True Belief Knowledge?," *Analysis,* 23 (1963), 121–123 and Clark, "Knowledge and Grounds: A Comment on Mr. Gettier's Paper," *Analysis,* 24 (1963), 46–48.

[5] Cf. "How Belief Is Based on Inference," *The Journal of Philosophy,* LXI (1964), 353–360.

[6] See note 3.

Clark Glymour

RELEVANT EVIDENCE

Scientists often claim that an experiment or observation tests certain hypotheses within a complex theory but not others. Relativity theorists, for example, are unanimous in the judgment that measurements of the gravitational red shift do not test the field equations of general relativity; psychoanalysts sometimes complain that experimental tests of Freudian theory are at best tests of rather peripheral hypotheses; astronomers do not regard observations of the positions of a single planet as a test of Kepler's third law, even though those observations may test Kepler's first and second laws. Observations are regarded as relevant to some hypotheses in a theory but not relevant to others in that same theory. There is another kind of scientific judgment that may or may not be related to such judgments of relevance: determinations of the accuracy of the predictions of some theories are not held to provide tests of those theories, or, at least, positive results are not held to support or confirm the theories in question. There are, for example, special relativistic theories of gravity that predict the same phenomena as does general relativity, yet the theories are regarded as mere curiosities.[1]

Prima facie, such judgments either may be conventional and properly explained entirely by sociological factors, or else they may have an underlying rationale and so may be explained as applications of general principles of scientific inference. At least with regard to the first kind of judgments, that is, those which are explicitly judgments of relevance, three different philosophical views are common: (1) the hypothetico-deductive method provides an obvious and well-understood rationale for such discriminations; (2) one or another system of inductive logic provides a rationale for such discriminations; and (3) there is no rationale for the judgments in question, and they must really be entirely the result of convention.[2] All three opinions are, I believe, quite wrong; there are principles that explain and provide a rationale for scientific judgments of relevance, but they are not exactly hypothetico-deductive principles nor are they principles of a probabilistic kind. The principles that provide a rationale for judgments of relevance also provide a partial rationale for other central features of scientific method; notably, they also explain why some theories are not supported by determinations of the accuracy of predictions derived from them. One consequence is that, although theories may be underdetermined by all possible evidence of a specified kind, they need not be so radically or so easily underdetermined as some writers, including myself,[3] have thought.

Consider the first of the above positions: One might suppose that some hypotheses in a theory are, in conjunction with initial con-

Reprinted by permission from the author and *The Journal of Philosophy,* Vol. LXXII, no. 14, (August 14, 1975), pp. 403–426.

ditions, *essential* to the deduction of a sentence that is decidable by experiment or observation. Such hypotheses would then be tested by the appropriate experiments or observations whereas other hypotheses in the theory—those not essential to the deduction—would not be so tested. An account of this kind is satisfactory only if the notion of an "essential" hypothesis can be made precise; and there are good reasons to believe that such a clarification is not trivial and perhaps not even possible, for the difficulties in making precise the notion of essential hypotheses are exactly those which meet any attempt to provide a criterion of cognitive significance of the kind long sought by the positivists. The positivists proposed to divide the predicates of a theory into two disjoint classes, one of which would comprise the "observation terms" of the theory. A sentence in the language of the theory was to be deemed significant if it was testable, and testability was to be defined solely in terms of the consequence relation holding between, on the one hand, sentences, or classes of sentences, in the language of the theory, and, on the other hand, sentences whose only nonlogical terms were observational. Every attempt to provide such a criterion has failed, and the catalogue of failures is familiar.[4] But if we could specify in precise logical terms what it is for a hypothesis, in conjunction with initial conditions, to be essential to the deduction of an experimentally decidable sentence, then taking the observation terms to be those nonlogical terms occurring in the experimentally decidable sentence or in the statement of initial conditions, we would have an account of testability of the kind the positivists required. We must expect that all the technical sorts of objections that told against empiricist criteria of cognitive significance would tell against any attempt to give a hypothetico-deductive account of epistemic relevance. Some of those who were themselves once part of the positivist tradition saw this connection fairly clearly and drew very strong holist conclusions from the failure of significance criteria. David Kaplan[5] reports that when Carnap was presented with a class of counter-examples (devised by Kaplan) to his last attempt at a significance criterion, "he reflected that he had been quite wrong for about 30 years, and that his critics who had been arguing that theories must be accepted or rejected as a whole (he mentioned at least Quine and Hempel) were very likely correct." And Hempel, at the end of his negative review of attempts at empiricist significance criteria, proposed that theories be evaluated in terms of their clarity and precision, and by such holist canons as simplicity, explanatory and predictive power, and the extent to which they have, as a whole, been confirmed by experience.

Which brings us to the second position. Hempel's own qualitative theory of confirmation[6] has the property that, if e is an evidence statement and p any sentence, consistent with e, that is not a logical consequence of a sentence all of whose nonlogical terms occur in e, then e confirms neither p nor the negation of p. But most of the evidence for complex theories is stated in terms that use only fragments of the vocabularies of the theories. For example, the positions of the planets on the celestial sphere supports Kepler's laws, but this evidence is stated in terms of times, ascensions, and declinations: the notions of a period of an orbit, a mean distance from the sun, and so on, do not occur in the statement of such evidence. Accordingly, despite the fact that his intent was to give an account of epistemic relevance,[7] Hempel's theory cannot explain why such evidence provides support for the theory as a whole or for particular hypotheses within the theory. Quantitative theories of confirmation using logical measure functions—Carnap's m^* for example—do better, but they share some of the limitations of Hempel's system; for example, if a hypothesis and an evidence statement share no nonlogical vocabulary, then the second generally cannot confirm or disconfirm the first.

