ON AN INFORMATION-THEORETIC MODEL OF EXPLANATION*

JAMES WOODWARD†

Division of Humanities and Social Sciences
California Institute of Technology

This paper is an assessment of an attempt, by James Greeno, to measure the explanatory power of statistical theories by means of the notion of transmitted information (I_t). It is argued that I_t has certain features that are inappropriate in a measure of explanatory power. In particular, given a statistical theory T with explanans variables S_i and explanandum variables M_j, it is argued that no plausible measure of explanatory power should depend on the probability $P(S_i)$ of occurrence of initial conditions in the systems to which T applies or the magnitudes of the conditional probabilities $P(M_j/S_i)$, in the manner in which I_T does.

In a series of papers, James Greeno (1970, [1970] 1971), has suggested that we measure the explanatory power of statistical theories by means of the notion of transmitted information. In effect, Greeno's idea is that we think of explanation by means of a theory T with explanans variables $\{S_1,S_2,. . .,S_n\}$ and explanandum variables $\{M_1,M_2,. . .,M_M\}$ as the selection of a message S_i from the set of explanandums variables and its transmission through a noisy channel, where the characteristics of the channel are defined by the conditional probabilities $P(S_i/M_j) = P_{ji}$. If the prior probability of each message $S_i = P_i$, then the average information or uncertainty associated with the selection of an S_i from among the explanans variables will be

$$H(S) = -\sum_{i=1}^{n} P_i \log P_i. \qquad (1)$$

Similarly, we may represent the uncertainty regarding which S_i was transmitted when M_j is received as $H(S/M_j) = \Sigma_{i=1}^{n} P_{ji} \log P_{ji}$. If we let the probability of occurrence of $M_j = P_j$ and average the above quantity over all possible M_j, we have

*Received August 1984; revised May 1985.

†Portions of this paper were written while I was a Visiting Fellow at the Center for Philosophy of Science at the University of Pittsburgh in 1983. I am grateful to the Center and the University of Pittsburgh for their support. I would also like to thank Brian Barry, Ron Giere, James Greeno, Joseph Hanna, and Wesley Salmon for helpful comments on an earlier draft of this paper.

Philosophy of Science, 54 (1987) pp. 21–44.
Copyright © 1987 by the Philosophy of Science Association.

$$-\sum_{j=1}^{m} P_j \sum_{i=1}^{n} P_{ji} \log P_{ji}. \qquad (2)$$

We may think of (2) as the "equivocation"—the average amount of information lost as a message is transmitted through the channel.

It is thus natural to think of the difference between (1) and (2) as the average amount of information per message transmitted through the channel:

$$I_t = H(S) - \sum_{j=1}^{m} P_j H(S/M_j). \qquad (3)$$

A bit of algebraic manipulation shows that I_t may also be written as

$$I_t = H(M) - \sum_{i=1}^{n} P_i H(M/S_i), \qquad (4)$$

where

$$H(M) = -\sum_{j=1}^{m} P_j \log P_j,\ H(M/S_i) = -\sum_{j=1}^{m} P_{ij} \log P_{ij},\ \text{and } P_{ij} = P(M_j/S_i).$$

We may also write I_t as

$$I_t = H(S) + H(M) - H(SXM), \qquad (5)$$

where

$$H(SXM) = \sum_{i=1}^{n} \sum_{j=1}^{m} - P_i P_{ij} \log P_i P_{ij}.$$

Thus, (4) suggests that we may also think of I_t as a measure of the reduction of our a priori uncertainty $H(M)$ about which message will be received (which explanandum-event will occur) provided by our knowledge of P_i and the conditional probabilities P_{ij}.

Since everyone will agree that, in some sense, the explanans of a successful explanation provides information about the explanandum and since many find it natural to associate explanation with the reduction of uncertainty about the occurrence of an explanandum-event, there is a prima-facie case for adopting I_t as a measure of explanatory power. Moreover, as Greeno shows, I_t has other apparently intuitive features—for example, when the explanans and explanandum variables are all independent in the sense that $P_j = P_{ij}$ for all j, then $I_t = 0$; while for a given $H(S)$ and $H(M)$, I_t will reach a maximum when the conditional probabilities P_{ij} are all either 1 or 0, and deductive-nomological explanation is possible.

While Greeno's is perhaps the most fully developed and defended information-theoretic model of explanation, a number of other writers have made similar suggestions. For example, Roger Rosenkrantz and Richard Jeffrey have both proposed that we measure explanatory power by more complex functions of I_t (Rosenkrantz 1970; Jeffrey 1970). Moreover, Wesley Salmon has emphasized the similarities between Greeno's model and his own S-R model of explanation and has attempted in part to motivate adoption of the S-R model by an appeal to information-theoretic considerations (Greeno [1970] 1971, p. 32ff.; Salmon 1971, pp. 11–17; and 1977, p. 154). Given the apparent naturalness and attractiveness of the idea that we measure the explanatory power of a theory by means of the information it transmits, it is surprising that these proposals have received so little critical attention.

In this paper, I shall attempt to assess the idea that we measure the explanatory power of a theory by reference to the information it transmits. My remarks will focus almost entirely on Greeno's model. I shall argue that an acceptable measure of explanatory power should not be a function of (i) $H(S)$ or $H(M)$, or (ii) a function of $H(SXM)$, at least when these quantities are understood as they are in Greeno's treatment. While I believe that the criticisms I develop will also apply, at least in part, to many of the other proposals mentioned above, I shall not attempt to argue for this claim here. Finally, in an appendix to this paper, I shall take up the suggestion that I_t is an appropriate measure of explanatory power because it resembles a measure of explanatory power (proportion of the explained variance in the dependent variable M) that is sometimes adopted by social and behavioral scientists. As we shall see, objections that are very close to those I shall advance against I_t have also been advanced against the use of the proportion of variance explained as a measure of explanatory power. This fact seems to me to provide additional support for the objections I shall urge against I_t.

Before attempting to assess Greeno's model there are two initial interpretive problems that must be addressed. The first has to do with Greeno's conception of a theory and the second with the nature of the explananda within Greeno's model. Greeno tells us that

> formally a theory is a triple $\langle \Omega, A\ P \rangle$ where Ω is the domain, A is a Borel field of the subsets of Ω, and P is a probability function with domain A. ([1970] 1971, p. 90)

This definition, which is of course just the usual definition of a probability space, means that for Greeno a theory consists not just in a specification of the *conditional* probabilities $P(M_j/S_i)$, but also consists in a specification of the *marginal* probabilities $P(S_i)$ and $P(M_j)$ of occurrence of the various possible values of the independent and dependent variables.

Given two theoretical structures that employ the same explanans and explanandum variables and that specify exactly the same conditional probabilities $P(M_j/S_i)$ but that are applied to two quite different populations, with different probabilities or frequencies of occurrence of the values of the independent variable, Greeno will speak in this case not of one theory being applied to two different populations, but rather of two different theories. Indeed, Greeno must say this, since, as we shall note explicitly in a moment, and as is perhaps already clear from a casual inspection of (3), the explanatory power, as measured by I_t, of the above two theoretical structures can differ greatly.

This conception of a theory has another, closely related consequence. As Greeno's discussion makes clear, he thinks of the marginal probabilities $P(S_i)$ and $P(M_j)$ as identical with or estimated from actual relative frequencies within the population under investigation.[1] This means that the relevant conditional probabilities $P(M_j/S_i)$ for this population are also determined by facts about the above frequencies. In particular, if a certain possible value S_i of the explanans variables does not occur in the population, so that $P(S_i) = 0$, then of course the conditional probability $P(M_j/S_i)$ will be undefined, on the standard [Kolmogorov] axiomization of the probability calculus.

