjected to ordeals are pronounced guilty or innocent depending on whether they pass or fail the test. This suggests that hypotheses may be pronounced true or false depending on whether they pass or fail the test. Analogies, of course, are not arguments. But by developing the analogy we may be led to a better understanding of tests of theoretical hypotheses.

One way in which we scientific philosophers of the twentieth century have advanced beyond our predecessors is that we now accept the idea that no empirical data could possibly determine *with certainty* that any theoretical hypothesis is *true*. Opinions still differ over whether *falsity* may be so determined. I shall take the liberal position that neither truth nor falsity can be determined with certainty. In the language of testing, no test of a theoretical hypothesis can be completely reliable.

Almost all contemporary students of scientific method would agree that the relevant notion of reliability is to be understood and explicated using concepts of *probability*. But just what role probability plays is a matter of deep disagreement. Many philosophers assume that probability is to be introduced as a measure applied to hypotheses themselves. Thus any test would result in the assignment of a probability to the hypothesis in question. That no test is completely reliable means that this probability is always less than one. I am convinced that this is not the way to go. No model of science that places the relations between evidence and hypotheses *within* the probability calculus will prove adequate. Since I cannot argue so general a thesis here, however, I shall proceed with the constructive task of developing an alternative account. This account *uses* probability without attempting to make scientific inference itself a species of probability calculation.[14]

The way probability enters our account is through the characterization of what constitutes an "appropriate" test of a theoretical hypothesis. So far we have concluded only that a test of a hypothesis is a process whose result provides the basis for our "verdict" either that the hypothesis is true or that it is false. This general characterization, however, is satisfied by the procedure of flipping a coin and calling the hypothesis true if heads comes up and false if tails. This procedure has the virtue that our chances of reaching the *correct* conclusion are fifty-fifty regardless of the truth or falsity of the hypothesis. But no one would regard this as a satisfactory way of "testing" hypotheses. It does, however, suggest that an "appropriate" test would be one that has *higher* probabilities for leading us to the correct conclusion. We shall follow this suggestion.

Thinking about tests in this way throws new light on the classical objections to the method of hypothesis. Let us assume for the moment that our powers of deduction and observation are perfect. This will allow us to concentrate on the nature of tests themselves. At least some realists among the classical astronomers may be viewed as advocating a testing procedure that recommended calling a hypothesis true if and only if it saves the phenomena. Following this procedure, the chances of calling a hypothesis false if it is in fact true are (ideally) zero. A true hypotheses cannot have false consequences. The defect of the procedure is that the chances of calling a false hypothesis true are at best simply not known. One might even argue that this probability is high on the ground that there are many, perhaps even infinitely many, false hypotheses that would also save the phenomena. The odds seem overwhelming that the hypothesis in question is one of these. What is needed to improve the procedure, therefore, is some way of increasing the chances that a false hypothesis will be rejected. This must be done in such a way, however, that the probability of rejecting a true hypothesis is not increased. It would be trivially easy to design a procedure guaranteed to reject false hypotheses: simply reject *any* proposed hypothesis, regardless of the evidence. Unfortunately this procedure is also guaranteed to reject any true hypothesis as well.

The above considerations suggest characterizing an *appropriate test* as a procedure that has *both* an appropriately high probability of leading us to accept true hypotheses as true and to reject false hypotheses as false. Alternatively, an appropriate test of a hypothesis is a procedure that is reasonably *unlikely* to lead us either to accept a false hypothesis or to reject a true one. This characterization still requires considerable elaboration and refinement, but it makes clear the kind of account we seek.

One immediate task is to clarify the interpretation of probability assumed in the above characterization of a good test. By a "procedure" I mean an actual process in the real world. If such a procedure is to have probabilities for leading to different results, these must be *physical* probabilities. Our account thus presupposes an acceptable physical interpretation of probability, something many philosophers regard as impossible. Here I would agree with those who reject attempts to reduce physical probability to relative frequency, and opt for some form of "propensity" interpretation. But since this is again too big an issue to be debated here, I shall proceed under the assumption that there is *some* acceptable physical interpretation of probability.[15] Moreover, we must assume that we can at least sometimes have good empirical grounds for judging the relevant physical probabilities to be high.

5. Example: Fresnel's Model of Diffraction

An example may help to flesh out the relatively abstract outline presented so far. This example is appropriate in many ways, one being its historical association with the re-emergence of the H-D method of testing in the early eighteen-hundreds after a century in the shadows of Newtonian methodological orthodoxy.

Wave models of optical phenomena had been developed by Hooke and Huygens at the end of the seventeenth century. Particle models were favored by Newton and the later Newtonians. At that time, the evidence for either type of model was genuinely ambiguous. Each type of model explained some phenomena better than the others. The then recently discovered phenomenon of polarization, for example, was an embarrassment to both, though perhaps more so to wave theorists. In general, particle models dominated eighteenth-century theorizing, perhaps partly because of greater empirical success but also, I think, because of the general triumph of Newtonianism. In any case, for most of the eighteenth century there was little serious work on wave models until Thomas Young took up the cause around 1800. The scientific establishment, including the French Academy of Sciences, was then dominated by particle theorists. Laplace, for example, published a particle model of double refraction in 1808. But interest in optics was obviously high, since the Academy prizes in 1810 and 1818, for example, were for treatments of double refraction and diffraction respectively.

The diffraction prize eventually went to Augustin Fresnel for a *wave* model. In Fresnel's models, diffraction patterns are produced by the

interference of secondary wave fronts originating at the edges of an object placed between a point light source and screen. Fresnel developed special cases of this general model for a single straight edge, a narrow body with parallel edges, and a narrow slit. The calculated patterns agreed well with known experimental results.