Several contemporary accounts of scientific inference suppose it to proceed by the formation of conditional probabilities by means of Bayes's theorem in the theory of probability. That is, it is assumed that there are prior probabilities assigned to all hypotheses in question, and the new or posterior probability of (or degree of belief in) a hypothesis *h* on new evidence *e* is just the conditional probability of *h* on *e* (and whatever old evidence there may be). Richard Jeffrey has generalized this strategy so that it need not be assumed that the evidence statement, *e*, is certain.[8] A test that results in evidence *e* is taken to be relevant to hypothesis *h* if and only if the posterior probability of *h*, that is, its conditional probability on *e*, is different from the prior probability of *h*. Analyses of this sort may perhaps be made consistent with the sorts of judgments of relevance described at the outset, but I think we should doubt that they explain such judgments or provide a rationale for them. In order to determine the conditional probability of *h* on *e* by Bayes's rule we must know the prior probabilities of *h* and of *e*, and we must know the conditional probability of *e* on *h*. Frequentists maintain that such prior probabilities are objective frequencies; more particularly, Reichenbach proposed that the prior probability of a theory or hypothesis be taken as the frequency of success in a suitable reference class of theories of the same kind as the theory in question. He gave, unfortunately, no account of how the success of past theories might, without circularity, be determined, nor did he indicate with any concrete examples just how the required groupings might be effected. Reichenbach himself seems to have understood his account as a proposal for future practice: "Should we some day reach a stage in which we have as many statistics on theories as we have today on cases of disease and subsequent death . . . the choice of the reference class for the probability of theories would seem as natural as that of the reference class for the probability of death."[9] Whatever the merits or difficulties with the proposal, one thing is clear: it cannot provide a rationale for those detailed judgments of relevance which scientists now make and have long been making, nor can it explain the great agreement scientists in the same field show about such matters. For we simply do not have statistics of the kind Reichenbach envisioned, nor do we have any idea of what their values would be or even of how to collect them.

Subjective probability theorists, who regard the probabilities of hypotheses as measures of our degrees of belief in them, are not affected by such criticism. But of course, on a strict subjectivist view, the assignments of prior probabilities are quite arbitrary so long as they accord with the requirements of the theory of probability. If, then, judgments of relevance are to be explained ultimately in terms of prior probability distributions, and those distributions are without rationale, the judgments of relevance will also be without rationale.[10] The bare subjectivist account seems to be a version of the third position above: judgments of relevance are conventional.

The conventionalist view would presumably attribute the agreement about relevance to such factors as the education of graduate students: young scientists are told by old scientists what is relevant to what. All relativity texts say that certain experiments do not test certain hypotheses because that was what all relativity textbook writers were taught. There are two difficulties: these suppositions do not explain how judgments of relevance came to be established in the first place, and they do not explain how it is that, with very little controversy, judgments of relevance are made in new cases. The latter fact, especially, suggests that, if scientific education determines scientific judgments about the relevance of evidence to theory, it must do so by teaching, explicitly or tacitly, principles and not merely cases. On the other hand, the conventionalist view has for its support the fundamental consideration that no plausible principles are known that would warrant the discrimination in question. I shall try to remove that support.

II

It is widely thought that, save in exceptional circumstances, universal hypotheses are supported or confirmed by their positive instances. If the hypothesis contains anomalous predicates—"grue," for example—then it will fail to be confirmed by positive instances, and, likewise, if the hypothesis is entailed by some well-confirmed theory, and a positive instance of the hypothesis is inconsistent with that theory, then the instance may serve to reduce our reasons to believe the hypothesis. But, barring circumstances such as these, we expect that universal hypotheses will be confirmed by their positive instances, and, in particular, we expect that a quantitative hypothesis stated as an equation will be confirmed by a set of values for the magnitudes[11] occurring in the hypothesis if the set is a solution to the equation. Now the trouble is that our experiments, observations, and measurements do not appear to provide us with positive instances of the hypotheses of our theories; in the quantitative case, for example, the magnitudes we determine by experiment or observation are generally not those, or not all of those, which occur in our theories concerning the phenomena observed.

Scientists seem to know very well how to get values of magnitudes occurring in their theories from values of magnitudes determined experimentally. Their strategy is to use hypotheses of the very theory to be tested to compute values of other magnitudes from experimentally determined magnitudes. To take a very simple example, suppose our theory consists of the single hypothesis that, for any sample of gas, so long as no gas is added to or removed from the sample, the product of the pressure and volume of the gas is proportional to the temperature of the gas. In other terms, under the given conditions

$$PV = kT$$

where k is an undetermined constant. Suppose further that we have means for measuring P, V, and T, but no means for measuring k. Then the hypothesis may be tested by obtaining two sets of values for P, V, and T, using the first set of values together with the very hypothesis to be tested to determine a value for k

$$k = \frac{pv}{t}$$

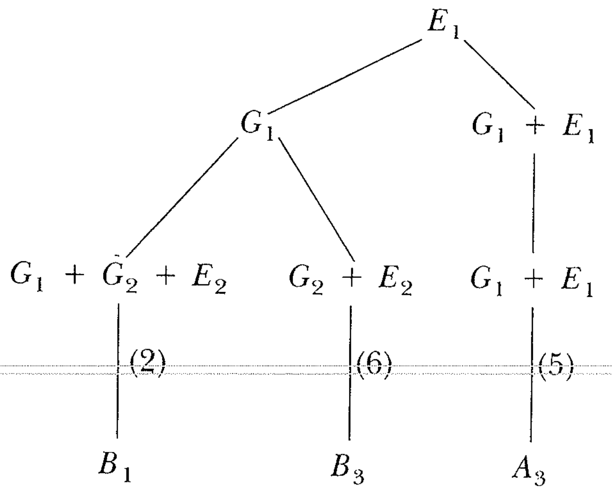
and using the value k thus obtained together with the second set of values for P, V, and T either to instantiate or to contradict the hypothesis.

In the example the very hypothesis to be tested was used to determine, from experiment, a value for a quantity occurring in it, and the determination was very simple. Cases of this kind abound in scientific literature,[12] but in general the situation is considerably more complicated. Typically, the theory in question will contain a great many hypotheses, and a given experiment or collection of experiments may fail to measure values of more than one quantity in the theory. To determine a value for one of the latter quantities the use of several hypotheses in the theory may be required, and the determination may proceed through the computation of values for intermediate quantities, or combinations of such. Such a determination or computation may be represented by a finite graph. The initial, or zero-level, nodes of the graph will be experimentally determined quantities; n-level nodes will be quantities or combinations of quantities such that, for each n-level node, some hypothesis of the theory determines a unique value of that node from suitable values of all the $(n-1)$-level nodes with which it is connected. The graph will have a single maximal element, and that element will be a single quantity. We permit that two connected or unconnected nodes may correspond to the same quantity or combination of quantities. I will call such a graph a *computation.*