Greeno's conception of what a theory is thus differs sharply from what we might regard as the standard conception among philosophers and natural scientists, according to which a theory is understood as telling us nothing about the actual distribution of values of the explanans variables, but is understood as specifying instead what we might call hypothetical conditional probabilities, which give the probability distribution of the values of the explanandum variable conditional on every possible value of the explanans variable regardless of whether all such values actually occur in the population under discussion. It is in this sense of "theory," that, say, quantum mechanics and general relativity are theories.

Indeed, typical statistical theories in the natural and social sciences have yet another feature not represented in Greeno's analysis. Such theories do not merely specify a probability distribution for the various values of the explanandum variable conditional upon various possible values of the explanans variable, but typically involve much more general laws and theoretical principles from which these conditional probabilities are derivable in a unified and systematic way, given an appropriate specification of initial and boundary conditions. Elementary quantum mechanics does not just consist of a vast collection of separate claims about the probabilities that particular physical systems will undergo certain transitions—if it did, physicists would probably not be inclined to think of it as pro-

[1] See, for example, the remarks on college going behavior in his [1970] 1971, 94.

viding very deep explanations. Rather, elementary, non-relativistic quantum mechanics provides, by means of Schrödinger's equation, a unified general procedure for obtaining such probabilities for a very large number of different systems and for understanding how they will evolve through time. Similar remarks hold for population genetics and statistical mechanics, and for many examples of statistical theories in the social sciences.[2]

Within Greeno's discussion, questions about how we should conceive of theories and the grounds that are relevant to assessing their explanatory power are closely interconnected. For example, one of the criticisms I shall develop below of the use of I_t as a measure of explanatory power is most naturally expressed, in terms of the standard notion of theory, as follows: when the same theory (standard sense) is applied to two different populations, the explanatory power of the theory with respect to those populations should not depend on the probability or frequency of occurrence of initial conditions in those populations in the way that I_t does. Phrased in terms of Greeno's more idiosyncratic conception of a theory, the criticism is that there is no reason to think that two different theories, which employ the same explanans and explanandum variables, which specify the same conditional probabilities $P(M_j/S_i)$, and which differ only in the marginal probabilities $P(S_i)$, $P(M_j)$, differ in explanatory power as I_t requires. Put this way, my criticism is, among other things, an argument against Greeno's notion of a theory and in favor of the standard conception. If two theories, in Greeno's sense, which differ only in the marginal probabilities $P(S_i)$ they assign, do not differ in explanatory power, we lose the most obvious motive for regarding them as two different theories. Similarly, I shall argue below that, in assessing the explanatory power of a theory, we should focus on certain facts about the hypothetical conditional probabilities (in the sense described above) I_t specifies and not the actual conditional probabilities $P(M_j/S_i)$, as I_t requires. Here again, to make this criticism is in effect to object to Greeno's notion of a theory, according to which a theory will not include a specification of the former probabilities.

There is one other feature of Greeno's model that requires brief comment before we proceed. The models of statistical explanation associated with such writers as Hempel and Salmon, which are most familiar in the philosophical literature, are models of what Greeno calls "single explanations" of singular explananda. While the details, of course, vary greatly, the intent of such models is to delineate the conditions under which the fact that a particular (kind of) individual, belongs to some class or pos-

[2] For a clear and accessible discussion of some examples of statistical theories in the social sciences having this character, see Goldberg, 1983.

sesses certain properties is explained by the provision of appropriate statistical information. Thus, to use Greeno's example, the fact that a certain boy Albert, from an affluent, urban background, is convicted of a major crime, might be explained by the provision of appropriate statistical information about the probability of conviction among boys like Albert (and, in Salmon's case, other statistical information as well). Greeno's discussion makes it clear that his measure I_t is *not* intended to evaluate single statistical explanations in this sense. Rather, the measure I_t is a "global" measure meant to apply to entire theories, which evaluates how well a theory does overall in providing explanations of all of the explananda in the domain in which it applies. Understood in this way, Greeno's model differs from the more familiar models of Hempel and Salmon not in providing (or relying on) a novel account of the structure of particular statistical explanations, but rather in being directed at a different *object* of evaluation—the entire theory. Using Greeno's own example, the idea is roughly this: Suppose we are given a theory T of juvenile delinquency which assigns a high probability to boys like Burt, from poor, urban homes, becoming juvenile delinquents and a low probability to boys like Albert, from affluent, urban homes, becoming juvenile delinquents. Then I_t is not intended to suggest—and it would be a mistake to conclude—that T provides a better explanation of Burt's delinquency than Albert's because of this difference in conditional probabilities. However, given theories T_1 and T_2 designed to apply to different populations of boys, I_t will tell us that, other things being equal, T_1 has more explanatory power than T_2 if the marginal and conditional probabilities specified by T_1 differ from those of T_2 in the appropriate way—if, for example, the conditional probabilities for boys becoming juvenile delinquents in various conditions (S_i) are closer to 1 and 0 than the corresponding conditional probabilities for T_2. In short, we replace the project of evaluating single explanations on the basis of probability assignments to particular explananda, with the project of evaluating entire theories on the basis of the overall pattern of probability assignments they make.[3] The measure I_t is taken by Greeno to be relevant to the latter project, not the former.

[3] I thus take Greeno's model to be one in which what is explained—the "values" of the explanandum variable M—is, roughly, membership in a class of individuals, or possession of a property by an individual. However, what we evaluate, in using I_t, is how well entire theories do in explaining all of the explananda, so construed, in their domain. In his comments on an earlier version of this essay, the anonymous referee suggests a different interpretation. According to this interpretation, what is explained is (something like) facts about relative frequencies (values of $P(M_i)$ for various M_i), and these are explained by reference to an explanans consisting of the frequency distribution of the S_i—that is, $P(S_i)$ for $i = 1, \ldots, n$—and the conditional probabilities $P(M_j/S_i)$ for $i = 1, \ldots, n$. Such explanations might be regarded as deductive in structure, involving derivations of $P(M_j)$ by way of the relationship $P(M_j) = \Sigma_{i=1}^{n} P(S_i) \cdot P(M_j/S_i)$. Alternatively, one might think of the conditional probabilities as expressing probabilistic laws; in which case these and the frequency dis-

1. Dependence on $H(S)$ and $H(M)$. A fundamental intuition about explanation that many people share is this: Given a scientific theory (in the standard sense of a theory) which applies to a variety of systems characterized by different initial conditions, what matters, in explaining why one of those systems, R_i, is in a certain state, are the initial conditions *IC* that actually obtain in that system and the general pattern of causal and nomological relationships specified by the theory. Information about the relative frequency with which initial conditions of type *IC* occur among all of the systems to which the theory applies is (precisely because it is not information which is causally or nomologically relevant) simply irrelevant for the purpose of explaining the behavior of R_i. If in fact 90 percent of all bodies which have fallen or ever will fall to the surface of the earth fall for less than 3 seconds, and only 5 percent fall for more than 3 and less than 4 seconds, this information is simply irrelevant in determining how well Newtonian mechanics explains the behavior of those bodies. If we are given an electron with a certain kinetic energy in a potential well of a certain shape, then quantum mechanics provides just as good (or as bad) an explanation of why the probability of making a momentum measurement within a certain range on the electron is P_i if electrons in such potential wells and with such kinetic energy happen to occur very frequently or very rarely. Moreover, such information is equally irrelevant in determining how well a theory explains the behavior of a

tribution for the S_i are thought of as explaining why the actual frequency of M_j is close to the predicted frequency $P(M_j)$ by showing that this outcome is, in a sufficiently large population, "probable." The referee also suggests that this alternative conception of the explananda with Greeno's model vitiates the criticisms I advance in sections 1 and 2. I shall confine myself to two brief comments.