Fresnel's memoir was referred to a Commission in which well-known advocates of particle models, Laplace, Poisson, and Biot, held a majority. The commission was apparently not fully convinced by the evidence Fresnel had presented, and Poisson devised a further test. He applied Fresnel's model to the case of a shadow produced by a circular disk and deduced that the resulting diffraction pattern would have a bright spot at the center of the shadow. Even from superficial accounts of this incident it seems clear that no one involved had ever seen such a spot. Moreover, it seems that Poisson and his fellow Commissioners did not expect the spot to appear. It certainly was not a consequence of any current particle models that such a spot should exist. The experiment was performed by François Arago, and apparently also by Fresnel. The spot appeared as predicted, and the Commissioners yielded.[16]

Now let us consider this episode in light of the framework outlined earlier. Assuming sufficient familiarity with the kind of *model* Fresnel proposed, the next question is just what *hypotheses* were at issue. There are a number of possibilities. (i) The specific set-up in Arago's laboratory fits the model. (ii) Any similar set-up with a circular disk, etc., would fit the model. (iii) All types of diffraction phenomena fit this sort of model. (iv) All optical phenomena fit a similar wave model. Which of these hypotheses were tested by Arago's experiments, and which did the Commission accept in awarding Fresnel the prize?

The episode, I suggest, is best understood taking the hypothesis of most direct concern to be the third of the above four: Fresnel's models capture diffraction phenomena in general. Following the terminology suggested earlier, this hypothesis could be designated "Fresnel's theory of diffraction." Of course everyone was also concerned with whether Arago's set-up fit the model, and this is a consequence of Fresnel's theory. The second hypothesis is also a consequence of Fresnel's theory, but once people were convinced that the model applied to Arago's apparatus, this generalization was not problematic. Enough was known of the general stability of optical phenomena by that time that this simple generalization could legitimately

be taken for granted once it was firmly established for a single case. The emphasis placed by empiricist philosophers on such generalizations is quite misleading.

One reason for focusing on Fresnel's theory of diffraction rather than on a broader wave theory of light is that Arago's experiments provide an appropriate test of Fresnel's theory, but not of the broader theory—in spite of the fact that the former is a logical consequence of the latter. This follows from our characterization of an appropriate test of a hypothesis, as we shall now see.

At the time of Arago's experiments, techniques for dealing with optical phenomena were sufficiently well developed that it was very probable that the spot would be observed—given that the Fresnel-Poisson model does fit this situation. So, given that Fresnel's theory is true, it was very unlikely that it should mistakenly have been rejected. This aspect of the test was entirely appropriate to the circumstances. But what if Fresnel's theory had been false? How probable was it that the testing process should have yielded the predicted spot even if Fresnel's models did not really capture diffraction phenomena? To answer this question we must first decide just how much of the episode to include within the "testing process."

According to common interpretations of the discovery/justification distinction, the decisive testing process began when Poisson constructed a Fresnel-style model for the circular disk and deduced that the spot should appear. Nothing that happened earlier is relevant to the confirmation of any of the hypotheses we have considered. I expect that many who reject a discovery/justification distinction would nevertheless agree with this conclusion. And indeed, this view of the matter follows naturally from the doctrine that there is a "direct" evidential relationship, analogous to deduction, between hypothesis and evidence. But this is not our view. On our account, the relationship between hypothesis and evidence is mediated by the testing process, and there is no a priori reason why incidents that occurred before the actual formulation of the hypothesis should not be relevant to the character of this process. In particular, the process by which a hypothesis is selected for consideration might very well influence its content and thus the likelihood of discovering a further consequence to be true.

The Commissioners apparently did not regard Fresnel's success in explaining the diffraction pattern of straight edges as decisive. Why? Was

this just prejudice? Or did they have good reasons for not regarding these familiar patterns as being part of a good test of Fresnel's models? I think the latter is the case. From Fresnel's own account it is clear that the straight-edge pattern acted as a constraint on his theorizing. He was unwilling to consider any model that did not yield the right pattern for straight edges. Thus we know that the probability of *any* model he put forward yielding the correct pattern for straight edges was near unity, independently of the general correctness of that model. Since the straight-edge pattern thus had no probability of leading to rejection of any subsequently proposed hypothesis that was in fact false, this pattern could not be part of a good test of any such hypothesis.

We could regard agreement with the straight-edge pattern as a test of a hypothesis if we knew the probability that Fresnel should pick out a satisfactory model using this, together with similar data, as a constraint. At best this probability is simply unknown. And given the frequency with which even experienced scientists come up with unsatisfactory models, there is reason to judge such probabilities to be fairly low. In either case we fail to have a good test of the hypothesis.

The case with the spot is quite different. We know that this result did *not* act as a constraint on Fresnel's choice of models. Suppose, then, that Fresnel had come up with a model that applied satisfactorily to straight edges and the like but was *not* correct for diffraction phenomena in general. The corresponding theory would therefore be false. What is the probability that any model selected in this way should nevertheless yield the correct answer for the disk? In answering this question we must also take into account the fact that the disk experiment was specifically chosen because it seemed to Poisson and others that no such phenomenon existed. So the consequence selected for the test was one that knowledgeable people thought unlikely to be true. Given all these facts, it seems clear that the test was quite likely to lead to a rejection of any false theory that Fresnel might have proposed. And this judgment about the test was one that could easily be made by everyone involved. My view is that they did make this judgment, implicitly if not explicitly, concluded that Poisson's proposed test was quite adequate, and, when the result came in, acted accordingly.

