Acta Biotheoretica 31, 145–164 (1982) 0001–5342/82/030145–20 $00.20/0

On inference in ecology and evolutionary biology: the problem of multiple causes

RAY HILBORN* and STEPHEN C. STEARNS**

*Institute of Animal Resource Ecology, University of British Columbia, Vancouver, B.C., Canada V6T 1W5; **Biological Laboratories, Reed College, Portland, Oregon 97202, USA

(*Received 8-X-1981*)

Summary. If one investigates a process that has several causes but assumes that it has only one cause, one risks ruling out important causal factors. Three mechanisms account for this mistake: either the significance of the single cause under test is masked by noise contributed by the unsuspected and uncontrolled factors, or the process appears only when two or more causes interact, or the process appears when there are present any of a number of sufficient causes which are not mutally exclusive. In ecology and evolutionary biology, experiments usually test single factor hypotheses, and many scientists apparently believe that hypotheses incorporating several factors are so much more difficult to test that to do so would not be practical. We discuss several areas in ecology and evolutionary biology in which the presupposition of simple causation has apparently impeded progress. We also examine a more mature field, the study of atherosclerosis, in which single factor studies did significantly delay progress towards understanding what now appears to be a multifactor process. The problem has three solutions: either factorial experiments, dynamic models that make quantitative predictions, response-surface methods, or all three. In choosing a definition for 'cause', we make a presupposition that profoundly influences subsequent observations and experimental designs. Alternative definitions of causation should be considered as contributing to potential cures for research problems.

1. Introduction

This paper examines how one presupposition about causation – that every 'properly analyzed' effect has a single cause – can bias inference, and in fact has retarded the analysis of population cycles in small rodents and of life-history traits. We argue that changes in concepts such as 'cause' can alter the scientist's perception of what constitutes a research problem just as powerfully as do changes in specific, empirical hypotheses. The idea of 'cause' appears here in two quite different roles. First, when one makes a statement like 'every effect has a single cause', one is making a *metaphysical* assumption about reality. Second, when one chooses to define 'cause', one makes a separate *analytical* step based on, but distinct from, the metaphysical assumption, because several definitions could be consistent with one assumption. Separate from the steps of assumption and definition is the problem of research strategy: how best to choose hypotheses and how best to test them. As for a general definition of 'cause', it is practically synonymous with 'adequate explanation' or 'the goal of science' (Nagel, 1961), and as such is beyond the scope of this paper. While we attempt no general definition, we do consider a series of narrower, technical definitions below.

*Order of authorship decided by the flip of a coin.

We begin by stating a widespread presupposition about research strategy: simple hypotheses and simple experiments are to be preferred to complex ones. This principle suggests that we should direct our energies to testing and rejecting simple hypotheses *before* considering complex ones. As scientists, we must face, but often ignore, the question of when to abandon simple explanations for complex ones. This question has three parts. What are the alternatives to testing a complex hypothesis if simple ones have failed? What are the scientific costs of failing to examine complex hypotheses? And when should we switch from simple to complex explanation?

These questions do trouble the research community. In the rest of the introduction, we examine the divergent views of several leading ecologists and evolutionary biologists to document the point that these issues elicit strongly stated positions whose differences suggest that important implicit assumptions have not yet been thoroughly examined.

In *Sociobiology*, Wilson (1975, p. 26) states: 'The single greatest difficulty encountered in the construction of multiple hypotheses is making them competitive instead of compatible... If more than one is true, some method must eventually be devised to assess their importance. The subject thereby gains one magnitude in difficulty'. Wilson suggests that progress in science will be slower if we must resort to complex hypotheses that are not mutually exclusive, and implies that we should avoid them if possible. (We note that two complex hypotheses may also be mutually exclusive.)

Krebs and Myers (1974, p. 271) also display a strong aversion to multi-factor hypotheses:

> The multiple-factor hypothesis is particularly dangerous as a methodological argument. If taken at its face value as a vague armchair theory, the multiple-factor hypothesis is certainly true. The factors which affect a lemming population in Alaska are certainly different from those affecting a vole population in Kansas. But if we adopt this hypothesis as our research strategy, we lose one of the most important checks on scientific speculation – the testability of hypotheses.