First, I think that a number of the passages in which Greeno is being most careful and explicit support my interpretation. For example, Greeno tells us that the intended interpretation of $\{M\}$ "is a variable or set of variables whose values are to be explained" and his most explicit example is one in which the variable M in a theory of juvenile delinquency takes the possible values (D_1 = no convictions, D_2 = minor convictions only, D_3 = major convictions) (p. 90). Second, I concede that there are a number of points in Greeno's discussion at which it is simply unclear what is being explained—it might be particular events of property possession by individuals, or kinds of events, or relative frequencies, or perhaps other possibilities as well. I think there is a very good reason for this. The appropriateness of I_t as a measure of explanatory power does not seem to depend, to any very large degree, on how we construe the explanada in Greeno's model, and this is why Greeno can be casual about such questions. (What distinguishes Greeno's model is not that it embodies some special conception of what is being explained—relative frequencies rather than more particular explananda—but rather the global character of the measure I_t.)

The criticisms I shall advance will be rather robust under different interpretations of what is explained within the context of Greeno's model. As I show in note 9, the criticisms I shall develop in sections 1 and 2 work just as well (or badly) if we think of the explananda of Greeno's model as relative frequencies. As I show in the appendix, they work equally well or badly if we think of statistical theories as explaining proportions of the variance of the dependent variable.

population composed of different kinds of systems. If how well Newtonian mechanics explains the behavior of falling bodies of type R_i does not depend on the frequency of occurrence of bodies of type R_i, there is no reason to suppose that the explanatory power of Newtonian mechanics with respect to a population of systems, a certain proportion of which are R_i, depends on the value of that proportion. Put in terms of Greeno's conception of a theory, the idea is that theories that differ only in the probabilities of occurrence of the initial conditions in the systems to which they apply and that employ the same laws linking explanans and explanandum variables should not differ in explanatory power. Given two quantum-mechanical theories that differ only in the fact that one specifies a population in which electrons occur in potential wells of form V_1 and V_2 equally often, and the other of which specifies a population in which such potential wells occur with unequal frequency, there is no reason to suppose that such theories differ in explanatory power.

The idea that the adequacy of the explanations provided by a theory should not depend on the frequency of occurrence of the initial conditions to which the theory makes reference (whether one thinks of explanations of the behavior of particular systems or explanations of the behavior of a whole population, composed of different kinds of systems) is widely recognized as a methodological requirement. For example, the biologist Richard Lewontin seems to appeal to essentially this methodological requirement when he criticizes the use of the analysis of variance to make straightforward inferences about causal and explanatory relationships, on the grounds that the results of such an analysis reflect not only lawlike or functional relations among variables (which is what interests us when our concern is with scientific explanation or tracing causal connections) but also the actual distribution of those variables in a population (Lewontin [1974] 1976, p. 189). As we shall see in more detail in the appendix, the sociologist Hubert Blalock appeals to essentially the same idea when he argues in favor of the use of unstandardized regression coefficients and against the use of correlation coefficients or path coefficients in sociological theorizing. To take another example, it is plausible to think that part of what is wrong with teleological modes of explanation, or explanations that purport to explain an occurrence by locating it within a sequence of "normal" development, is that they violate this requirement—certain outcomes are treated as more readily explainable simply because (in effect) they are the outcomes of initial conditions that occur more frequently or usually than others. Part of what is distinctive about a modern, post-Aristotelian conception of scientific explanation is that it takes cases represented by initial conditions that occur rarely or even not at all neither more nor less seriously for purposes of explanation than cases that occur frequently.[4] (Thus, it is recognized that an adequate theory of motion must account for the case, which may never occur in a universe full of gravitating masses, of a particle moving under the influence of no external force.) Moreover, just as the explanatory power of a scientific theory ought to be independent of information about the frequency of occurrence of initial conditions, so it also ought to be independent of information about the frequency of occurrence of explananda-phenomena, since these are in part a function of the frequency of occurrence of initial conditions.

It is easy to see that the use of I_t as a measure of explanatory power violates these methodological requirements. To see this we need only recall that I_t is a function of $H(S)$ and $H(M)$ and these depend upon the probabilities or relative frequencies of occurrence, P_i, of the initial conditions or explanans variables S_i and the relative frequencies of occurrence, P_j, of the outcome or explanandum-variables M_j. Indeed, it is easy to show that the expression $-\Sigma_{i=1}^{n} P_i \log P_i$ will attain its maximum when $P_1 = P_2 = \ldots = P_n$ and will go to zero when one of the P_i is equal to one and the remainder equal zero. Given a set of theories (in Greeno's sense) $\{T^1, T^2, \ldots, T^k\}$, each with explanans variables S_1 and S_2, and the same conditional probabilities, the member of the set with the greatest explanatory power will be the theory for which $P(S_1)$ and $P(S_2)$ are as close as possible to one half, and those theories for which $P(S_1)$ and $P(S_2)$ become more unequal will progressively diminish in explanatory power.

Greeno is of course aware of this feature of his account, and even suggests that it is a virtue. He writes I_t as (4) and remarks:

> (4) says that the information transmitted cannot be greater than the uncertainty of the explananda. In effect, if we can do quite a good job explaining the explananda without the theory, then the theory cannot help as much as it could if the explananda were more uncertain. For example, it could be quite interesting and useful to investigate the factors that relate to whether a person from a middle-class family attends college or not—some do and others do not. On the other hand, there would be less opportunity to increase the information transmitted by investigating factors that might relate to college attendance by sons of upper-class families, since almost all of these young men attend college. ([1970] 1971, p. 94)

Elsewhere Greeno notes that if all members of a population share the property M_1, so that $P(M_1) = 1$, then $I_t = 0$ regardless of the conditional probabilities P_{ij}. Greeno remarks:

[4]For a more detailed development of this idea, see my 1980.

> Theories can be trivial for many reasons, but one way to produce triviality would be to try to explain a characteristic that is shared by all members of a domain. For example, physical anthropologists are not interested in explaining why the members of some primitive tribe have two legs. ([1970] 1971, p. 93)

There are at least two difficulties with this response.

(a) Even if it were true that a theory that attempts to explain why all members of a domain possess a certain property will always be trivial, it is not at all clear why the explanatory power of a theory should depend on $H(M)$ in the manner that Greeno's measure suggests. Consider two theories (in Greeno's sense) T and T' both of which identify the same factors S_i as causally relevant to going to college and both of which correctly specify the same conditional probabilities of someone going to college, given these S_i. It is not at all clear why we should believe that T provides a better explanation than T' if exactly half of the kind of population with which T is concerned go to college and $^3/_4$ or $^1/_4$ of the population with which T' is concerned go to college. (Certainly this information is not causally or nomically relevant to anyone's college-going behavior.)

(b) Moreover, it is false that any theory that attempts to explain a property shared by all members of a domain is inevitably trivial. If all human beings in a certain population are born with two legs, modern genetics will explain why this is so. This theory will remain explanatory even if it were the case that (that is, people's genetic structure happened to be such that) every human being who ever lived was born with two legs.

I suspect that Greeno has been misled here by a perfectly sound ıdea: to explain why some particular A is B, and *a fortiori*, to explain why all A's and B's, one must (among other things) find some factor such that if that factor were to be different, then A would not be a B. We think that the presence of a certain genetic structure in human beings helps to explain why they are born with two legs because we think that if this genetic structure were appropriately different, those persons would have been born without two legs. It is important to understand, however (and this is where Greeno is misled), that this requirement on explanation is a requirement about what would happen under certain *hypothetical* or *counterfactual* conditions, and not a requirement that demands the actual occurrence of those conditions. What is essential to explaining the occurrence of human beings with two legs is not that human beings with one or no legs should actually occur, but that we be able to specify the conditions under which they *would* occur. In just the same way, what is important to explaining why a particle moves in a certain way under the influence of a force is that we be able to specify how it would move in the absence of any force, and not that situations in which the particle is free of all forces should actually occur.