In thinking about this and similar examples it is crucial to remember that the probabilities involved are physical probabilities inherent in the actual scientific process itself. If one slips into thinking in terms of probability relations among hypotheses, or between evidence and hypotheses, one

will necessarily misunderstand this account of the nature of empirical testing. In particular, one must not imagine that to estimate the probability of not finding the spot one must be able to calculate the probability of this result as a weighted average of its probabilities relative to all possible alternative theories. No such probabilities are involved. Rather, one needs only to estimate the chances that *any* model generated and tested in this way should fail to cover the general class of diffraction phenomena and nevertheless give the right result for an experiment devised as the disk experiment was in fact devised. My contention is that the participants' knowledge of the whole situation justified concluding that this chance was low.

There remains the question of why Arago's experiment does not provide an appropriate test of a more general wave theory of all optical phenomena. Let us regard this more general theory as a conjunction of theories restricted to various types of optical phenomena: reflection, refraction, polarization, etc. A Fresnel theory of light would say that Fresnel's models are adequate for this whole range of phenomena. Now since Fresnel's theory of diffraction is one conjunct in this larger theory, and since the test is a diffraction experiment, the probabilities of mistaken acceptance or rejection of a more general theory are identical to those for the more restricted theory. Is the experiment not then an equally good test of the broader theory? A positive answer would be a strong objection to this account of empirical testing since it seems intuitively clear that it would not have been correct to accept the broader theory on the basis of these experiments.

The objection fails, however, because the *appropriateness* of a test need not be solely a function of the absolute *magnitude* of the relevant probabilities. It may depend also on what other tests of the theory might be possible, and in this case there were much better tests available. Since Fresnel's model was selected using mainly diffraction phenomena as constraints, the falsity of a general Fresnel theory of light would be much *more likely* to be demonstrated by experiments on phenomena *other* than diffraction. In particular, one would want a phenomenon to which such wave models had not yet been applied. Many such phenomena were familiar at the time. In fact, no such experiments were necessary because it was almost immediately apparent that there were many phenomena, e.g., polarization, for which Fresnel's model gave no account whatsoever.

Many recent philosophers have objected to the H-D method because it

satisfies a "converse consequence condition." That is, if T is confirmed by the truth of some consequence, O, then, if T′ implies T, T′ is equally confirmed. In particular, T and H, for any H, is confirmed. And, granted that a logical consequence of any hypothesis is at least as well confirmed as the hypothesis itself, by confirming T we can equally well confirm any H whatsoever. The above discussion shows that such objections are based on an oversimplified view of the H-D method—indeed, a version to which few if any serious defenders of the H-D method ever subscribed.[17]

6. The Role of Novel Predictions

As we have seen, many champions of the H-D method have suggested that successful predictions are sufficient for the confirmation of hypotheses; some, such as Peirce, have taken them to be necessary as well. Critics argued that successful predictions were neither necessary nor sufficient. From our present perspective we can see why the defenders were on the right track even though their critics were technically correct. First let us give the critics their due.

That successful predictions are *not sufficient* is easily seen by imagining other possible sequences of events in the Fresnel example. Suppose that Biot had repeated Fresnel's calculations for at straight edge and persuaded Arago to repeat these measurements. Of course Biot's prediction would have been verified, but no one would have regarded this replication as providing a decisive test of Fresnel's hypothesis regarding this experiment or of the general adequacy of his approach to diffraction phenomena. Why? Because the imagined process would not have been a good test of the hypothesis. The process had a high probability of supporting Fresnel's hypothesis if it were true. But it also had a high probability of supporting the hypothesis even if it were false. The many previous experiments with straight edges had provided ample evidence for the empirical generalization that this type of experiment yielded the indicated diffraction pattern. So regardless of the truth or falsity of Fresnel's hypothesis, it was highly probable that the hypothesis would be supported. This violates our conditions for a good test of a hypothesis. Both Mill and Keynes used this sort of example in their analyses of the prediction criterion, though each within a quite different framework.

This same counterexample also shows why many H-D theorists have insisted on *novel* predictions. If a predicted result is not novel, there will be a more or less well-justified low-level empirical hypothesis linking the type

of experiment and the type of result. This makes it likely that the test will justify the hypothesis no matter whether it is true or false. Thus it is difficult to have a good test unless the prediction is novel. This point seems to have been missed by empiricist critics of the H-D method such as Mill and Keynes, although perhaps the true value of novelty was also not sufficiently understood by either Whewell or Peirce.

That successful predictions are *not necessary* is also easily demonstrated by an imaginary variation on the same example. It has been claimed that the bright spot in the center of a circular shadow was observed in the early part of the eighteenth century by J. N. Delisle.[18] It seems pretty clear that none of the principles in the case had ever heard of these supposed observations. But suppose they did occur and were published. Imagine, then, that Laplace, but not Fresnel, knew of these results, and upon reading Fresnel's memoir recalled Delisle's unexplained observations. It would not have taken him long to apply Fresnel's method to the case and conclude that Fresnel's model explained Delisle's results. I think the Commission would have been equally convinced that Fresnel's theory was correct. But whether or not they would have been, they should have been. It was about as improbable that Fresnel, ignorant of Delisle's results, should have developed an inadequate model that nevertheless explained Delisle's results, as it was that an inadequate model should have happened to predict correctly the result of Arago's later experiment. In either case the conditions for a good test are satisfied.