Note how the concepts of 'hypothesis' and of 'research strategy' are confused in this quote, and how 'factor' appears to be identified with 'cause'. That such confusion can occasionally afflict even prominent ecologists may account for the origin of the problems we discuss.

These criticisms of multi-factor hypotheses clash with the recognition that many processes have complex causes. As Ashby (1956, p. 5) stated:

> Science stands today on something of a divide. For two centuries it has been exploring systems that are either intrinsically simple or that are capable of being analyzed into simple components. The fact that such a dogma as "vary the factors one at a time" could be accepted for a century, shows that scientists were largely concerned in investigating such systems as *allowed* this method; for this method is often fundamentally impossible in complex systems.

Simon (1962), Wimsatt (1974), Southwood (1980), and Strong (1980) have extended the implications of Ashby's observation. Should we abandon research on important questions because the dominant hypotheses may be

compatible and complex? Should we switch research areas when we can find no critical experiments to distinguish between them? We think not. We must consider the implications of multi-factor hypotheses, and the options we have to rejecting them.

For the purposes of the following discussion, we define 'factor' as an operationally definable element of reality that we believe has an influence on the effect we are analyzing. The object of the analysis is to discover the status we should assign to that influence. Under what circumstances should we promote 'factors' to the status of 'causes'?

If we have three single factor hypotheses – A, B, and C – and we have rejected each as the necessary and sufficient cause of the effect of interest, we may broaden the search and generate a new single factor hypothesis D. We may construct combinations of A, B, or C. We may redefine the meaning of 'cause' to gain new insight, or we may abandon the field and work on a new problem.

Assume that A and C in fact interact to produce our effect. If we already know what we are trying to discover, we will test a complex hypothesis involving the interactions of A, B, and C. If we accept on faith that every effect has a single necessary cause, we will fail to test the multi-factor hypothesis, and for this pay a price.

The price of such a mistake is delay. If we concoct a new hypothesis D, and test it, or if we construct *ad hoc* modifications of A, B, or C, or if we switch fields, we delay the understanding of the process of interest. Later in this paper we discuss the history of research on atherosclerosis and find that adherence to single causes in the initial interpretations of a major public health study (the Framingham study) misled the research community and the public.

We claim that the decision to examine explanations based on multiple causes should be made earlier than is generally the case – not necessarily as the first step in a new field, but perhaps as the first step in the investigation of a new case in an already developed field. For example, Holling (personal communication) argues that studies of population dynamics should now begin with the assumption that populations are regulated by the interaction of predation and food. Evidence from dozens of species compels us to formulate working hypotheses that are multi-factorial at the outset.

Lewontin (1974a, p. 401), after discussing those cases in which alternative causes *can* be discriminated in human genetics, stated:

> The second problem of causation is quite different. It is the problem of the *analysis* into separate elements of a number of causes that are interacting to produce a single result. In particular, it is the problem of analyzing into separate components the interaction between environment and genotype in the determination of phenotype. Here, far from trying to discriminate individuals into two distinct and mutually exclusive etiological groups, we recognize that all individuals owe their phenotype to the biological activity of their genes in a unique sequence of environments and to developmental events that may occur subsequent to, although dependent upon, the initial action of the genes. The

analysis of interacting causes is fundamentally a different concept from the discrimination of alternative causes. The difficulties in the early history of genetics embodied in the pseudo-question of 'nature versus nurture' arose precisely because of the confusion between those two problems in causation. It was supposed that the phenotype of an individual could be the result of *either* environment *or* genotype, whereas we understand the phenotype to be the result of *both*.

These are fundamental issues of research strategy that go deeper than arguments over the definition of terms. Our experience in ecology and evolutionary biology suggests that progress has been impeded in several important areas by adhering to the single-cause paradigm, and by avoiding hypotheses that are not easily separable. We also suspect this is true in many areas of the natural and social sciences.

This paper considers the pathologies that result from avoiding multi-factor hypotheses. We first define the nature of multi-factor hypotheses by using the concepts of necessity and sufficiency. We then use seven brief examples to examine the problems that arise when one tries to analyze multiple-caused processes with methods developed for distinguishing between mutually exclusive single causes.