We count ourselves unable to explain why some A is B when we think it *impossible* (as a matter of logic, or perhaps, of the fundamental laws of nature) for an A to be other than B or when we have no sense for the conditions under which A would be other than B. Thus, we do not think it possible to provide a (causal or nomological) explanation of why some equation that reduces to a mathematical identity holds.[5] Similarly, if, as Kripke and others have argued, genuine property identities are necessary truths, it is not surprising that we regard such identities as providing a natural stopping point in explanation. In the same way, we regard ourselves as unable at present to explain (in terms of some deeper theory) why the field equations of general relativity hold, because we do not know how those equations would have been different had certain other conditions been different. Contrary to what Greeno claims we will, given that all A's are B's, think it possible to explain why some A is B (or why all are) as long as it is possible for A's to be non-B's (whether because the above generalization is an accidental truth or because it is a derivative law of nature, resulting from the application of a more fundamental law to initial conditions that could have been otherwise), and when we know the conditions under which this possibility will obtain.

Essentially the same point can be made about a similar defense of I_t as a measure of explanatory power advanced by Salmon. Salmon maintains that it is appropriate for I_t to be zero, when $P_{ij} = P_j = 1$ for some j, because such a case "corresponds to the case in which dedutive-nomological explanation becomes vacuous through failure of relevance conditions, as in the example of the man who takes birth-control pills" (1971, p. 15). But a purported explanation of why X, who is male, fails to get pregnant in terms of his failure to take birth-control pills is *not*, as Salmon seems to suggest, analogous to the cases discussed by Greeno in which one undertakes to explain why all members of an actual population possess a certain property by citing some other property shared by all members of that population. If we were to randomly select a population of males, give them birth-control pills, and then compare this population with another randomly selected population of males who were not given birth-control pills, we would notice no difference in the incidence in preg-

[5] In his "Explanations, Tests, Unity, and Necessity" (Glymour 1980), Clark Glymour describes a number of cases in which an apparently contingent relationship between variables is shown to reduce to a mathematical identity when appropriate substitutions are made. He notes that such derivations represent especially satisfying explanations; we are not tempted to raise further questions about why the mathematical identity should hold.

nancy in the two populations. If we let S = X takes birth-control pills and M = X gets pregnant, then with respect to these two hypothetical populations, it will be the case that $P(M/S) = P(M/\bar{S})$. By contrast, a randomly selected population all of whom share a genetic factor, G, which is causally responsible for two-leggedness, T, and a randomly selected population all of whom have been genetically altered so that they lack G will differ in the incidence of T. With respect to these hypothetical populations it will not be the case that $P(T/G) = P(T/\bar{G})$. Instead, we will have $P(T/G) > P(T/\bar{G})$. It is this difference between the behavior of the conditional probabilities in these two cases that reflects the difference in their status as explanations. Thus, the reason why the fact that X, who is male, takes birth-control pills fails to explain his failure to get pregnant is *not*, as Greeno's account would suggest, that because all males fail to get pregnant, no explanation is possible for this explanandum. (The absence of appropriate reproductive organs does, one would think, help to explain why X fails to get pregnant, even though X is a member of a population all of whom fail to get pregnant.) Rather, it is the fact that $P(M/S) = P(M/\bar{S})$ that indicates that the birth-control-pills explanation is vacuous.[6,7]

We thus see again, from a slightly different perspective, that the mere fact that $P(M/S) = P(M) = 1$ for some actual population does not, as both Greeno and Salmon appear to suggest, show that an explanation of M in terms of S is vacuous. Even in the rather special case in which all members of a population share a certain property, it is not, as I_t and Greeno's understanding of what a theory is would suggest, information about the actual probability of occurrence of that property in the population which matters in accessing an explanation of why the members

[6] It is true enough that if $P(S) \neq 1$, then given that $P(M/S) = P(M)$, we can infer, by elementary algebra, that $P(M/S) = P(M/\bar{S})$. But in the case in which $P(S) = 1$ for the actual population under investigation, we can have $P(M/S) = P(M)$ (with $P(M/\bar{S})$ undefined) for that actual population even though the conditional probabilities $P(M/S)$ and $P(M/\bar{S})$, which could obtain in the two hypothetical populations all or none of whose members possess S, are such that $P(M/S) > P(M/\bar{S})$.

[7] To introduce even hypothetical conditional probabilities of form $P(Y/X)$ is of course to think of both X and Y as random variables, with a well-defined joint probability distribution. One might well wonder what such an assumption means when X is an independent variable, which is capable in principle of experimental manipulation or which is not in any obvious sense viewable as the result of a chance process. (Consider "time" as an independent variable in a dynamical system.) Such skepticism about the use of even hypothetical conditional probabilities to express lawlike or nomological relations is fully in the spirit of my argument in this paper, but would lead us into a number of issues that I lack the space to explore here. Here my point is that *if* one wishes—as virtually all philosophers of science who write on probabilistic causation or statistical explanation do—to express facts about causal or nomological connections between X and Y by means of conditional probabilities, what matters is the relationship between the hypothetical conditional probabilities $P(Y/X)$ and $P(Y/\bar{X})$.

possess this property, but rather the sort of hypothetical information about conditional probabilities described above.

Greeno also attempts to defend the dependence of I_t on $H(S)$. He writes I_t as (3) and then remarks:

> equation (3) relates to the fact that it is impossible to obtain any explanatory power by using an explanans that has only one value in the population. ([1970] 1971, p. 94)

Here too, a correct intuition underlies this claim: an explanans variable that *can* (for example, as a matter of logic) assume only one value cannot by itself (causally or nomologically) explain anything. But an explanans variable S_i that as a matter of fact assumes just one value in a population (for example, a common human genetic structure) can perfectly well explain why all members of the population have some property M_1 as long as it is possible for S_i to assume a different value and as long as, given this different value, the members of the population would not have M_1.

2. Dependence of I_t on $H(MXS)$. On Greeno's model, as we have seen, explanatory power is also in part a function of the values of the conditional probabilities P_{ij}. For example, for explanans variables $\{S_1,. . .,S_n\}$ with fixed probabilities of occurrence, I_t will be at a maximum when all the P_{ij} are either 1 or 0, and at a minimum when all the P_{ij} are equal. However, while the overall pattern of conditional probability assignments affects the overall explanatory power of a theory, Greeno rejects, as we have seen, the claim (let us call it (H)) that with respect to single statistical explanations, an explanation of some particular M_j in terms of some S_i is better the higher the conditional probability $P(M_j/S_i)$. Now, as we have already in effect noted (pp. 31–33), if the claim that the explanatory power of a theory depends on the overall pattern of probabilities P_{ij} is ever to be prima-facie plausible, those conditional probabilities cannot be just the actual conditional frequencies that obtain in the population under investigation. For one thing, $P(M_j/S_i)$ construed as such an actual frequency will be undefined when S_i does not occur in the population, and yet facts about the probability of M_j were (contrary to fact) S_i to occur in the population—facts about hypothetical conditional probabilities—can be quite relevant to the assessment of explanatory power. However, there are difficulties with the use of I_t as a measure of explanatory power which will remain even if the conditional probabilities occurring in I_t are construed as hypothetical conditional probabilities. In what follows, I shall argue that considerations that are very similar to those that tell against views like (H) as a thesis about single statistical explanations also make it problematic that the overall explanatory power of a statistical theory depends on the conditional probabilities P_{ij} *in the* *manner required by I_t*,[8] even when those conditional probabilities are construed as hypothetical conditional probabilities. To see this, let us recall one of the primary difficulties with the simplest form of (H): this is the thesis that an explanation of M_j in terms of S_i is better the higher the hypothetical conditional probability $P(M_j/S_i)$. Consider a system A consisting of a flux of electrons, each moving with the same kinetic energy along the X-axis and encountering the same one-dimensional potential barrier. Suppose that M_1 is the event of some particular particle being transmitted, and M_2 is the event of some other particle being reflected. How are M_1 and M_2 to be explained? Well, elementary quantum mechanics certainly explains why these events occur with certain probabilities, and in (very roughly) the following fashion. Given the Hamiltonian governing the incoming particles, and making various other assumptions about symmetries, what happens at boundaries, etc., one solves the time-independent Schrödinger equation for the wave functions representing the transmitted and reflected waves. The probability that a particle will be transmitted or reflected is then given by the square of the absolute value of the transmitted or reflected wave. Let us call the common initial conditions to which the above derivations appeal S_1. Suppose that when we carry out the above derivations, we obtain $P(M_1/S_1) = 0.75$ and $P(M_2/S_1) = 0.25$, and that these conditional probabilities are very close to the frequencies of transmission and reflection actually observed in the system under discussion. Suppose that we also concede, for the sake of argument, that the above derivations are properly thought of as expla-