The graph associated with the computation of the constant in the ideal-gas law is obvious, but it may not be clear what happens in a more complicated case. Let us consider a theory developed in a recent psychological paper;[13] since our considerations are almost entirely structural, we need not concern ourselves with much of the detail regarding the interpretation—which happens to be complicated—of the quantities occurring in the theory. The theory consists of the following set of linear equations, together with their consequences (with respect to real algebra):

$$\begin{aligned}
&(1)\quad A_1 = E_1\\
&(2)\quad B_1 = G_1 + G_2 + E_2\\
&(3)\quad A_2 = E_1 + E_2\\
&(4)\quad B_2 = G_1 + G_2\\
&(5)\quad A_3 = G_1 + E_1\\
&(6)\quad B_3 = G_2 + E_2
\end{aligned}$$

The *A*s and *B*s are supposed to be quantities that we know how to estimate experimentally. Suppose then that we do an experiment that gives us values for the quantities A_1, B_1, A_3, and B_3. Naturally we could use equation (1) to compute a value for E_1 immediately from the experimental value of A_1. But it is also possible to compute a value for E_1 from the values of B_1, A_3, and B_3 in the following way:



As we have seen, a given set of data may permit the computation of a value for a quantity in more than one way. If the data are consistent with the theory, then these different computations must agree in the value they determine for the computed quantity, but, if the data are inconsistent with the theory, then different computations of the same quantity may give different results. Further, and most important, what quantities in a theory may be computed from a given set of initial data depends both on the initial data and on the structure of the theory. In the example above we supposed given values for A_1, B_1, A_3, and B_3. These permit us to compute values for E_1 and for G_1, but, as the authors of the paper from which we have taken the equations put it, "two of the parameters, G_2 and E_2, occur only together in the expectations with the same coefficients, and are therefore inseparable. We can therefore estimate only G_1, E_1, and $(G_2 + E_2)$" (*ibid.,* 317). That is, we cannot, with this theory, get values of G_2 and of E_2 with these data. Similar things happen with other sets of possible initial values. If we have values of A_1, B_1, A_2, B_2 only, then we cannot compute values for G_1 or for G_2. If, initially, we have values for A_2, B_2, A_3, B_3 only, then we cannot compute values for any of the quantities that appear on the right-hand side of the preceding equations.

It is clear, then, I hope, how scientists may use hypotheses in their theories for the determination of values of quantities that are not in fact measured or estimated by standard statistical methods. The examples already given suffice, I believe, to show that the strategy is in fact used explicitly in some cases. The question is, to what end is this strategy used? More particularly, if experiment permits the computation of values for all quantities occurring in a hypothesis, and these values accord with the hypothesis, does the positive instance thus obtained support or confirm the hypothesis? The answer cannot always be affirmative. Consider the example just discussed; suppose we determine A_1 by experiment and use the hypothesis:

$$(1) \qquad A_1 = E_1$$

to compute a value for E_1. We then have values for both A_1 and E_1, and these values are in accord with hypothesis (1) and provide a positive instance of that hypothesis. But clearly it would be wrong to think that this instance provides any support for the hypothesis. Intuitively, the difficulty is that the value of E_1 has been determined in such a way that, no matter what the value of A_1, it could not possibly fail to provide a positive instance of the hypothesis. To test a hypothesis we must do something that could result in presumptive evidence against the hypothesis. So a plausible necessary condition for a set I of values of quantities to test hypothesis h with respect to theory T is that there exist computations (using hypotheses in T & h) from I of values for the quantities occurring in h, and there exist a set J of possible values for the same initial quantities such that the same computations from J result in a negative instance of h—that is, the values of the quantities occurring in h which are computed from J must contradict h. Actually, it is not necessary that all the quantities occurring in h be computable from the initial data, for some of them may occur vacuously. For example, to test an equation of the form

$$a(x^2 + y) + bx - ay = 0$$

we do not require a value for y. The quantity y is vacuous in the equation because, given any value v of x for which there exists a value u of y such that (v, u) is a solution to the equation, then (v, z) is also a solution for all possible values, z, of y. The generalization to cases with more quantities is obvious.

There is another useful condition which, for many theories, is equivalent to that just given. Suppose a hypothesis is equivalent to an equation of the form:

$$X(Q_1 \cdots Q_j) = 0$$

where X is some functional form, and where it is understood that two hypotheses are equivalent if every set of values which is a solution of one is a solution of the other and vice versa. Suppose further that a value for every quantity occurring in the hypothesis can be computed (by using hypotheses of a given theory T) from a set of values for experimentally determined quantities $E_1 \ldots E_k$. Now, for any quantity Q_i occurring in the hypothesis, the computation for Q_i specifies Q_i as a single-valued function of the quantities whose nodes are immediately connected to the Q_i node in the graph of the computation. Similarly, the quantities at the nth-level nodes are, each of them, specified as single-valued functions of the quantities at the $(n - 1)$-level nodes with which they are connected. Thus, ultimately, by composing all these functions, Q_i itself is specified as a single-valued function $f_i(E_1 \ldots E_k)$ of the experimentally determined quantities $E_1 \ldots E_k$. Replacing each Q_i in the hypothesis by $f_1(E_1 \ldots E_k)$ we obtain the equation

$$X(f_1(E_1 \ldots E_k), \ldots, f_i(E_1 \ldots E_k))$$

in which the only quantities are those experimentally determined. We shall say that this equation *represents* the hypothesis for this set of computations. For example, if the hypothesis is (1) above, that is,

$$A_1 = E_1$$

and the only computation is that of E_1 illustrated previously, then the representative of the hypothesis for this computation is

$$A_1 = (A_3 + B_3 - B_1)$$

Now the following is obvious: If the representative of a hypothesis for a set of computations holds identically, that is, if every set of possible values for the quantities occurring in the representative is a solution of the representative, then the computations cannot test the hypothesis, because the necessary condition given before will not obtain. Something more is true. If the functional form X of the hypothesis, and the functions f_i, are composed of operators that determine unique values for all possible sets of values of the quantities they operate on, then the hypothesis will be tested by a set of computations from initial data if the equation representing the hypothesis is not an identity.