In the first, latitudinal clines in clutch size in birds, we present an hypothetical example to show how single-factor experiments fail to uncover multiple-factor relationships. In the second, intraspecific competition in population dynamics, we point out that simple hypotheses can hide complex assumptions. In the third, population cycles in small mammals, we suggest that understanding has been delayed by the decision to continue to use single-factor experiments that were not giving clear results. In the fourth, the evolution of age at maturity, we note that complex hypotheses can be stated and tested as rigorously as can simple ones.

In the fifth, atherosclerosis in humans, we show how competing single-factor schools have coalesced to form one multi-factor hypothesis. We chose this example to demonstrate that multi-factor approaches can be necessary at levels of biological organization below that of the organism. In the sixth, the dynamics of chromosome inversions in a grasshopper, we point out that one weakness of the single-cause viewpoint stems from insufficient dimensionality. In the seventh, reaction yields in chemistry, we refer to experimental methods designed specifically to deal efficiently with multiple causes. These methods need not be 'an order of magnitude' more difficult, as Wilson suggested, but are relatively straightforward.

2. Some models of causation

2.1 Single level causation

If we wish to explain only the presence or absence of some effect Z, and have three possible causes A, B, and C, all of which are located on the same level of biological organization, there are three ways these can combine to produce Z:

I. One factor may be necessary and sufficient.
II. Several factors may be necessary, and none sufficient.
III. Several factors may be sufficient, and none necessary.

If we account explicitly for all combinatorial interactions, there are many possibilities. The three factors may be grouped as follows: A; B; C; A or B; A or C; B or C; A and B; A and C; B and C; (A or B) and C; (A or C) and B; (B or C) and A; (A and B) or C; (A and C) or B; (B and C) or A; (A or B or C); (A and B and C). Any one of these groups of factors could be: (1) necessary and sufficient; (2) necessary but not sufficient; (3) not necessary but sufficient; (4) neither necessary nor sufficient.

Clearly, only mechanism I, which corresponds to combinatorial outcome (1) for the first three groupings, is single factor causality; II and III are multi-factor. If we do the simplest experiment first, we would use four treatments: (a) only factor A, (b) only factor B, (c) only factor C, (d) neither A, B, nor C. These experiments would test all hypotheses under I. To isolate one of the 68 cases (17 groups x 4 conditions), we only need four more treatments: (e) A and B, (f) A and C, (g) B and C, and (h) A, B, and C.

This simple example shows that embracing multi-factor hypotheses does not necessarily imply an order of magnitude increase in difficulty. For this case, only doubling the number of treatments is necessary to distinguish among 68 possibilities, rather than 4.

2.2 Hierarchical causation

Biologists have long recognized that the natural world has hierarchical organization. This perception has been formalized as a set of academic specialties that correspond to levels in the hierarchy. Continuing research brings changes in the boundaries of academic specialties, but some boundaries are natural, not man's inventions. Levels of organization, such as cells, organs, organisms, populations, must be distinguished from levels of analysis, such as genetic, biochemical, physiological, ecological, or evolutionary. Levels of organization represent natural objects; the choice of level of analysis represents a research strategy.

Given this hierarchical view of nature and of scientific explanation, then every time we achieve an explanation of an effect at one level in terms of effects at other levels there is a possibility of a particular kind of multi-factor causation that has been analyzed by Mackie (1965). Suppose that A, B, and C, which we conceive as being at, say, the higher level, each are effects that appear only when there are certain antecedent conditions at the lower level. Let us indicate these conditions for A as HIJ, for B as MNO, and for C as TUV. This notation is intended to communicate the notion that H, I, and J are each necessary but insufficient antecedent conditions for A, that it takes the combination of all three to produce A.