[8] I should emphasize here that my disagreement is with the claim that the explanatory power of statistical theories depends on the magnitudes of the conditional probabilities P_{ij} in the specific way claimed by I_t. I do not make the (patently false) general claim that the explanatory power of a statistical theory has nothing to do with the overall pattern of conditional probabilities it ascribes. Indeed, I have already urged above that a necessary condition for some particular value of S_i to be explanatory with respect to some M_j is that $P(M_j/S_i) \neq P(M_j/\bar{S}_i)$. However, one can agree to this without adopting the specific claims about explanatory power represented by I_t. Again, if we adopt the customary distinction between potential and actual explanations, and require that for a theory to be actually explanatory, it must be at least approximately true, then clearly the explanatory power of a theory T will depend on whether the conditional probabilities it posits or predicts are at least approximately true or empirically correct. (Of course, in Greeno's case, the conditional probabilities P_{ij} are apparently just the conditional frequencies actually exhibited in the population under investigation, so these can hardly fail to be empirically accurate. However, my discussion here is meant to include the more general case in which we have a theory (standard sense) that predicts various conditional probabilities, and questions about the empirical adequacy of these arise.) This is thus another respect in which explanatory power of T depends on the values it ascribes to the conditional probabilities it specifies. But here too, one can agree to this without endorsing the specific measure embodied in I_t. One might—as I do—take the view that given two theories T_1 and T_2 applying to different populations, as long as the conditional probabilities they ascribe are empirically accurate (where this is a matter to be ascertained by the use of standard procedures for statistical testing, such as tests of significance), the explanatory power of T_1 and T_2 should not depend on whether those values happen to be close to 1 or 0.

nations (let us call them respectively (Ex. 1) and (Ex. 2)) of the *occurrence* of M_1 and M_2, and not just explanations of the probabilities with which they occur.

The defender of (H) wishes to hold that, simply because $P(M_1/S_1) = 0.75$ and $P(M_2/S_1) = 0.25$ the quantum mechanical explanation envisioned above of M_1 is better than the explanation of M_2. This claim is highly counter-intuitive. What is the relevant difference between the explanans of (Ex. 1) and the explanans (Ex. 2) that grounds this alleged difference in explanatory power? The explanans of (Ex. 1) is, as I have constructed the case, exactly the same as the explanans of (Ex. 2). Each explanans makes reference to the same initial and boundary conditions, described in the same way, and each invokes the same claim about general nomological connections (Schrödinger's equation) and the same general procedure for calculating probabilities from the ψ function that is the solution to that equation. In short, each seems to provide exactly the same information about causal and nomological connections. It is true that when we calculate the probabilities of M_1 and M_2, we arrive at different values; but why should this be taken to show that the above explanation is a better explanation of barrier penetration than of barrier reflection? The claim that the above explanations differ in explanatory power seems flatly inconsistent with the claim, which I see no reason to reject, that the explanatory power of an explanans depends upon the information about (relevant) causal and nomological connections it provides.

This objection to (H) is a familiar one—indeed similar (although not, I think, identical) objections to (H) have been repeatedly raised by Salmon (see 1971, 1977). However, if this line of objection is at all cogent, a very similar line of objection will apply to any measure of explanatory power which is a function of $H(SXM)$. Let P be a population of quantum mechanical systems one half of which are A (as characterized above) and the remaining half of which are of kind A'', which differ from systems of kind A in initial conditions and hence conditional probabilities for transmission and reflection. Let P' be a population of quantum mechanical systems half of which are A', where systems that are A' are identical with A except the initial conditions (the kinetic energy of the incident flux of electrons, the potential function) are now such that $P(M'_1/S'_1) = 1/2$ and $P(M'_2/S'_1) = 1/2$, where S'_1 are the initial conditions in A' and M'_1, M'_2 are the transmission and reflection of an electron in A'. Let the remainder of the systems in P' be of kind A''. Now, contrast the theory (Greeno's sense of theory) T' that consists in a specification of the marginal probabilities with which systems of kind A' and A'' occur in populations of sort P' together with the relevant conditional probabilities with the theory T that specifies the corresponding information with respect to populations of sort P. For the purposes of I_t the only relevant difference between T' and T

will have to do with the differences in the conditional probabilities of transmission and reflections in systems of kind A' and systems of kind A. In particular, the defender of I_t as a measure of explanatory power is committed to the view that simply because of this difference in conditional probabilities $P(M_1/S_1)$, $P(M_2/S_1)$ and $P(M_1'/S_1')$, $P(M_2'/S_1')$ for the two systems A and A', theory T has, in population P, more overall explanatory power than T' in population P'.

It is difficult to see what justification there could be for this claim. Suppose that one thinks of quantum mechanics as a theory in a more standard sense of theory, according to which the same theory can apply to populations of systems with different frequencies of occurrence of initial conditions and according to which the same theory can explain transmission and reflection in both A and A', even though the conditional probabilities for these are different in the two systems. From this perspective, the quantum-mechanical explanations (call them (Ex. 1′) and (Ex. 2′)) given for transmission and reflection in system A' are exactly like those given for system A except for the brute fact that A and A' differ in initial conditions and hence involve different probabilities for transmission and reflection. In explaining transmission and reflection in A' in terms of quantum mechanical theory, we will invoke the same general information about nomological connections and the same general procedure for calculating probabilities that we invoke in explaining transmission and reflection in A. We have already agreed that within A, the brute fact that initial conditions are such that $P(M_1/S_1) = 0.75$ and $P(M_2/S_1) = 0.25$ does not show that M_1 is better explained that M_2. Why should information about the relative magnitude of the conditional probabilities for transmission and reflection not matter at all when we confine ourselves to A, and yet matter crucially when we compare how well quantum mechanics explains the behavior of the systems in population P (which contains systems of kind A) with how well it explains the behavior of the systems in population P' (which contain systems of kind A')? It is not easy to understand what relevant difference in the causal or nomological information provided by quantum mechanics as applied to these two populations grounds this alleged difference in explanatory power. Moreover, there is not the slightest suggestion to be found in quantum mechanics textbooks that there is such a difference in explanatory power.[9]

[9] This is perhaps the appropriate place to comment briefly on an issue raised by the anonymous referee (see note 3). Would it make any difference to the cogency of the above criticisms if the explananda within Greeno's model were regarded as relative frequencies? It is hard to see why this should be the case. Consider—to return to an example of Greeno's—two theories, T and T', both of which specify the same explanans variable (S_1, S_2) (social class) as relevant to college-going behavior (M_1,M_2) and both of which specify the same conditional probabilities $P(M_j/S_i)$, but which are such that T specifies marginal probabilities $P(S_1) = P(S_2) = 1/2$ in the populations P' to which it applies and