We have, in effect, an account of theory testing, and one that naturally evolves from a few elementary observations: *ceteris paribus,* hypotheses are supported by positive instances, disconfirmed by negative; instances, whether positive or negative, of a hypothesis in a theory are got by using the hypotheses of that theory itself (or, conceivably, some other) to make computations from values got from experiment, observation, or independent theoretical considerations; the computations must be carried out in such a way as to admit the possibility that the resulting instance of the hypothesis tested will be negative. Hypotheses, on this account, are not generally tested or supported or confirmed absolutely, but only *relative to a theory.* The general idea is certainly not new. Herman Weyl,[14] for example, seems to have had it:

> The requirements which emerge from our discussion for a correct theory of the course of the world may be formulated as follows:
>
> 1. *Concordance.* The definite value which a quantity occurring in the theory assumes in a certain individual case will be determined from the empirical data on the basis of the theoretically posited connections. *Every such determination has to yield the same result* . . . Not infrequently a (relatively) direct observation of the quantity in question . . . is compared with a computation on the basis of other observations. . . .
>
> 2. It must in principle always be possible to determine on the basis of observational data the definite value which a quantity occurring in the theory will have in a given individual case. This expresses the postulate that the theory in its explanation of the phenomena, must not contain redundant parts (121/2).

Again, in "Testability and Meaning"[15] Carnap proposed to regard hypotheses as confirmed by observation statements if the hypotheses, or instances of them, could be deduced from premises consisting of the observation statements and certain special hypotheses. The special hypotheses—bilateral reduction sentences—were in effect allegedly privileged hypotheses of a theory; privileged in being immune from disconfirmation and in being analytic. But the appeal to analytic truth is quite independent of the main idea, namely, to confirm hypotheses by deducing instances of them by means of other hypotheses in the same theory.

III

Before turning to the questions with which we began, some objections to this account of theory testing need to be considered.

One objection is that the foregoing account is an account of testing for quantitative theories only; it does not seem to apply to qualitative theories or to theories construed as deductively closed, axiomatizable sets of first-order sentences. But the account is straightforwardly extended to first-order theories, and thereby to qualitative theories if the logical form of their hypotheses is known.

By a "quantity" we will mean an open atomic formula. By a "value" for a quantity we will mean an atomic sentence or its negation containing the same predicate constant as the quantity. It certainly must be allowed that, if initial data I (that is, a set of values for quantities) and theory T are consistent, then I disconfirms h with respect to T if T and I together entail $\sim h$ but T alone does not. Conversely, if T and I are consistent and T and I entail h but T alone does not, then I must count as confirming h with respect to T. The more typical and more complicated cases arise when T and I together neither entail nor refute h unless T does so alone. For these cases we may give a quasi-Hempelian analysis:

I confirms h with respect to T if

(i) T and I are consistent with each other and with h.

(ii) There exists a set, call it S, of values for quantities such that there are computations from I of the values in S and, further, such that S entails the development (in Hempel's sense[16]) of h for the individual constants occurring in members of S.

(iii) There exists a set J of possible values for the initial quantities such that the same computations (as in ii) from J given values of the quantities in S that entail the development of the negation of h.

I disconfirms h if I confirms the negation of h.

I should like briefly to note some features of this account. If I is inconsistent with T, then I neither confirms nor disconfirms any hypothesis with respect to T; but in that case I may nonetheless confirm or disconfirm various hypotheses with respect to sub-theories of T. Hempel's consistency and equivalence conditions are satisfied so long as the theory is kept fixed. The same initial data may, however, confirm inconsistent hypotheses with respect to different theories. Because of condition iii, Hempel's special consequence condition is not satisfied, and neither, of course, is the converse consequence condition.

On Hempel's theory, $\sim R(a)$ confirms both $\forall x \sim Rx$ and $\forall x(Rx \supset Bx)$, but, on the account just given, it does not, because no value of $R(x)$ will, by itself, entail the development of the negation of the second hypothesis, and so condition iii is not met. The "paradox" of the ravens arises in the new account just as in Hempel's, but it is at least confined: if initial data Ra,Ba confirm a hypothesis of universal conditional form with respect to theory T, it is not always the case that $\sim Ra, \sim Ba$ also confirm that hypothesis with respect to T. For example, if the hypothesis is $\forall x(Cx \supset Dx)$ and the theory is $\forall x(Rx \supset Cx)$ & $\forall x(Dx \equiv Bx)$, then the first set of initial data, Ra,Ba confirms the hypothesis, but $\sim Ra, \sim Ba$ does not confirm the hypothesis.

Although I think that most of the features of the foregoing account for first-order theories are plausible enough, I shall not defend them now. There are a variety of ways in which the general strategy I have outlined in the previous section might be extended to formalized theories, and the quasi-Hempelian account just given is only one of them. One can, for example, try to preserve the consequence condition by replacing iii with a radically weaker condition, e.g.,

(iii*) If h has a representative for the set of computations in ii, the representative is not a valid formula.

but then one will have to allow that $\sim Ra$ confirms $\forall x(Rx \supset Bx)$. Again, it is straightforward to adapt the general strategy to a Popperian viewpoint, so that hypotheses of universal form may be tested but hypotheses of existential or mixed form never are. The point is that the account *can be* extended to formalized theories, and the extension need not be much less plausible—I think not any less plausible at all—than accounts of confirmation that are confined to "observation" statements.

A serious difficulty, urged by Professor Hempel, is this: typically, the hypotheses of a theory of themselves determine nothing about experimental or observational data; something definite about experimental outcomes can be inferred from the theory—or values of theoretical quantities can be infered from the data—only if special, empirically untested, assumptions are made. Hempel calls such assumptions "qualifying clauses" or "provisos." One example, alleged by several writers, is that no observable consequences about the motions of heavenly bodies follow from Newton's three laws and the law of universal gravitation unless one makes some assumption about what forces are acting, e.g., that only gravitational forces act between the bodies of the solar system.

There may indeed be many cases in which a theory can be applied to a system only if it is assumed that the system has some property of a kind that is not determined experimentally; even when that is so, however, one must still be able to say what hypotheses in the theory are tested by the experimental results on the supposition that the qualifying clause is met, and our account proposes an answer to that question. Of course, one wants to know something more about when it is reasonable to assume that qualifying clauses are satisfied, and what role they may play in the assessment of a whole theory, but that is beyond our scope at present.