Now consider a level above that upon which A, B, and C are found where we are studying an effect, Z, that is influenced in some fashion by A, B, and

C. In particular, let us suppose that Z appears either when A is present, or when B is present, or when C is present, or when they are present in any combination: that is A, B, and C are each sufficient but not necessary for Z. How are we to describe the relationship among several levels? That is, what should we call the clear but complex relation between, say, H, which is a necessary but insufficient antecedent condition for A, to Z? Mackie (1965) has described H (or N, or V) as *an insufficient but non-redundant part of an unnecessary but sufficient condition* (such as A). For brevity, this is termed an *inus* condition. Schaffner (personal communication) notes that '... what is usually termed a cause is an *inus* condition. But our knowledge of causal irregularities is seldom *fully* and *completely characterized: we know some of the inus* conditions but rarely all possible ones.'

At first encounter, the *inus* condition may appear to confuse more than it clarifies, but we suggest that this is a misleading first impression. It has the distinct advantage of making explicit the idea of several sufficient factors on one level (A, B, or C), none of which is necessary to produce an effect on a higher level, Z. It implies that experiments on two levels are required, first to establish the sufficiency of A, B, or C to produce Z, then to establish the necessity of H, I, and J to produce A by their interaction. While the *inus* condition is certainly not the only possible logical relationship of factors among levels, it does indicate that complex causal networks can be stated clearly and explicitly and approached, at least in principle, experimentally. Notions of causation viewed in this way also contribute to the unification of science by making the establishment of connections among levels an implicit research program.

The introduction of hierarchies suggests that there are another three ways in which phenomena may be multiply caused: through the interaction of several factors on the same level, through the interaction of several factors on antecedent levels, or through a long chain of single causes proceeding through many levels. No biologist contemplates, except as an abstract exercise, the tracing of causation back to fundamental physical particles. These examples suggest that in any partial tracing there should be numerous opportunities for encountering multi-factor causation on or between levels.

3. Clutch size: single-factor experiments can mislead

When a measurable process is caused by two or more factors, there is a real danger of rejecting each factor in turn, and concluding that none is a cause. In this section we illustrate the underlying problem using analysis of variance to interpret hypothetical data bearing on a real problem, latitudinal clines in avian clutch size.

Most birds that breed in the Northern Hemisphere over a wide range of latitudes have larger clutches further north. Several explanations for these clines have been advanced, among them (1) higher food densities in northern

region during the spring and summer, (2) higher predation in southern areas favoring more, smaller clutches, and (3) longer daylength in northern regions allowing more foraging time. Most students of clutch size would agree that these hypotheses are not mutually exclusive and could interact to produce the clines.

Suppose that a scientist tries to test this multiple causal hypothesis. He or she travels around the world gathering data on clutch size, levels of predation, levels of food, and daylength, then uses analysis of variance to present the results (Table 1). The conclusion drawn is that all three factors

Table 1. 3 way factorial ANOVA

Source[a]	df	SS	MS	F[b]
Predation	1	8	8	8
Food	1	8	8	8
Daylength	1	8	8	8
Residual	4	4	1	
Total	7	28		

[a]Note: interaction terms omitted for clarity
[b]$p < 0.05$

are important in determining clutch size. These results, together with data showing that higher latitudes have more food, lower predation, and longer daylength, form a plausible explanation.

Now imagine a second scientist also interested in latitudinal clines in clutch size, but one who believes that one should begin with the simplest possible hypothesis. That hypothesis might be that longer daylength in northern latitudes is responsible for the clines. This scientist collects only data on clutch size and daylength, visits the same locations, and uses the same statistical methods, but gets different results (Table 2). The total

Table 2

Source	df	SS	MS	F
Daylength	1	8	8	2.4 NS
Residual	6	20	3.33	
Total	7	28		

variation observed is the same in both cases, but because food and predation levels were not measured in the second case, the variability due to these sources is assigned to the residual term, and the test for significance of daylength effect is not significant. The second scientist concludes that daylength is not an important determinant of clutch size.

Another scientist might test the simple hypothesis that food determines the latitudinal clines, and because the previous study shows that daylength is not important, does not measure daylength. This scientist would find no

significant effect of food. Another scientist might do the same for predation. Each individual test of one factor rejects it, because the variability due to other causes, when not removed by control or measurement, masks the effect of the single factor. This is not a criticism of the single-factor approach *per se*, but a comment on poor experimental procedure, and a warning against hasty rejection of alternate hypotheses where proper controls simply are not possible.