Consider, by contrast, the kind of case in which it *would* be natural to say that a theory T (in the standard sense) explains one set of quantum phenomena A better than another set A'. For example, one would say this if T provided less accurate predictions (or no predictions at all) about some of the phenomena in A (or the probabilities with which they occur) than about the phenomena in A'. Again, one might say this if T in conjunction with appropriate information about initial and boundary conditions permitted the derivation of both A and A', but the derivations provided with respect to A were relevantly different (for example, involved different laws, or different assumptions about initial conditions) than the derivations provided with respect to A', and the latter derivations seemed more ad hoc, less coherent, less general, or more reliant on assumptions for which there was no independent theoretical rationale. (Similar considerations would, of course, be involved in a comparison of the explanatory power of two competing theories T_1 and T_2.) Thus, for example, one might say that Bohr's early quantum theory of 1913 provided a better explanation of spectral emissions of hydrogen (where it at least made accurate predictions) than it did of non-classical barrier penetration (which it did not predict at all). Again, one might say that Bohr's theory does not explain very well, if at all, why an electron in a potential well will occupy only discrete energy levels (Bohr's "quantum conditions" are imposed ad hoc, without any real justification besides the fact that they yield experimentally correct results), while modern quantum mechanics provides a much better explanation of this phenomenon (the quantization of

T' specifies $P(S_1) = 3/4$ and $P(S_2) = 1/4$ in the populations P to which it applies. Then, on the proposed new construals, using I_t as a measure of explanatory power, we must concede that T provides a better explanation of the relative frequency of college-going behavior in P than T' provides of the relative frequency of college-going behavior in P'. But this claim seems just as arbitrary and unwarranted as the earlier claim that T provides a better explanation of individual college-going behavior in P than T' provides with respect to P'. Both T and T' agree exactly on which variables are relevant to college-going behavior and on how each such variable is relevant (as expressed in the appropriate conditional probabilities). Both T and T' provide, we may suppose, equally accurate predictions about the relative frequency of college-going in P and P'. As before, it is difficult to see what causally or nomologically relevant information T provides that T' fails to provide and that might ground the claim that T provides better explanations than T'. Similarly, an argument parallel to the argument made in section 2 shows that under the proposed new construal, with frequencies as explananda, one still faces the problem that the explanatory power of a statistical theory depends on the values of the conditional probabilities P_{ij} in an unintuitive way. Given a theory T that specifies conditional probabilities $P(M_1/S_1) = P(M_2/S_2) = 3/4$ and $P(M_2/S_1) = P(M_1/S_2) = 1/4$ in the population P to which it applies and a second theory T' which specifies conditional probabilities $P(M_1/S_1) = P(M_2/S_1) = P(M_1/S_2) = P(M_2/S_2) = 1/2$ in the population P' to which it applies, the marginal frequencies $P(S_i)$ being the same in both populations, (and such that $P(M_j/S_i) \neq P(M_j)$ for any i, j), then use of I_t still requires us to conclude that the relative frequency of college-going behavior is better explained in P than in P', even if the relative frequencies are equally accurately predicted in both cases.

allowable energy levels arises in a natural way out of the imposition of certain boundary conditions on Schrödinger's equation).

In short, as a look at any standard history of quantum mechanics will confirm,[10] we assess the explanatory power of statistical theories by means of the same considerations that we use to assess the explanatory power of nonstatistical theories. The sort of differences in the overall pattern of probability assignment reflected in the term $H(SXM)$ do not by themselves reflect differences in the explanatory power.[11]

3. Conclusion. One common line of criticism of the S-R model in effect focuses on the point that not every case of providing grounds for thinking an explanandum obtains (or odds for betting on whether it will obtain) constitutes an explanation for that explanandum. The criticisms of Greeno's model I have advanced exploit the point that a similar conflation of "reason-seeking" and "explanation-seeking" questions seems

[10] See, for example, Jammer (1966), especially chapters 5 and 6.

[11] This is perhaps the appropriate place to comment on an alternative use of I_t as a measure of explanatory power, briefly suggested by both Greeno and Salmon in conversation: Suppose that we measure the explanatory power of a theory T' concerned with domain D by adopting an index which measures how "close" the information carried by T' is to the information carried by T, where T is an ideal theory, which specifies the correct statistical generalizations governing all the objects in D. The underlying intuition about explanation is that the more nearly T' approximates the correct probabilities specified by T the better the explanation T' provides.

As Greeno points out, this sort of conception of explanatory goodness can be used to provide at least a partial motivation for the adoption of I_t as a measure of explanatory power in the case of theories which are designed for a domain in which we think that the ideal theory would be nearly deterministic. Given this conception of explanatory goodness, and a deterministic domain, it becomes intelligible why, as I_T requires, explanatory power should become larger (given equal marginal probabilities) as the conditional probabilities P_{ij} become more unequal and nearer to 0 or 1, for this is an indication of how close the theory under assessment is to the true deterministic theory governing the domain. (Greeno thinks that ideal theories in psychology will typically be nearly deterministic, and thus regards I_T as an appropriate measure of explanatory power for psychological theories, the domain for which it was originally designed.)

Salmon suggests that a similar conception of explanatory goodness might be used to motivate an information-theoretic measure of explanatory power for a theory T' dealing with a domain governed by a nondeterministic ideal theory T, where here of course the measure in question will need to be some function of $I_{T'}$, which is relativized or "normalized" with respect to I_T.

Since this sort of attempt to motivate an information-theoretic measure of explanatory power represents a significant departure from the motivations suggested by Salmon and Greeno in the texts considered above, I shall not undertake to assess it here. I will note, however, that the most obvious candidates for an appropriately normalized measure, covering the nondeterministic case $I_T{}'/I_T$ and $|I_{T'} - I_T|$ will not do, since if the true theory T has roughly equal conditional probabilities, these expressions will decrease in value as we move from a seriously false theory to a theory closer to the truth; while the opposite will be the case for a true theory T with highly unequal conditional probabilities. This of course does not show that it may not be possible to find some more complex measure of the same general sort that behaves appropriately.

inherent in any attempt to measure explanatory power by transmitted information. What I_t measures is simply the average extent to which our uncertainty regarding the arrival of some message is reduced by the operation of a communications channel. From the perspective of the information theorist, who is interested in questions about, for example, the maximum rate that information can be transmitted with arbitrarily high reliability through a given channel, there is no reason to distinguish among the various ways or grounds in which this reduction in uncertainty is achieved. The basic problem is that not all grounds that are relevant to how much our uncertainty is reduced matter equally for purposes of explanation and not all reductions in uncertainty constitute explanations. There is, for example, a clear sense in which, if the same message is always received from a channel, the channel does not transmit any information. It is thus perfectly reasonable that, even if the conditional probabilities P_{ij} are all one or zero, I_t should be zero when $H(M)$ is zero. But for the purpose of explanation, what matters is not just any information about (the frequency of occurrence of) the explanandum-phenomenon, but rather information that is causally or nomologically relevant. Our prior information about the frequency of occurrence of some explanandum phenomenon is not causally or nomologically relevant information; and so even in the case of an explanandum-phenomenon that occurs with relative frequency equal to one, we learn something relevant for the purposes of explanation when we have identified the laws and initial conditions that are causally relevant to that phenomenon. We do not, it is true, learn anything more about the frequency with which the explanandum-phenomenon occurs, but this just highlights again the difference between explaining and providing grounds for expecting or betting.

A similar point can be made about the dependence of I_t on $H(SXM)$. As long as we are simply concerned with the total average amount of information transmitted, this dependence is perfectly reasonable; for everyone will agree that a channel in which, for a given input message S_i, the probability of some single associated output message M_j is very high and the probability of any other output message quite low, will be ideal and that a channel in which these probabilities are nearly equal will not be very efficient at transmitting information. But unless we make the illegitimate further move of identifying providing grounds for expecting or betting rates with explaining, there is no reason to think that this fact shows that the explanatory power of a statistical theory depends on $H(SXM)$.