It is not clear to me how often such qualifying clauses are really essential. Consider Newton again. In book III of the *Principia* Newton uses his first two laws to deduce from Kepler's laws that there is a centripetal force acting on the planets in inverse proportion to the square of their distances from the sun. He further shows, using terrestrial experiments and the third law, that this centripetal force between two bodies must be proportional to the product of their masses. Now, as deductivists like Duhem[17] have insisted, these deductions do not result in an instance of the gravitational-force law because that law requires that the gravitational force acting between *any two* bodies be proportional to the product of the masses and inversely proportional to the square of the distance between them; but the total gravitational force acting on any planet must be the sum of the forces due to the sun and to the other bodies in the solar system, and hence the total gravitational force acting on a planet ought not to be inversely proportional to the square of its distance from the sun. Newton's conclusions are inconsistent with his law. Duhem's objection fails entirely, however, if we recognize that Kepler's laws need not be taken as strictly correct initial data, but rather as very good approximations subject to whatever errors there may be in the observations of planetary positions and times. The question then becomes whether the planetary perturbations are sufficiently small that the deviation in the total force acting on a planet from that calculated by Newton using Kepler's laws is less than the error of the computed result due to error in the initial data. Such a determination in turn, requires, besides some idea of the error of the observations, an estimate of the relative masses of the planets to the sun. For any planet with a satellite, the ratio of the planet's mass to the sun's can be estimated from data independent of those used to compute the circumsolar force; Newton is thus able to argue without circularity that the gravitational interaction of the planets is very small in comparison with the solar force.[18]

In effect, the method of testing described in this paper is Newton's method, save that in Newton's case the matter is complicated by the use of empirical laws as initial data and the use of approximations. Not only Newton, but Newtonian scientists of the eighteenth and nineteenth centuries claimed to deduce their laws from the phenomena. Perhaps they overstated their case, but they had, nonetheless, a case to state. The scorn heaped on their method by Duhem is undeserved.

Another objection is that the account is, after all, just the old hypothetico-deductive account. For, if a set of initial data confirms a hypothesis with respect to a theory according to the preceding account, then surely there is a valid deduction of some of the propositions in the initial data set from premises consisting of the rest of the propositions in the initial data set, the hypothesis tested, and the theorems of the theory that are used in the computations. Further, if the data disconfirm the hypothesis, the negation of some proposition in the initial data set must be deducible in an analogous way. And surely H-D theorists would agree that in some contexts only some particular hypothesis or hypotheses from among all those which might appear in such deductions are in fact tested.

It is true, I think, that any test can be converted into a deductive argument in the way suggested; but the converse is not true. Not all deductions of singular statements from putative laws and initial conditions can be transformed into tests. For example, suppose hypothesis h is tested by data I with respect to theory *T*. For each predicate occurring in *h* or in *T* but not occurring in *I*, choose two new, distinct predicates, and replace each occurrence of each predicate, *P* say, by the disjunction of the two new predicates associated with *P*. Then *h* is changed into a new hypothesis *h**, and *T* is changed into a new theory *T**, and, further, if there is a valid deduction of a proposition in *I* from the rest of *I*, *h*, and theorems of *T*, then, by the substitution theorem, there is also a valid deduction of that proposition in *I* from the rest of *I*, *h**, and *T**. But, in general, *I* will not test *h** with respect to *T**. That is exactly as it should be, for no scientist would take evidence to support a theory like *T** when another like *T* was available. The H-D method has us deduce singular statements from laws; the new procedure, in effect, has us deduce *instances* of laws from singular statements and other laws. The two are not the same. I have no doubt that H-D advocates agree that sometimes data test certain hypotheses and not others; what I doubt is that their principles afford any explanation of those judgments.

IV

We still have to consider what the account of theory testing can contribute to the questions with which we began. What grounds can there be for claims to the effect that one or another experiment has no bearing on one or another hypothesis within a theory? In general terms our answer is clear enough: depending on the nature of the experiment or observation and the structure of the theory in question, a given hypothesis may or may not be tested according to the scheme outlined in previous sections. In particular cases, detailing the application of the scheme may be very complex, and the psychoanalytic and relativity examples mentioned at the outset are certainly too complex to discuss here.[19] It is, however, fairly easy to see how the account of theory testing can explain the claim that observations of a single planet do not, of themselves, provide a test of Kepler's third law.

Kepler's first and second laws specify features of the motion of any planetary body moving about the sun. The third law, however, relates features of the orbits of any two bodies; specifically it claims that the ratio of the periods of any two planets equals the 3/2 power of the ratio of their mean distances from the sun. The parameters that uniquely determine the Keplerian orbit at any time can be estimated from several observations of the planet on the celestial sphere; in fact, three suitably chosen observations suffice for the computations,[20] and a fourth observation of a single planet permits a test of Kepler's first and second laws. But, however many observations we may have of the location of a single planet on the celestial sphere, those are not, by assumption, observations of the location of any *other* planet on the celestial sphere. To test Kepler's third law, we need estimates of the periods and mean distances from the sun of at least two planets. But from the observations of one planet alone we cannot compute, using Kepler's laws and their consequences, the parameters of the orbit of any other planet. We can, of course, compute under those circumstances the *ratio* of the square of the period to the cube of the mean distance from the sun for any planet whatsoever, but only by *using* Kepler's third law itself. So, even if we count such a ratio as one quantity, the representative of Kepler's third law (see p. 412 above) for the requisite computations will be a trivial identity, and hence the third law will not be tested.

The account of theory testing helps to account for a good deal more about scientific methodology. A standard methodological principle is that a theory is better supported by a variety of evidence than by a narrow spectrum of evidence. The substance of the principle is, however, unclear so long as we lack some account of what constitutes relevant variety. One view, which I believe is incorrect, is that what constitutes a relevant variety of evidence for a theory is entirely determined by what other theories happen to be in competition with the first.[21] On the contrary, if, as I have argued, a given piece of evidence may be evidence for some hypothesis in a theory even while it is irrelevant to other hypotheses in that theory, then we surely want our pieces of evidence to be various enough to provide tests of as many different hypotheses in that theory as possible, regardless of what, in historical context, the competing theories may be. There is a further complication. In assessing a theory we are judging how well it is supported with respect to itself, and this reflexive feature of theory testing makes for certain difficulties. If a hypothesis is confirmed by observations and computations using another hypothesis in the theory, then it is always possible that the agreement between hypothesis and evidence is spurious: both the hypothesis tested and some hypothesis used in the computations of the test may be in error, but the errors in one hypothesis may be exactly (or exactly enough) compensated for by the errors in the other. Conversely, a true hypothesis may be disconfirmed by observations and computations using other hypotheses in the theory if one or more of the hypotheses used in the computations are incorrect. The only means available for guarding against such errors is to have a variety of evidence, so that as many hypotheses as possible are tested in as many different ways as possible. What makes one way of testing relevantly different from another is that the hypotheses used in the one computation are different from the hypotheses used in the other computation. Part of what makes one piece of evidence relevantly different from another piece of evidence is that some test is possible from the first that is not possible from the second, or that in the two cases there is some difference in the precision of computed values of theoretical quantities.