The error made by the second scientist is to assume that one need not control for factors demonstrated not to be significant. Failure to find significance is weak evidence that a factor is not important for two reasons: its significance may be masked as outlined above, it may interact with another factor to produce the process of interest, or both.

This example demonstrates the kind of error that can arise at a simple level from the unreflective application of analysis of variance. Lewontin (1974a) points out that the capacity of analysis of variance to discriminate *causes* is severely limited by the localization of the data used in space and time. Only longer-term studies of the dynamics of a process in which antecedent conditions are clearly controlled can result in more certain statements about causes. This criticism does apply to the hypothetical case we just developed, but it does not invalidate our point.

4. Competition: simple hypotheses hide complex assumptions

In proposing a simple hypothesis, we may have to make a long series of complex assumptions. In proposing a complex hypothesis, we may have to make a few, simple assumptions. We owe this insight and the following example to P.J. den Boer (personal communication).

Consider the hypothesis 'population density is regulated around its equilibrium value by intraspecific competition'. This is a simple hypothesis with respect to causes, but a very complex hypothesis if we consider its implicit assumptions: first, that some essential resource is in continuously short supply; second, that individuals are sufficiently similar in their interactions to allow them to be simply summed; third, that there is some 'equilibrium density' that all individuals are roughly equally able to tolerate and thus influence each other; fourth, that the population can be considered closed; fifth, that the habitat can be considered homogeneous; sixth, that abiotic factors can be considered not to change enough to alter the competitive relationships; seventh, that evolutionary change is not important over the time scale required to attain 'equilibrium density'; and so forth.

Thus hypotheses can be made simple either in their assumptions or in their statements about causes, but probably not in both ways at the same time. Rather than set out to test a single-factor hypothesis that hides very complex implicit assumptions, we should prefer other hypotheses with simpler assumptions, some of which may very well be multiple-factor

hypotheses whose overall complexity, taking account both of assumptions and of causation, is in fact less.

5. Population cycles in small mammals: has understanding been delayed by single-factor experiments?

A field that has been dominated by single-cause thinking is the study of population regulation in small mammals. Many small mammals undergo periodic cycles of abundance, and *the* cause of these cycles has been intensively sought for over fifty years. The research was initiated by Elton and still continues.

Chitty (1960) summarized the history of studies on lemmings and voles (microtines). The initial hypothesis of Elton and Chitty was that epidemic disease decimated vole populations when they reached high numbers. Subsequent work showed that disease was neither necessary nor sufficient to cause vole declines. Data on predation, food shortage, poor weather and adreno-pituitary exhaustion as causes of vole declines suggested that none of these factors were necessary. Most importantly, the changes in survival of voles and lemmings during declines are often sex specific, and there are other features of the populations, notably high body weights prior to the decline, that are difficult to explain with the extrinsic mechanisms mentioned above. Chitty wrote (1960, p. 107), 'We therefore had to consider a third type of explanation involving two or more factors in combination including at least one necessary and sufficient condition.' Chitty (1967) proposed that rapid genetic selection for differing forms of social behavior was a necessary condition for the cycles, and this hypothesis has guided the work of one research group for the last 20 years.

Chitty (1960, 1967) has advocated the rigorous application of the traditional scientific method as set out by Cohen and Nagel (1934) and Popper (1959): he always tries to falsify his hypotheses. The difficulty of controlling food supply and predation pressure in natural populations has resulted in a pragmatic but unfortunate decision: once a factor has been rejected as not a necessary cause of the declines, this factor is neither controlled nor measured in subsequent experiments. Others (Lidicker, 1978; Pearson, 1966; Pitelka, 1973; Keith, 1974) have argued that small mammal cycles are caused by an interaction of food, predation, and social behavior. Chitty and Krebs object to such hypotheses because they do not think they can be tested, a view exemplified in the earlier quote from Krebs and Myers (1974: *cf.* Introduction).

The essence of this approach is the search for *necessary* causes. It recognizes that disease, food, predation, and weather undoubtedly influence small mammal cycles, but rejects them as necessary causal agents. This has led to a progressive restriction in the definition of the process of interest. Whereas Elton originally hoped to explain the cycles, Chitty now hopes only

to explain the cause of poor survival during the sharp declines that occur for one to three months every two or four years. He sees this as the key to the problem, the unitary, 'properly analyzed' process to be explained by a single necessary cause.