Returning to the champions of the prediction rule, it is clear that they overstated their case. But they were fundamentally correct in thinking that the fact that a result was predicted successfully may be *relevant* to the confirmation of a hypothesis. The conditions that define an appropirate test of a hypothesis are themselves contingent empirical hypotheses about the actual process by which a particular hypothesis is tested. This is due to the fact that the relevant probabilities are *physical* probabilities embodied in the testing process. Judging a process to constitute a good test of a hypothesis thus requires judging that the relevant physical probabilities are high. All sorts of empirical facts about the case may be relevant to these judgments—including facts about when a specified consequence of a hypothesis became known, and to whom.

Let us imagine yet another variant on the Fresnel example. Suppose that someone named Delisle really did observe the spot and that Fresnel, but not other principles in the case, knew of Delisle's results right from the

beginning of his work on diffraction. So in addition to explaining the standard diffraction patterns for straight edges, etc., Fresnel was all along seeking a model that would also explain the existence of the spot—and he succeeded. But suppose further that Fresnel suppressed all mention of disks and spots in his memoir, and the rest of the story proceeds as in real life. On our analysis this would be a case of scientific dishonesty, even fraud. The Commission would have had every reason to think that a good test had been performed. But they would have been deceived, for in fact it would have not been a good test. It is not in fact unlikely that a model designed to accommodate a given result should in fact do so. And this is true no matter whether the corresponding hypothesis is true or false. It is possible, therefore, to be justifiably mistaken about whether a given experiment constitutes a good test or not.

We can go further. One might object that it does not matter whether Fresnel *knew* of Delisle's result, but only whether he *used* this knowledge in selecting a model of diffraction. This is in principle correct. The probability of fit between model and observation is not influenced if the observations play no role in the selection of the model. But it is exceedingly difficult to be confident that no such selection pressure existed when the result was known to the developer of the model. One can always be more confident of the goodness of a test if the result was not previously known. So it is a sound methodological rule to prefer results that were genuinely predicted even though this is not strictly necessary for a good test. The methodological good sense of the champions of prediction was better than their justifications.

Finally, our account of testing theoretical hypotheses reveals the methodological insight in Popper's claims that to test a hypothesis severely, one must attempt sincerely to refute it. Far from introducing irrelevant psychological considerations into a supposed logical relationship between evidence and hypothesis, Popper is highlighting (though in an exaggerated and misleading fashion) one aspect of good tests. Not only is the process by which the *hypothesis* was selected relevant to our judgment of the quality of a test; the process of selecting the *prediction* to be investigated is also relevant. This process can be one that makes it more or less likely that the test will reveal the falsity of the hypothesis—if it is indeed false. In particular, if a knowledgeable scientist such as Poisson investigates a model with the express intent of discovering an inadequacy, and finds a consequence he thinks likely to be false, that is good reason to

think that the test has a high probability of discovering a false consequence—if there are any to be discovered. Of course a scientist need not be attempting to refute the hypothesis in question; he may just be trying to devise the best possible test. But the knowledge that a given consequence was selected for investigation in a well-informed attempt to refute the hypothesis is relevant to the judgment as to how good the test might be.

7. The Logic of Tests

A good test of a hypothesis is a physical process with specified stochastic properties, namely, a high probability of one outcome if H is true and another if H is false. That this process has one outcome rather than the other, however, also has *epistemic* consequences. If a good test has a favorable outcome, we are to "conclude" that H is true, "accept" H as being true, or some such thing. One must provide some account of the rationale, or "logic," of this step from the physical outcome to the epistemic conclusion. Here I shall follow those who regard the epistemic step as a kind of *decision*. This opens the way to a decision-theoretic analysis of scientific inference. But since we have renounced probabilities of hypotheses, our decision theory must be "classical," or "non-Bayesian," decision theory.

Casting the problem in decision-theoretic terms, we realize immediately that what really needs to be justified is not so much the decision to accept (or reject) any particular hypothesis, but the general *decision rule* that tells us, for each possible outcome of the experiment, whether to accept or reject the hypothesis. It is so obvious which of the four possible decision rules is correct that most traditional accounts of the H-D method do not even note the epistemic step from physical outcome to accepted conclusion. To understand the logic of the step, however, it is useful to consider the full range of possibilities.

In any test of a theoretical hypothesis there are four possible epistemic results, two correct and two incorrect. The correct ones are accepting H if it is true and rejecting H if it is false. The incorrect ones are rejecting H if it is true and accepting H if it is false. Now to conceptualize the problem in decision-theoretic terms we must assume that it is possible to assign some kind of "value" to these possible results. For the moment we shall not worry what kind of value this is or whether it has the formal properties of utility. And just for convenience (it makes no difference to the argument), we shall suppose that both correct results have the same value (which we

may set arbitrarily at 1) and that both incorrect results also have the same value (which we may set arbitrarily at 0). Finally, let α be the probability that the prediction is false even though H is true and β be the probability that the prediction is true even though H is false. For the moment it will not matter much what these probabilities are so long as they are strictly between zero and one half. With these assumptions we can represent the "decision" to accept or reject H in a two by two matrix (Figure 1).

	Hypothesis true	Hypothesis false
Accept Hypothesis	$Pr = 1 - \alpha$ $V = 1$	$Pr = \beta$ $V = 0$
Reject Hypothesis	$Pr = \alpha$ $V = 0$	$Pr = 1 - \beta$ $V = 1$

Figure 1.

Each of the four possible outcomes is labeled with its respective value and probability (conditional on the hypothesis being true or false).