Kepler's laws again provide a simple example. Kepler did not determine elliptical orbits for planets as simply the best fit for the data; on the contrary, he gave a physical argument for the area rule—his second law—and used the area rule together with the data to infer that the planetary orbits are ellipses. Seventeenth-century astronomers were able to confirm Kepler's first law only by using his second, and they were able to confirm his second only by using his first. Understandably, there remained considerable disagreement and uncertainty as to whether the two laws were correct, or whether the errors in one were compensated for by the errors in the other. Not until the invention of the micrometer and Flamsteed's observations of Jupiter and its satellites, late in the seventeenth century, was a confirmation of Kepler's second law obtained without any assumption concerning the planet's orbit.[22] I doubt that this example is singular; quite the reverse: it seems unlikely to me that the development and testing of any complex modern theory in physics or in chemistry can be understood without some appreciation of the way a variety of evidence serves to separate hypotheses.

At the outset it was observed that some theories are regarded chiefly as curiosities and rarely taken seriously, despite the fact that they account for all the evidence accounted for by some theory taken very seriously and are not known to be irreconcilable with any other phenomena. In many cases this kind of scientific discrimination can plausibly be explained as the result of applying the principles of evidential relevance that we are concerned to describe.

Some years ago Walter Thirring[22] published a special relativistic theory of gravitation. Thirring's theory supposes that spacetime has a flat metric η_{uv} like that of special relativity and that gravitation is due to a tensor field, ψ_{uv}, that has no effect on the metric. Writing down equations for these quantities, Thirring was able to show that his theory accounts for many of the phenomena that are usually taken to confirm general relativity. His theory is almost universally regarded as a curiosity; such an assessment might of course result from mere prejudice or from any of a variety of obscurely motivated methodological opinions, e.g., the view that a phenomenon confirms a theory only if the theory literally *predicts* the phenomenon. But I think the account of relevant evidence developed in the preceding sections best explains this assessment, and also best explicates what physicists typically say in justifying that assessment. What they say is that Thirring's theory is defective because his metric, η_{uv}, is not "observable."[24] A better word would be 'determinable', and, if we understand the authors in that way, then the complaint makes perfect sense. Free-falling particles do not follow geodesics of Thirring's metric, η_{uv}, nor do clocks measure time according to it, nor rods distance. What, according to the theory, such systems measure are geodesics, time, distances, as determined by the quantity:

$$\eta_{uv} - f\psi_{uv}$$

where f is a suitable function. By making compensatory changes in ψ_{uv}, an infinite variety of different flat metrics η_{uv} can be made compatible with all data about rods, clocks, test particles, etc. This is not just experimental uncertainty, or a failure to obtain perfect accuracy in our measurements. We noted earlier that, if in a theory a quantity A is replaced throughout by an algebraic combination of new quantities B, C, D, then hypotheses formerly tested by various initial data may be turned into hypotheses not tested by those data, because values for B, C, D cannot be computed even approximately. That is in effect what happens in Thirring's theory: the general relativistic metric, g_{uv}, which is determinable in principle from the behavior of material objects, is replaced by an algebraic combination—$(\eta_{uv} - f\psi_{uv})$—of new quantities. The result is that values for the new quantities cannot be computed from the relevant initial data, and so, although it might be possible to determine evidence *against* Thirring's theory, it is not possible to determine evidence *for* its central hypotheses because they cannot be instantiated. The physicists' principle is that we should prefer theories whose hypotheses are positively tested by our evidence to theories that, even though consistent with our data and affording an explanation of it, are not positively tested by it. The principle is a good one.

Theories with quantities whose values cannot be determined by the evidence are, in an intuitive way, less simple than theories without undeterminable quantities or with fewer of them. Still, it is a mistake to see this discrimination as no more than a manifestation of our preference for simple theories; I think we do better to try to understand whatever rational preference there may be for simpler scientific hypotheses as derivative from our preference for better tested theories, and the account presented here provides at least a partial rationale for our attachment to simplicity. Quine, for one, seems to think differently:

> Yet another principle that may be said to figure as a tacit guide of science is that of sufficient reason. A lingering trace of this venerable principle seems recognizable, at any rate, in the scientist's shunning of gratuitous singularities. If he arrives at laws of dynamics that favor no one frame of reference over others that are in motion with respect to it, he forthwith regards the notion of absolute rest and hence of absolute position as untenable. This rejection is not, as one is tempted to suppose, a rejection of the empirically undefinable; empirically unexceptionable definitions of rest are ready to hand, in the arbitrary adoption of any of various specifiable frames of reference. It is a rejection of the gratuitous. This principle may, however, plausibly be subsumed under the demand for simplicity, thanks to the looseness of the latter idea.[25]

Though it is perfectly correct that we can always make determinable an undeterminable quantity in a theory merely by adding a further hypothesis, that is not enough. For it is not in the least obvious that we can always add a hypothesis which will be tested by the evidence available or which will not be tested negatively either by the evidence available or by evidence easily produced. In Newtonian theory there is no way to compute which unaccelerated trajectories through space-time are truly at rest with respect to absolute space. One can easily add to the theory untestable hypotheses about the rest frame—e.g., that the center of mass of the universe is at rest; and one can easily add hypotheses one has every reason to believe false or at best contingently true—e.g., that inertially moving cabbages are at rest. Doing better is hard. Identification of the rest frame with the reference system in which particular physical systems—whether cabbages or the Sun—are at rest is unsatisfactory, for such correlations cannot be even approximate laws because the physical systems can be accelerated. The aether, were there one, would perhaps have done the job, but there is not one, and the importance of that fact in the history of physics underscores the point: what theoretical magnitudes we can determine depends on what lawlike hypotheses are available to us, and that, in turn, depends on what kinds of things there are.