In a recent review of the behavior-selection hypothesis of population regulation, Krebs (1978a) outlined a series of experiments that could be performed to test it. Krebs implies that if these experiments gave results contrary to the behavior-selection hypothesis, we should abandon it and seek another. We see a major danger in such a move. Once a factor has been eliminated as a necessary cause, it is also eliminated from subsequent experiments designed with other single causes in mind, and is neither controlled nor monitored. Food supply and predation are both known to affect survival of voles, and without controls for these factors, experiments to test the effect of social behavior on declines in numbers may prove inconclusive because variation induced by other factors will mask the effects of social behavior and genetics. Although these declines may be largely due to social behavior, the social behavior hypothesis could be rejected using the same criteria that have been used to reject the disease, food, predation, and weather hypotheses.

We contend that already they may have rejected many of the factors that cause cycles. Moreover, 'social behavior may change in the course of a cycle, and may be critical to understanding the cyclic change, but it is merely mediating an effect of environment on population performance... Even if [someone] showed that there was a consistent [cycle] in social behavior... that would not be an explanation of cycles. It would not even be an adequate explanation of poor survival in a sharp decline... What are the selective factors triggering the observed change in social behavior?' (Pitelka, personal communication). Whereas Chitty and Krebs imply that when they eventually discover the cause of the cycles it will be a single cause operating on a single level of biological organization, Pitelka implies that the cause will be complex and hierarchical, extending over several levels of biological organization.

Recent studies by Krebs and his students have found both predation and food supply to have major effects on population numbers in *Microtus townsendii*. Krebs (1978a) concluded '*Microtus townsendii* populations in southwestern British Columbia may be limited by both heavy carnivore predation (coupled with botfly parasitism) and by spacing behavior.' Taitt (1978) demonstrated that changes in food supply, moderated by spacing behavior, lead to changes in population density. Neither of these results is contrary to Chitty's contentions, but what is new is the recognition that while none of these factors may be necessary for the declines of interest, the study of food, predation, and spacing behavior will probably be required as controlled background to understand the declines.

We do not yet understand small mammal population dynamics. It now appears that Chitty was correct in pointing to intrinsic mechanisms, in

this case spacing behavior, as one major and perhaps necessary factor. However, the inability to establish this point definitively has undoubtedly been delayed by inadequate control of food and predation. As Andersson and Hansson (1974, p. 126) noted, 'strong inference seems to be most useful when there is a choice between mutually exclusive alternatives, whereas a main problem in population regulation is to obtain a quantitative estimate of the influence of different processes, in which case a simple refutation approach is not sufficient.'

6. Age at maturity: complex hypotheses can be rigorous

The field of life-history evolution is dominated by complex causation. '[F] or any given trend in life-history traits, e.g., an increase in clutch size, there are several plausible hypotheses, not mutually exclusive, that could explain the trend, either singly or in combination' (Stearns, 1976, p. 37). For example, there are several hypotheses in the literature on the circumstances under which natural selection favors the evolution of delayed maturity. Stearns and Crandall (1981) selected two that operate at the same level of biological organization (changes in age-specific mortality and fecundity): (a) organisms that delay maturity produce offspring with lower juvenile mortality, and (b) organisms that delay maturity gain fecundity.

They built models based on the assumption of either the first hypothesis, the second hypothesis, or both, for the two hypotheses are not mutually exclusive. In doing this, they assumed a stable age distribution, an exponentially growing population, population growth rate as the definition of fitness, and optimization as an appropriate procedure. They then estimated some parameters in the models from field data reported in the literature on lizards and salamanders, and used their models to predict the ages at which those nine populations should mature. For the quality-of-young hypothesis (a), the equations converged to solutions in all nine cases, and the correlation of predicted with observed ages at maturity was 0.93. For the fecundity-gain hypothesis (b), the equations converged to solutions in six cases, and the correlation of predicted with observed ages at maturity was 0.90. For the combined hypothesis, the equations converged to solutions in all cases, and the correlation of predicted with observed ages at maturity was 0.96.