The meta-decision problem of choosing a decision rule for making the object-level decision is represented as a four-by-two matrix (Figure 2). The obvious decision rule to accept H if and only if the prediction is true is represented as (A, R), and the others are represented accordingly. The outcomes are labeled with the appropriate *expected* values of applying the rule conditional on the truth or falsity of the hypothesis.

	Hypothesis true	Hypothesis false
(A, A)	1	0
(A, R)	$1 - \alpha$	$1 - \beta$
(R, A)	α	β
(R, R)	0	1

Figure 2.

We are now in a position to consider in a systematic way why the obvious rule is correct.

We have arrived in "meta-meta land." What principle do we use to justify a decision rule to guide our actual decisions? The least controversial principle that might possibly apply—namely, dominance—fails to be applicable. Since (A, A) is best if H is true and (R, R) is best if H is false, no one choice of decision rule dominates all others. But the next least controversial principle does apply—namely, *maximin* (it would be "minimax" if we measured our values as losses rather than gains). Maximin recommends the intuitively obvious decision rule (A, R). The justification for this recommendation is that following (A, R) guarantees the greatest "security level," that is, the highest minimum expected gain whether the hypothesis is true or not. In either case, our expected gain can be no lower than the smaller of $(1 - \alpha)$ or $(1 - \beta)$. Since α and β are both less than one-half, no other decision rule can guarantee a greater security level.

The trouble with an appeal to the maximin principle is that it justifies too much. It sanctions the (A, R) rule as long as α and β are less than one-half, which means that any test of a hypothesis is acceptable as long as the probabilities of accepting true hypotheses and rejecting false ones are strictly greater than one-half. This is certainly not in general acceptable for tests of scientific hypotheses. We need a somewhat stronger principle that can force the probabilities of correct acceptance and rejection to be appropriately greater than one-half.

One general decision strategy that has some theoretical backing and seems well suited to the present context is *satisficing*.[19] This strategy may be understood as a strengthening of maximin. To apply this strategy we must assume that there is some way to set a value that is the minimum regarded as a "satisfactory" outcome of the decision process. This is the *satisfaction level* for the particular decision problem at hand. This level is not a function of the given decision matrix but is imposed "from the outside" either by a decision maker or by the decision context. The decision problem will have a *satisfactory* choice only if the security level is at least as great as the imposed satisfaction level. Otherwise there simply will not be any choice sanctioned by this strategy.

Satisficing has not been much studied by decision theorists and philosophers because it does not conform to the standard conditions for a solution to the fundamental meta-decision problem. That problem is: For any decision matrix, devise a general rule that selects a uniquely rational choice

from among all the given possible options. Any rule that does not guarantee some choice of options cannot be a solution to this problem. By invoking satisficing when it does not reduce to maximin, we are rejecting the standard formulation of the fundamental problem of decision theory. And in making satisficing the basis for our acceptance or rejection of hypotheses we would similarly be rejecting a standard, if often implicit, formulation of the basic theoretical problem of scientific inference: For any hypothesis and any evidence, define a uniquely rational function that gives the degree to which the evidence "supports" the hypothesis. On our account, the decision matrix corresponding to an appropriate test must have satisfactory payoffs for the evidence to count either for or against the hypothesis. If the evidence is not part of an appropriate test, the hypothesis is simply neither accepted nor rejected. Neither conclusion is justified.

The demand for a unique solution to any decision problem may have some merit in the context of *practical* decision making. Here one often cannot refuse to make some choice. But science is different. It is not required that scientists be able to say at any moment, given whatever happens to be known at that time, whether some specified hypothesis should be accepted or rejected. One may simply say that sufficiently good tests are lacking. The other side of the coin is that scientists are not helpless in the face of inadequate tests. They may devise and carry out tests they have explicitly designed to be adequate according to the standards of their field. The scientific enterprise, after all, is not simply a matter of evaluating hypotheses in light of available information, but an active seeking out of information to answer definite (and often theoretical) questions.

My analysis of the Academy Commission follows the above outline. The evidence presented in Fresnel's memoir did not constitute a sufficient test of Fresnel's theory, and the Commission, rightly, could not decide whether it was correct or not. So they devised a new experiment, one that was sufficient. What made the new experiment sufficient was primarily that it had an adequately high probability of rejecting a mistaken model.

Probabilities, however, are not the only component in a decision-theoretic analysis of the logic of testing. We must also assume the existence of "epistemic" or "scientific" values. I say "values" rather than "utilities" because a satisficing strategy, like maximin, can be applied with as weak a value structure as a mere ordering—a full-fledged utility function is not necessary. This is an additional virtue of satisficing as a general decision strategy. Still, those who question this approach may ask whether there are any such things as scientific values and, if so, whether they can play the

role assigned to them by this analysis. Here I can only sketch a reply.

In principle we need appeal to only one scientific value, truth. That is, correctly accepting a true hypothesis or rejecting a false one is valued over the reverse. This seems very difficult to deny, except as part of a radical theory of science in which truth plays little or no role.[20] Otherwise the decision-theoretic analysis requires only that the satisfaction level of expected value be strictly less than the value assigned to accepting truth (or rejecting falsehood). Because the probability of a mistake is strictly greater than zero, our expected payoff is necessarily less than maximal no matter whether H is true or not. If our satisfaction level were equal to the value of a correct decision, no amount of data or care in experimental design would be good enough for us ever to make either choice. Thus we really do not need to delve deeply into questions about the value of truth, e.g., *how much* is truth valued over error. So long as truth is *more* valued, we can assume an arbitrary scale. The interesting question is what determines the satisfaction level. And given a fixed value scale, this question reduces to asking what determines an acceptable level of risk. So we are back to probabilities.