APPENDIX

I have suggested above that, whether one thinks of statistical theories as explaining individual outcomes or relative frequencies, I_t depends upon the actual distribution of values of the explanans variables in a population in a way that makes it unsuitable as a measure of explanatory power. In this appendix, I examine the relationship between I_t and another population-specific measure of explanatory power—the product-moment correlation coefficient r_{xy}. As we shall see, r_{xy} (or actually r_{xy}^2), seems to embody a conception of explanation that closely resembles the conception embodied in I_t and to share many of the distinctive features of I_t. But for just this reason correlation coefficients and such related measures as path coefficients seem unsuitable as measures of explanatory power.

The product-moment correlation coefficient r_{xy} is a symmetric measure of the degree of linear relatedness between two random variables, X and Y, both measurable on an interval scale. The coefficient r_{xy} varies from $+1$ to -1 with $r_{xy} = 0$ corresponding to the case in which there is no linear relationship between X and Y and $r_{xy} = 1$ (-1) corresponding to the case in which there is a perfect positive (negative) linear relationship between X and Y. By definition,

$$r_{xy} = \frac{S_{xy}}{S_x S_y}, \tag{1}$$

where S_{xy} is the sample covariance of X and Y, and S_x, S_y are, respectively, the standard deviations of X and Y.[12]

For our purposes, the best way to get an intuitive sense for this measure (and for some of its limitations) is to think of it in the context of linear regression. Suppose that we make n observations of the random variables X and Y, where the ith values of these variables are represented by x_i and y_i. Suppose these are related by the following linear relationship:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{2}$$

where α and β are fixed coefficients, and ε_i is a so called "error" or "disturbance" term. The linear regression equation of Y on X will be the path of the means of Y for various values of X—that is, the expected value of Y conditional on X, $E(Y/X)$. If one uses a least-squares estimate[13] of α and β and makes certain assumptions about the distribution of the error term (in particular that $E(\varepsilon_i) = 0$ for all i, and that x_i and ε_i are uncorrelated for all i),[14] then it is readily shown that

$$b = \frac{S_{xy}}{S_x S_x} \tag{3}$$

and

$$a = \bar{y} - b\bar{x} \tag{4}$$

[12] I follow the usual convention of using roman letters to represent sample values and Greek letters to represent population values. Hence, r_{xy} is the sample correlation coefficient,

$$S_y = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}}, \text{ and } S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for x and y. The population correlation coefficient, which is of course what we are ultimately interested in, is $\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y$. Similarly α and β in (2) are regression coefficients in the population, while a and b are estimates of these from the sample.

[13] The least-squares estimate will involve the choice of values for α and β so that the quantity $Q = \Sigma_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ is minimized; that is, the choice of the straight line such that the sum of the squares of the deviations of the actual values of y from this line is minimized.

[14] To establish confidence in intervals or to make tests of significance, the further assumption that the ε_i are normally distributed with constant variance for all levels of X is commonly made.

represent unbiased estimates for α and β. (Here S_{xy} and S_x are, as before, respectively the sample covariance between X and Y and the standard deviation of X, and $\bar{y}$ and $\bar{x}$ are the sample means for X and Y.) One can think of r_{xy} as a measure of the spread of the value of Y about the regression line. Suppose that we let y_{p_i} be the value of y_i that would be predicted for each value of x_i from the regression equation (that is, $y_{p_i} = a + bx_i$). Let y_i be the actual value of Y associated with x_i and $\bar{y}$ be the mean of Y. Then we can show that

$$r_{xy}^2 = \frac{\sum_{i=1}^{n} (y_{p_i} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{5}$$

The quantity

$$\frac{\sum_{1=1}^{n} (y_{p_i} - \bar{y})^2}{n}$$

is often described as that proportion of the variance in Y that is "explained" by the variable X, since it is that portion that is predictable from the values of X. Thus r_{xy}^2 is said to provide a measure of the proportion of the variance in Y that is explained by X to the total variance in Y. A high (low) value of r_{xy} indicates that a large (small) proportion of the variance in Y is being explained by X. If $r_{xy} = 1$, then the values of Y are perfectly predictable from X—there is no spread at all about the regression line.

While there are a number of important differences[15] between I_t and r_{xy}^2, it is also clear that the two measures behave in broadly similar ways. For example, if X and Y are statistically independent, in the sense that $P(Y/X) = P(Y)$ for all values of X and Y, then $r_{xy} = 0$ and, as we have already noted, $I_t = 0$. And while it is not true in general that if $r_{xy} = 0$, X and Y will be independent, this will be true if the distribution of X and Y is bivariate

[15] From the perspective of the recent philosophical literature on statistical explanation, one of the most striking differences is that while in the case of the information-theoretic model it seems most natural to think that what is being explained as either individual outcomes or (perhaps) the probabilities with which those outcomes occur, it is clear that with r^2_{xy} or regression analysis, neither of these are plausible candidates for what is being explained. To begin with, while Y is thought of as a random variable with a definite probability distribution, individual values of Y that lie off the regression line are not thought of as "explained" at all; they are rather part of the unexplained variance. For the same reason, it is inappropriate to think of the probability distribution of the values of Y for a given X as what is explained. While one may make certain assumptions about this distribution for the purpose of setting confidence intervals for α and β (for example, that for each level of X, the distribution of the various values of Y will be normal, with fixed variance), knowledge of the distribution of values of the independent variable does not provide an explanation of *why* these assumptions should hold—the distribution governing the spread of the value of Y about the regression line for a given value of X is due to the operation of the unmeasured "disturbance" term and represents precisely what is "unexplained."

Instead, the use of correlation coefficients and regression analysis seem to embody the quite different idea that what is explained is a "population level" parameter: (a certain proportion of) the variance of the dependent variable, or perhaps the mean value of the dependent variable for each level of X. This conception of what statistical explanations explain has, to the best of my knowledge, been left entirely unexplored in the philosophical literature on statistical explanation.

normal. While r_{xy} is a measure of the degree of linear relatedness between X and Y and I_t embodies no such specific assumptions about the form of the relationship between X and Y (indeed makes no assumption that X and Y are measurable on an interval scale), both measures embody the idea that explanatory power will be maximized in the case in which there is a perfect deterministic relationship between X and Y.[16] And perhaps most interestingly, the correlation coefficient r_{xy}, like I_t, is population specific in the sense that it is a function of the actual distribution of values of X and Y in the population under investigation, and not just the causal relationship between X and Y. The easiest way to see this is to rewrite r_{xy} (1) as

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{S_{xy}}{S_x S_y}\left(\frac{S_x}{S_x}\right) = b_{yx}\left(\frac{S_x}{S_y}\right). \tag{6}$$

One can think of b_{yx}—the coefficient of X in the regression equation (2)—as an indication of how much change on the average in Y would be produced by a unit change in X. As a number of writers have suggested, this coefficient, and the regression equation (2), can be regarded as reflecting (something like) the functional, causal, or nomological relationship between X and Y.[17] Further, (6) shows us that the value of r_{xy} is a function, not just of this causal relationship, but of the actual distribution of the values of X and Y—in particular the (square root of the) variances of X and Y—in the population under discussion. Given two populations in which exactly the same functional relationship between X and Y obtains (as indicated by b_{yx}) the value of r_{xy} can vary from 0 to 1, depending upon the amount of variation in X which happens to be present in the population in proportion to the amount of variation in Y that is due to the operation of the disturbance term (where this in turn depends upon the amount of variation that happens to be present in those causes of Y other than X).