There is another kind of case where judgments often attributed to a taste for simple things can at least partially be attributed instead to a taste for well-tested things. First a word about error. Suppose the measurements that comprise some body of evidence are subject to error and, though the exact error of any measurement is unknown, an upper bound to the error is known. Then each measurement may be regarded as determining an interval of possible values of the measured quantity, within which the true value must lie. This is, I believe, a typical circumstance in scientific measurement. Computations of theoretical quantities may proceed as before, but what is determined from the data is a *set* of values of any computed quantity. Again, a test of a theoretical relation is understood as before, but with the following complication: what is required for an instance of a hypothesis is, for each quantity in the hypothesis, a set of values for that quantity such that there can be drawn from the respective sets a collection of precise values—one for each quantity—satisfying the hypothesis. This is the obvious generalization of our account when error is present.

Suppose a theorist is entertaining hypotheses about the functional form of the relation between two quantities, X and Y, which he can determine experimentally. We assume that he has no well-established theory to guide him, and we suppose his measurements of X to be subject to some error of known bound. If getting values for X and for Y is difficult, costly, and tedious, our theorist will doubtless wish to draw his conclusions from but a few data points if that is possible. Suppose he has six points and, to within the tolerable error, they lie on a line: our theorist claims that the relation between X and Y is linear. Why does he think the linear hypothesis better than some other polynomial relation? In particular, the six data points are perfectly consistent with the hypothesis that

$$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4 + a_5X^5$$

and, because of the error, the coefficients of quadratic and higher powers of X need not be zero. Why the linear hypothesis rather than the fifth-degree hypothesis? Can there be any more to it than a taste for simple things?

A Popperian answer is that the simpler, linear hypothesis can be falsified by fewer data points than can the fifth-degree hypothesis. This cannot be exactly the right reason, for, *given* the six data points, the linear and fifth-degree hypothesis each require the same number of *additional* data points for a possible falsification, namely, one. The reason for the preference, I suggest, is straightforward: two data points permit a computation of intervals of values for the undetermined coefficients of the linear hypothesis, and four more data points permit four tests of that hypothesis; but the six values of X and Y permit only a computation of intervals of values of the constant coefficients in the fifth-degree hypothesis; they do not permit any test of it. The theorist should prefer the linear hypothesis for the straightforward reason that he has more positive evidence for it than for any other polynomial relation.

V

There are two theses which have recently gained such wide assent among empiricist philosophers that they deserve to be regarded as new dogmas of empiricism. I have in mind the claim that our theories may be underdetermined by all possible evidence, and the further claim that each theory is tested as a whole. Dogmas may of course be true, and, with suitable qualifications, these dogmas are. I should like to conclude by saying something about the qualifications.

For some theories, at some stages of their development, a set of quantities can plausibly be demarcated such that the evidence for or against the theory in question consists of values for these quantities for various systems. When such a demarcation can plausibly be made, it not only makes sense to ask whether the theory is uniquely determined by all possible evidence of the relevant kind, but, further, we can sometimes hope to get an answer to this question. Of course an answer, whether affirmative or negative, says nothing about what sorts of underdetermination may occur if novel kinds of evidence are discovered. For example, the state of absolute rest is undeterminable in Newtonian gravitational theory, but, had the combination of Newtonian theory with Maxwell's electrodynamics proved correct, optical experiments would have permitted a determination of the rest frame.[26] Again, for certain models of general relativity, it can be shown that no measurements of the quantities peculiar to that theory suffice to determine the global topology of space-time,[27] but, even if our universe is in fact one of these topologically underdetermined universes, it is still possible that other branches of physics—plasma physics for example—might provide evidence and theory sufficient to determine a unique topology.

If we confine consideration to a given kind of evidence, we can inquire whether evidence of that kind uniquely determines a best theory that explains it. Conceivably, all possible such evidence might fail to determine a unique theory for either of two kinds of reasons. First, there might occur two or more theories that are not intertranslatable but all of whose hypotheses are tested positively by the evidence so that every methodological demand met by one theory is met by the other. I know of no plausible candidates for this kind of case, but I see no reason why they should not exist. Second, there might occur two or more theories that differ only in hypotheses that cannot be tested, and, for some reason or other, every plausible theory accounting for the evidence also contains such a hypothesis. There are a great many examples of this kind of case, and analyzing when this sort of underdetermination arises is a standard problem in the social sciences.[28]

Demonstrating underdetermination is sometimes possible, but it is not as easy as some writers have supposed. Reichenbach,[29] for example, argued that, even in the context of classical physics, the theory of the geometry of space is underdetermined; for, given any geometry, we can suppose it to be the true one and explain the coincidence behavior of material bodies in terms of this geometry and the action of a "universal force." But, if one sets out actually to write down such a theory, one quickly discovers that it is obtainable only by dividing the Euclidean metric of Newtonian theory into two new quantities, just as Thirring divided the metric field of general relativity into two new quantities. The result is a theory which, on the same evidence, is less well tested than Newtonian theory. We cannot demonstrate underdetermination by substituting for one or more predicates of a theory a combination of new predicates, since the result of the substitution is a theory less well tested than the original.

Early in this century both Duhem and Frege urged that a theory must be tested as a whole. Reductive programs, like Carnap's *Aufbau,* would have avoided holism had they succeeded, but they did not succeed. Later, a number of philosophers, notably Carnap and C. I. Lewis, tried to avoid holism by putting analytic truth to work. They kept in common some version of the claim that, given a collection of analytic truths, or truths by convention, each hypothesis in a theory has its own, independent connections with experience. It is understandable that a new romance with holism should be the concomitant of estrangement from the distinction between analytic truths and synthetic truths.

Part of what has been said or suggested on behalf of holism is false, and part of it is true. It is true that a great part of a theory may be involved in the confirmation of any of its hypotheses, and it is further true that the assessment of any hypothesis in a theory in the face of negative evidence requires the assessment of all hypotheses in that theory. It is false that a piece of evidence is evidence indiscriminately for all hypotheses in a theory or for none of them, and it is false as well that theories must be accepted or rejected as a whole. For positive evidence may fail to provide any support for some hypotheses in a theory—support, that is, with respect to the theory itself—even while confirming other hypotheses. And, if the total evidence is of sufficient variety, evidence inconsistent with a theory may still leave us with a fragment that is best confirmed with respect to itself. If we are lucky, in some axiomatizations of the theory we may even be able to single out a particular axiom that deserves the blame. A naive holism that supposes theory to confront experience as an unstructured, blockish whole will inevitably be perplexed by the power of scientific argument to distribute praise and to distribute blame among our beliefs.