My view, which I can only affirm here, is that the satisfaction level, or level of acceptable risk, is not a function of individuals but of the *institution* of science. The institution decrees that tolerating some risk of error is better than having no theoretical conclusions at all. Yet something more should be forthcoming. It seems that different fields, or the same field at different stages of maturity, have different tolerances for risk of error. Again, it seems not just that scientists in some fields are more risk adverse. Rather, some *fields* are more or less risk adverse. I think there are objective reasons why this is so and why it is proper, but that would be a digression here. The main point is to recognize this as an important question for a theory of science.[21]

One important consequence of introducing values into an account of scientific inference is that it automatically relativizes the acceptance of a theoretical hypothesis to the scientific context characterized by those values. This relativization eliminates what has often been regarded as a fatal objection to the idea that hypotheses are ever "accepted" as being true. The objection is that it would be irrational to regard any hypothesis as true for all purposes or in any possible context. Thus Laplace may have been convinced of the truth of Fresnel's hypothesis, but it would have been irrational to stake the fortunes of the whole French nation on this conclusion. On our analysis, however, Laplace's commitment to the truth

of the hypothesis would be restricted to the scientific context, leaving open the question whether the hypothesis was sufficiently well-tested to presume its truth if other values were at stake. This makes the relationship between "pure" and "applied" science more complicated than is often supposed, but that, I would argue, is an added virtue of this approach.[22]

8. The Weight of Evidence

It is a commonplace that the evidence for a theory may be better or worse, and that evidence may accumulate. In short, evidence seems to come in degrees. Our account, however, leads to hypotheses being either "accepted" or "rejected." Does not this account, therefore, run counter to well-established methodological intuitions? I think not, but some explanation is in order.

In the first place, it is not strictly true that our account leaves no room for degrees of support. A test can be more or less stringent (or severe) depending on the probabilities of mistaken acceptance or rejection. The probability of correctly rejecting a mistaken hypothesis, what statisticians call the "power" of a test, is a particularly good measure of the severity of a test. For example, later tests of wave models tended to be better than the early experiments of Fresnel and Arago because they employed fairly precise *quantitative*, rather than merely qualitative, predictions. The famous 1850 experiment of Foucault and Fizeau on the relative velocity of light in air and water was a much better test because wave models gave a quite precise value for this ratio based on the known relative indices of refraction. The chances of a badly mistaken model yielding the right value to within the known experimental error had to have been very small indeed.

Second, the idea that a theory is a generalization over regions of a domain provides ample room for the idea of accumulating evidence for a theory. As more and more regions of the domain are brought under a given type of model, the evidence in favor of the generalization increases. This is how I would understand the history of the wave theory between 1820 and 1850. As wave models were refined, more and more optical phenomena were shown to fit: diffraction, polarization, double refraction, and so on. By 1850 it was reasonable to conclude that all optical phenomena could be accounted for by suitably specialized wave models.[23]

9. Approximation and Scientific Realism

Peirce said that every scientific conclusion should be doubly qualified by the phrase "probably and approximately." Thus far we have been primarily concerned with the role of probability in the testing of theoretical hypotheses, and, by implication, the way probability "qualifies" our conclusions. It is time to turn to the second of Peirce's qualifications. The need for some such qualification is once again well-illustrated by the Fresnel episode. At the time of his original memoir on diffraction, Fresnel's wave models employed *longitudinal* rather than transverse waves. It so happens that for diffraction phenomena the two types of models give at least qualitatively similar results. This is not true in general, however, and it was the switch to transverse wave models several years later that provided the key to a wave explanation of polarization and other previously difficult cases. But what are we to say of the Commission's conclusion after Arago's experiments with the opaque disk? Were they justified in concluding simply that Fresnel's theory is correct, even though this conclusion had later to be modified? Or should the conclusion have been softened by a qualifying "approximately"?

One reason for softening conclusions is the general realization that (with the possible exception of models of microphysical phenomena) no model is likely to capture the full complexity of the phenomena under investigation. Thus no hypothesis asserting a "perfect fit" between a model and a phenomenon is likely to be exactly true. We can be fairly confident on general grounds that there are likely to be at least minor discrepancies, even if these do not show up in existing experiments. Thus, if one desires a conclusion that one does not know in advance to be strictly false, it seems advisable to say that what one accepts as true is not ever H itself, but the more complicated conclusion that "H is approximately true."

This position is reinforced by our account of the testing of theoretical hypotheses. It is required that the testing process be such that a false hypothesis is very likely to lead to a failed prediction and thus to rejection of the hypothesis. But we could never be in the position of knowing that even a very slight difference between the model and the real system would very probably lead to a failed prediction. The most we could ever reasonably claim about a testing process is that it is very likely to detect a type and degree of deviation from the model. The *type* is given by the category of phenomena being studied. Arago's experiment could not possibly detect mistaken features of the model that would not show up in simple diffraction

experiments. Nor could it detect *degrees* of deviation beyond the resolving power of his experimental apparatus. If the spot had been of very low intensity, for example, he would have missed it even though the wave model was basically correct. Thus, if the experiment yields a positive conclusion, that conclusion must be correspondingly qualified. What we can say is that the real system exhibits the structure of the model in those respects relevant to the domain in question and to a degree that the experiments performed are capable of detecting. This is an elaboration of the simple phrase "H is approximately true."