This feature of correlation coefficients has led many writers to regard them as inevitably misleading measures of the degree of causal or explanatory connectedness between variables. Consider, for example, discussions of the influence of heredity and environment on some phenotypical trait such as IQ. It frequently has been noted that one may have two populations, in both of which IQ is fully determined by the same function of both genetic and environmental variables, and yet, if the first population is one which exhibits little genetic variation and considerable environmental variation, the correlations between IQ and genetic background will be quite low, while if the second population is one in which there is considerable genetic variation, and relatively little environmental variation, the correlation between IQ and genetic background will be quite high. As in the case of I_t, given a population which is perfectly homogeneous for some genetic trait G that plays a causal role in the production of phenotypical trait P, the correlation between G and P will be zero. In short, the presence of a high or low correlation between genetic structure and IQ in some population does not by itself tell us what the value of this correlation would be in some other population, exhibiting different amounts of genetic or environmental variation, or anything about how changes in environmental or genetic structure would affect IQ scores; and yet, it is precisely this information that is of interest for the purpose of causal understanding (or social policy). For just this reason, writers on genetics commonly caution that correlation coefficients are not to be understood as measures of the "extent" to which one variable causes or determines another.

Similar reservations regarding the use of correlation coefficients and path coefficients (which also possess many of the population-specific features of correlation coefficients) and a preference for unstandardized regression coefficient are expressed by Hubert Blalock. Commenting on a study (Miller and Stokes 1963) that attempted to determine the corre-

[16] The reader may recall, in this connection, note 11 in which it is suggested that that use of I_T as a measure of explanatory power would make most sense if it were understood as a measure of the closeness of T to an ideal, deterministic theory T'.

[17] This interpretation of regression coefficients is particularly urged by Blalock. See his 1964 and 1967.

lation between various attitudes of a representative and the attitudes of his constituency. Blalock writes:

> Suppose one finds stronger correlations between constituency's and representative's attitudes in the North than is true in the South. It is quite conceivable that the same laws are operative, giving the same value of b_{yx} [the unstandardized regression coefficient] in each region yet there may be more variation in constituency's attitudes in the North, and if extraneous factors operated to the same extent in both regions, this would account for the larger correlation. [While the amount of variation in X can be measured], uncontrolled and unknown disturbing influences cannot, and one would have no way of determining whether or not these also varied more in the North than in the South. It would therefore be more meaningful to compare slope estimates [regression coefficients] than the respective correlations. (1964, p. 14)

Elsewhere, Blalock remarks that the variance of X (the exogenous independent variable) and the amount of variance in Y produced by factors other than X

> may be taken as "accidental" from the point of view of one's theory. . . . [The theorist] cannot account for the numerical values of the variance in exogenous variables. In this sense, they are taken as accidental, and unique to each population even where the same causal laws are operative on all populations. (1964, pp. 146–47)

Thus while r_{xy} does indeed resemble I_t in that both measures depend upon facts about the actual distribution of the independent and dependent variables in the population under investigations—facts that are in part "accidental" from the point of view of the causal laws operating in those populations—this resemblance seems to represent a limitation, rather than a virtue, of both measures.

Finally, let me note another similarity between I_t and r^2_{xy}, construed as measures of explanatory power. In both cases, explanatory power is, in effect, identified with the increase in predictive ability or reduction in uncertainty associated with knowledge of the dependent variable, relative to some initial state of uncertainty. Suppose that the value of X is not known at all. Then the best prediction for the value of Y would be μ_y—the population mean—or $\bar{y}$ assuming only sample data are available. However, if one knew the value of X, x_i, one would predict the value of Y, y_{pi}, which lies on the regression equation. What r^2_{xy} measures is how this improvement in one's ability to predict $Y[(y_{pi} - \bar{y})^2]$ compares with the original total variation in or uncertainty regarding the value of Y. This idea is often appealed to in attempts to motivate r^2_{xy} as a measure of explanatory power. Quite apart from the obvious objection that knowledge of spurious correlations, symptoms, etc., may improve one's ability to predict, we have already noted, in connection with I_t, a central difficulty with this way of conceiving of explanatory power. The relative size of the improvement in one's ability to predict Y that comes with the knowledge of X depends upon factors that have nothing to do with any causal or nomological relation between X and Y, but rather have to do with the actual distribution of X and Y in the population. Given two populations in which exactly the same causal relation holds between X and Y, the relative improvement in one's ability to predict Y may be either quite high or low, depending on whether the amount of variation in X that happens to occur in the population is relatively large or small.

REFERENCES

Ayala, Francisco (1982), *Population and Evolutionary Genetics*. Menlo Park, California: Benjamin Kummings Publishing.

Blalock, Hubert and Blalock, Ann B. (1964), *Causal Inference in Non-Experimental Research*. Chapel Hill: University of North Carolina Press.

———. (1967), "Causal Inferences, Closed Populations and Measures of Association", *The American Political Science Review 61*: 130–36.

Cartwright, Nancy (1979), "Causal Laws and Effective Strategies." *Noûs 13*: 419–37.

Fetzer, James (1981), *Scientific Knowledge*. Dordrecht: D. Reidel Publishing.

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

Goldberg, Samuel (1983), *Probability in Social Science*. Boston: Birkhauser.

Greeno, James (1970), "Theoretical Entities in Statistical Explanation", in *PSA 1970,* R. Buck and R. Cohen (eds.). Dordrecht: D. Reidel Publishing, pp. 3–26.

———. [1970] (1971), "Evaluation of Statistical Hypotheses Using Information Transmitted", in Salmon (1971). (Originally published in *Philosophy of Science 37*: 279–83.)

Hanna, Joseph (1978), "On Transmitted Information as a Measure of Explanatory Power", *Philosophy of Science 45*: 531–62.

Hempel, Carl (1965), *Aspects of Scientific Explanation*. New York: Free Press.

———. (1968), "Maximal Specificity and Lawlikeness in Probabilistic Explanation", *Philosophy of Science 35*: 116–33.

Jammer, Max (1966), *The Conceptual Development of Quantum Mechanics*. New York: McGraw-Hill.

Jeffrey, Richard (1970), "Remarks on Explanatory Power", in *PSA 1970,* R. Buck and R. Cohen (eds.). Dordrecht: D. Reidel Publishing, pp. 40–46.

Lewontin, Richard [1974](1976), "The Analysis of Variance and the Analysis of Causes", in *The IQ Controversy,* N. J. Block and Gerald Dworkin (eds.). New York: Random House. (Originally published in *American Journal of Human Genetics 26*: 400–411.)

Miller, Warren, and Stokes, Donald (1963), "Constituency Influence in Congress", *The American Political Science Review 57*: 45–56.

Niiniluoto, I. (1981), "Statistical Explanation Reconsidered", *Synthese 48*: 437–72.

Pierce, John (1970), *An Introduction to Information Theory*. New York: Dover Publications.

Rosenkrantz, Roger (1970), "Experimentation as Communication with Nature", in *Information and Inference,* Jaakko Hintikka and P. Suppes (eds.). Dordrecht: D. Reidel Publishing, pp. 58–93.

Salmon, Wesley (ed.) (1971), *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

———. (1977), "A Third Dogma of Empiricism", in *Basic Problems in Methodology and Linguistics,* Robert E. Butts and Jaakko Hintikka (eds.). Dordrecht: D. Reidel Publishing.

Schrader, Douglas (1977), "Causation, Exploration and Statistical Relevance", *Philosophy of Science 44*: 135–43.

Shannon, Claude, and Weaver, Warren (eds.) (1949), "A Mathematical Theory of Communication", in *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Skyrms, Brian (1980), *Causal Necessity*. New Haven: Yale University Press.

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

Woodward, James (1979), "Scientific Explanation", *The British Journal for the Philosophy of Science 30*: 41–67.

———. (1980), "Developmental Explanation", *Synthese 44*: 443–66.