NOTES

1. See, for example, Ya. B. Zeldovich and I. D. Novikov, *Relativistic Astrophysics* (Chicago: University Press, 1971), pp. 66–71.

2. For the third position, see Kaplan, *infra;* the view is certainly suggested by many of Quine's remarks, but I find it nowhere explicitly in his writings. The second position is perhaps the most popular: cf. C. I. Lewis, *An Analysis of Knowledge and Valuation* (La Salle, Ill.: Open Court, 1946); H. Reichenbach, *Experience and Prediction* (Chicago: University Press, 1938); R. C. Jeffrey, "Probability and Falsification," unpublished; I can cite no texts for the first view, but philosophers at the University of Chicago and at Indiana University, where earlier versions of this paper were read, urged it. I am indebted to them for their criticism, and to the National Science Foundation for support of research. I owe special thanks to Richard Jeffrey and to Carl Hempel for reading and criticizing drafts of this essay.

3. "Theoretical Realism and Theoretical Equivalence," in R. Buck and R. Cohen, eds., *Boston Studies in the Philosophy of Science,* vol. VIII (Boston: Reidel, 1971).

4. Cf. Hempel, "Empiricist Criteria of Cognitive Significance: Problems and Changes," in *Aspects of Scientific Explanation* (New York: Free Press, 1965).

5. "Homage to Carnap," in Buck and Cohen, *op. cit.,* pp. xlvi–xlvii.

6. "Studies in the Logic of Confirmation," in *Aspects of Scientific Explanation, op. cit.*

7. See *ibid,* p. 5/6. Hempel was, of course, aware of the difficulty and entertained remedies. One remedy, the converse consequence condition, he rightly rejected, and subsequent attempts to revive it [cf. B. Brody, "Confirmation and Explanation," this JOURNAL, LXV, 10 (May 16, 1968): 282–299] have not proved fruitful.

8. *The Logic of Decision* (New York: McGraw-Hill, 1965), ch. 11.

9. *Theory of Probability* (Berkeley: Univ. of California Press, 1949). This passage is taken from S. Luckenbach, *Probabilities, Problems and Paradoxes* (Encino, Calif.: Dickenson, 1972), p. 44.

10. The standard Bayesian response to criticisms that turn on the arbitrariness of prior probabilities is by appeal to stable estimation theorems; i.e., to proofs that, under certain conditions whatever the prior distributions may be, the posterior distributions will be nearly the same given sufficient evidence. [Cf. Edwards, Lindman, and Savage, "Bayesian Statistical Inference for Psychological Research," *Psycholog-*

ical Review, LXX (1963).] But I know of no such theorems for the kind of case under consideration, that is, when the evidence statements are confined to a proper sublanguage of the language in which the hypotheses may be formulated.

11. I shall use the terms 'magnitude' or 'quantity' either to signify abstract objects, e.g., the type of the token 'kinetic energy', or else to signify properties under a description. The important point is that for my purposes "mean kinetic energy" . . . and "temperature" must count as different quantities even though temperature is mean kinetic energy.

12. For example, some of Jean Perrin's tests of equations of the kinetic theory are exactly of the kind illustrated. Perrin had, for instance, to use one of the equations to be tested to determine a value for a constant (Avogadro's number) it contained.

13. J. Jinks and D. Fulker, "Comparison of the Biometrical, Genetical MAVA and Classical Approaches to the Analysis of Human Behavior," *Psychological Bulletin,* LXXIII, 5 (May 1970). The equations given are taken from p. 316.

14. *Philosophy of Mathematics and Natural Science* (New York: Atheneum, 1963).

15. *Philosophy of Science,* III, 4 (October 1936): 419–471; IV, 1 (January 1937): 1–40; reprinted in H. Feigl and M. Brodbeck, *Readings in the Philosophy of Science* (New York: Appleton-Century-Crofts, 1953).

16. Cf. "Studies in the Logic of Confirmation," *op. cit.*

17. Cf. *The Aim and Structure of Physical Theory* (Princeton, N.J.: University Press, 1954), *passim.*

18. This discussion ignores many historical niceties. Newton assumes, for example, that the center of gravity of the solar system moves inertially, and this assumption, having no experimental support, is presumably just the sort of thing Hempel would call a "proviso." But Newton's argument does not in fact require the assumption. A more careful account of Newton's argument is given in my "Physics and Evidence," to appear in *Pittsburgh Studies in the Philosophy of Science.*

19. For a very qualitative application of the strategy to Freudian theory, see my "Freud, Kepler and the Clinical Evidence," in R. Wollheim, ed., *Freud* (New York: Doubleday, 1975). The explanation I should offer of why the field equations of general relativity are not tested by measurements of the gravitational red shift turns on the imprecision of these measurements and closely follows the account given by John Anderson in his *Principles of Relativity Physics* (New York: Academic Press, 1967), ch. 12.

20. The classic treatment is Gauss, *Theory of the Motion of Heavenly Bodies Moving about the Sun in Conic Sections.* A translation from the Latin is published by Dover, New York, 1963.

21. For this view see, for example, Peter Achinstein, "Inference to Scientific Laws," in R. Stuewer, ed., *Minnesota Studies in the Philosophy of Science,* vol. v (Minneapolis: Univ. of Minnesota Press, 1970), p. 95.

22. Cf. Curtis Wilson, "From Kepler's Laws, So-called, to Universal Gravitation: Empirical Factors," *Archive for the History of Exact Sciences,* VI (1969): 89–170.

23. "An Alternative Approach to the Theory of Gravitation," *Annals of Physics,* XVI (1961): 96–117.

24. Zeldovich and Novikov, *loc. cit.* Thirring makes essentially the same criticism of his own theory. More recent analyses have shown that the theory is in fact inconsistent.

25. *Word and Object* (Cambridge, Mass.: M.I.T. Press, 1960), p. 21.

26. A discussion of this case is given in M. Friedman, *Foundations of Space-Time Theories,* unpublished Ph.D. thesis, Princeton, 1972.

27. Cf. my "Topology, Cosmology and Convention," *Synthese,* XXIV, 2 (August 1972): 195–218.

28. Cf. Franklin Fisher, *The Identification Problem in Econometrics* (New York: McGraw-Hill, 1966).

29. *The Philosophy of Space and Time* (New York: Dover, 1957).