It is sometimes supposed that in moving from "H is true" to "H is approximately true," one is trading a probable falsehood for a vacuous truth. That this is not so is well illustrated by the Fresnel example. Models of the type Fresnel used for diffraction simply do not work for polarization experiments. And the reason is clear: longitudinal waves do not polarize. The correct conclusion, as was clear at the time, was that any supposed Fresnel theory of polarization was not even approximately true. Of course much more could and needs to be said about approximation, but the general charge of vacuity can be safely dismissed.

Even doubly qualified, our conclusions are still quite solidly realistic. With all qualifications understood, to accept H as approximately correct is to assert that there is a real structure underlying the phenomena, and that this underlying structure, as far as we can tell, reflects the structure of our model. Bas van Frassen has recently raised general objections that apply even to so qualified a version of scientific realism. Since the view of theoretical models embedded in my account is essentially van Fraassen's, it is important to see how his objections may be avoided.[24]

Reformulated in my terminology, van Fraassen's view is that theoretical hypotheses may *refer* to underlying entities and processes, but that no evidence can *justify* accepting claims about such theoretical goings on. Rather, the most we can ever justifiably claim is that our models are "empirically adequate," i.e., that they "save the phenomena." His main direct argument for so restricting our conclusions is that, for any given evidence, the conclusion of empirical adequacy is necessarily better supported than any realistic conclusion simply because it is logically weaker. The realistic conclusion implies empirical adequacy, but the reverse implication, of course, fails. Van Fraassen tends to state his argument by saying that the empiricist hypothesis of empirical adequacy is *more probable,* for any given evidence, than a realistic hypothesis. The

general point, however, does not depend on invoking (logical or subjective) probabilities of hypotheses.

That the empiricists' hypothesis is always *better* supported does not, however, imply that the realists' hypothesis cannot be *adequately* supported. Reformulating the same point, just because empirical adequacy is a *more acceptable* hypothesis, it does not follow that a realistic hypothesis is not *acceptable*. Van Fraassen assumes that only the more probable empiricists' hypothesis could be acceptable. That this assumption is not necessarily justified is reinforced by the logic of satisficing. If we assign *equal* value to the truth of both empiricist and realist hypotheses, the empiricist hypothesis, being more probable, would have greater expected value. But the realist hypothesis might still have "satisfactory" expected value, making it acceptable to a satisficer. But of course a realist would assign *greater* scientific value to true realistic hypotheses, which could give them greater expected value. So a realistic satisficer need not even be in the position of settling for second best.

These considerations do not, of course, settle the issues between realists and empiricists. My objective has been simply to show that one can adopt van Fraassen's formal account of theoretical models without committing oneself to his anti-realist arguments. Van Fraassen himself insists that this separation be possible, but it is useful to see how to make the separation for a particular realistic alternative.

10. Final Considerations

In conclusion, I would like briefly to mention two issues that would have to be treated in any generally adequate account of scientific inference but that cannot be examined in any detail here. One involves a technical elaboration; the other is philosophically much broader.

The above account treats only deterministic models. This leaves out stochastic models, which yield only statistical predictions. But the account also fails to be adequate to the actual testing of deterministic models. Experiments testing theoretical hypotheses typically require multiple measurements on complex apparatus that, when pushed to their limits of accuracy, yield a spread of values. Thus even tests of deterministic models typically involve testing statistical hypotheses. Testing theoretical hypotheses is therefore generally a two-stage affair. One begins by testing a statistical hypothesis or estimating a statistical parameter in order to determine whether the theoretical prediction has been fulfilled or not.

Only after deciding on the truth or falsity of the prediction can one reach a conclusion on the theoretical hypothesis. It turns out, however, that the above account of testing theoretical hypotheses can be extended in a completely natural way to incorporate the statistical level into the whole testing process. No fundamentally new principles need be involved.[25]

The broader issue concerns *justification* of the type involved in traditional philosophical discussions of "the justification of induction." On the above account, whether a proposed test of a theoretical hypothesis is a satisfactory test is itself an empirical question. The judgment that fulfillment of the prediction would be unlikely if the hypothesis tested is not (even approximately) true is an empirical judgment. This opens the way for a typical Humean regress.

Philosophers of science today are much less concerned with Hume's problem than in the recent past. Foundationist justifications, in particular, have largely gone out of fashion. In their place are discussions of the "rationality" of the scientific process. I regard this development as a "progressive problem shift." The new program, however, requires a sound theory of the scientific process before one can fruitfully inquire after the rationality of that process. We do not yet have such a theory. Indeed, one of the main defects of current theories is that they lack good accounts the empirical testing of theoretical hypotheses. We may hope this is just a temporary situation, part of the general over-reaction to positivism. In any case, only when we achieve a more balanced picture of the scientific enterprise itself shall we be in a position to develop better ideas about the nature of scientific justification and rationality.[26]

Notes

1. Bas C. van Fraassen, On the Extension of Beth's Semantics of Physical Theories. *Philosophy of Science* 37 (1970): 325-339; Frederick Suppe, Theories, Their Formulations and the Operational Imperative. *Synthèse* 25 (1973); 129-164; Patrick Suppes, What is a Scientific Theory?, in *Philosophy of Science Today* S. Morgenbesser, ed., (New York: Basic Books, 1967), pp. 55-67; Joseph Sneed, *The Logical Structure of Mathematical Physics* (Dordrecht, Holland: D. Reidel, 1971); Wolfgang Stegmüller, *The Structure and Dynamics of Theories* (New York: Springer, 1976).

2. Here I depart from the view I have taken in previous papers and in my elementary text, *Understanding Scientific Reasoning* (New York: Holt, Reinhart and Winston, 1979). I have generally used the term "theory" to refer to a generalized definition or a model—which has no empirical content. But this usage has met sufficient resistance from scientists and science students that I have decided to compromise in the interests of communication. The underlying view of the scientific enterprise, however, is the same. My view here parallels that of Sneed and Stegmüller, although the parallel is difficult to see through the forest of set theory in which they hide their views.

3. R. M. Blake, Theory of Hypothesis among Renaissance Astronomers, in *Theories of Scientific Method*, ed. E. H. Madden (Seattle: University of Washington Press, 1960).

4. René Descartes, *Principles of Philosophy*, Part III, Sec. 42, in *Descartes: Philosophical Writings*, trans. E. Anscombe and P. Geach (Edinburgh: Nelson, 1954), p. 223.

5. L. E. Loemker, *Leibniz: Philosophical Papers and Letters*, 2 vols. (Chicago: University of Chicago Press, 1956), Vol. 1, p. 288. See also the paper On the Method of Distinguishing Real from Imaginary Phenomena. In Loemker, Vol. 2, p. 604.

6. Christiaan Huygens, *Treatise on Light*, trans. S. P. Thompson (London: Macmillan, 1912).

7. William Whewell, *Philosophy of Discovery* (London: J. W. Parker, 1860), p. 273.

8. John Stuart Mill, *Logic* (8th ed.) (London, 1881), Book III, Ch. XIV, Sec. 6.

9. C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, 8 vols. (Cambridge, Mass.: Harvard University Press, 1931-1958), 2.739.

10. John Maynard Keynes, *A Treatise on Probability* (London: Macmillan, 1921), pp. 304-306.

11. E. G. Zahar, Why Did Einstein's Programme Supersede Lorentz's? *The British Journal for the Philosophy of Science* 24 (1973): 95-123 and 223-262. Alan E. Musgrave, Logical Versus Historical Theories of Confirmation. *The British Journal for the Philosophy of Science* 25 (1974): 1-23.

12. Still less am I defending the logical shadow of the hypothetico-deductive tradition recently criticized by Clark Glymour in *Theory and Evidence* (Princeton: Princeton University Press, 1980), pp. 29-48.

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

14. I have elaborated on the deep philosophical differences between these two approaches to scientific inference in Testing vs. Information Models of Statistical Inference. In *Logic, Laws and Life*, ed., R. G. Colodny (Pittsburgh: University of Pittsburgh Press, 1977), pp. 19-70. For further references see this article and the review essay Foundations of Probability and Statistical Inference. In *Current Research in Philosophy of Science*, ed. P. D. Asquith and Henry Kyburg, Jr. (East Lansing: Philosophy of Science Association, 1978).

15. I have myself developed a propensity interpretation in several papers beginning with Objective Single-Case Probabilities and the Foundations of Statistics, in P. Suppes, L. Henkin, A. Joja, and Cr. C. Moisil, eds. Logic, Methodology and Philosophy of Science, IV (Amsterdam: North-Holland, 1973), pp. 467-83. For later references see also A Laplacean Formal Semantics for Single Case Propensities, *Journal of Philosophical Logic* 5 (1976): 321-353.

16. The above account is based entirely on secondary sources such as E. T. Whittaker's *A History of the Theories of Aether and Electricity* (London: Thomas Nelson & Sons, 1910; rev. ed. 1951; Torchbook edition, 1960).

17. This objection was first advanced by C. G. Hempel in his classic Studies in the Logic of Confirmation, *Mind* 54 (1945): 1-26 and 97-121. It forms much of the basis of Glymour's discussion in *Theory and Evidence*.

18. The reference to Delisle's observations appears in a footnote in Whittaker's *History* (Torchbook edition, p. 108). However, Whittaker gives no references and there are no other entries under "Delisle" in the index. Here is a case where *historical* research might alter one's *methodological* appraisal of a scientific episode.

19. For an authoritative introduction to satisficing see H. A. Simon, *Models of Man* (New York: Wiley, 1957), pp. 196-206 and also chapters 14 and 15.

20. Such a view is developed in Larry Laudan's influential book *Progress and its Problems* (Berkeley: University of California Press, 1977).

21. In several previous papers I have attempted to distinguish two rough types or stages in scientific inquiry, "exploratory" and "confirmatory," and to argue that the satisfaction level, as reflected in the acceptable risk for a mistaken conclusion, is justifiably higher in confirmatory inquiry. See Empirical Probability, Objective Statistical Methods and Scientific Inquiry, in *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, W. L. Harper and C. A. Hooker, Vol. 2 (Dordrecht, Holland: D Reidel, 1976), pp. 63-101;

and Testing vs. Information Models of Statistical Inference, in *Logic, Laws and Life*, ed. R. G. Colodny, (Pittsburgh: University of Pittsburgh Press, 1977), pp. 19-70.

22. This point is developed is somewhat greater detail in both papers mentioned in n. 21.

23. The useful notion of a "domain" has been developed in a number of papers by Dudley Shapere. See, for example, Scientific Theories and their Domains, in *The Structure of Scientific Theories*, ed: Fred Suppe, (Urbana: University of Illinois Press, 2nd ed., 1977), pp. 518-555.

24. See n. 13.

25. This point is further developed in the two papers mentioned in n. 21 above.

26. My own version of a "nonfoundationist" justification of induction is developed in The Epistemological Roots of Scientific Knowledge, *Minnesota Studies in the Philosophy of Science*, Vol. VI, *Induction, Probability and Confirmation*, G. Maxwell and R. M. Anderson, eds., (Minneapolis: University of Minnesota Press, 1975), pp. 212-261.