This example demonstrates that simple hypotheses can be combined to form complex ones that are just as subject to quantitative tests as are simple ones. In this case, because correlations of 0.90, 0.93, and 0.96 are all high and not significantly different, weight must be given to the number of cases explained. The quality-of-young hypothesis explains nine cases, the fecundity-gain hypothesis explains seven cases, and their combination explains nine cases. Thus either could be sufficient, but neither necessary, in seven cases, and only one is sufficient in the remaining two (its necessity is not established by these results).

7. Atherosclerosis: a case study in complex causation

We know of no ecological problems that have been so completely worked out that we could use their history to demonstrate to the satisfaction of a determined critic that single factor hypotheses have delayed progress. This is in part a comment on the state of ecology, but it is also a comment on the ease with which a single-causationist can deflect the force of an argument by invoking the idea that any phenomenon that appears to have multiple causes has not been 'properly analyzed'. We therefore use the history of scientific explanations of atherosclerosis to illustrate a story that is sufficiently complete to make our point.

What would we not know today if atherosclerosis had been viewed as having a single cause? We first present a brief summary of current understanding of atherosclerosis drawn mostly from Ross and Glomset (1976), Gresham (1976), and McGill (1977). Atherosclerosis is a disease of the arteries characterized by the formation of nodules or plaques involving the arterial wall (Critchley, 1978). Several prominent workers now suggest that atherosclerosis has multiple causes that are not necessarily mutually exclusive. For example, 'it should be apparent that the three hypotheses [on the proliferation of smooth muscle cells in plaques] are not necessarily mutually exclusive; in fact, in some interesting ways they are complementary' (Ross and Glomset, 1976, p. 425; see also McGill, 1977, p. 56, and Beaumont and Beaumont, 1978, p. 144).

An atherosclerotic plaque is formed by the proliferation of smooth-muscle cells, the deposition of lipid, and the accumulation of collagen, elastic fibers, and proteoglycans. On one level, epidemiological risk factors are correlated with incidence of atherosclerosis: smoking (Astrup and Kjeldsen, 1974), plasma lipids, the most prominent being cholesterol (increased incidence) and high-density lipoprotein (decreased incidence: Gotto, 1979), sex (males have higher risk: McGill and Stern, 1979), age and hypertension (e.g., Kannel and Gordon, 1971), diabetes (Gries *et al.*, 1979), auto-immune responses (Beaumont and Beaumont, 1978), stress, race, heritable factors, and a lack of exercise (McGill, 1977). While risk factors are in a sense causes, similar to *inus* conditions, we concentrate on the mechanisms that produce the plaque, for plaque production is precisely the sort of narrowly-defined process that one might expect to have a single proximate cause no matter how many risk factors affect it.

In fact, at this second level we still see multiple causation. The hypotheses on plaque formation include two on smooth-muscle cell proliferation – the Response-to-Injury Hypothesis (Duguid, 1949; French, 1966; Mustard and Packingham, 1975; Ross and Glomset, 1976) and the Monoclonal Hypothesis (Benditt and Benditt, 1973; Benditt, 1977) – and several hypotheses on how changes in the arterial wall allow an influx of blood chemicals leading to lipid and collagen deposition (Astrup and Kjeldsen, 1974).

This example suggests two points. First, are we dealing with a single phenomenon that has multiple causes, or several phenomena each of which has a single cause? Cohen and Nagel (1934) comment: 'When a plurality of causes is asserted for an effect, the effect is not analyzed very carefully. Instances which have significant differences are taken to illustrate the same effect. These differences escape the untrained eye, although they are noticed by the expert.' We disagree. Their comment does not apply to all important and interesting cases, as had been clearly argued by Stebbing (1931) before their book was published.

Atherosclerotic plaques are recognized by workers in the field as objects identifiable by any properly trained pathologist. In that sense the class of atherosclerotic plaques forms a unitary phenomenon. If we have to subdivide such classes into single objects to obtain the unitary phenomena sought by Cohen and Nagel, we lose generality.

At least two processes are involved in the dynamics of just one plaque component, the proliferation of smooth-muscle cells – injury to the arterial wall allows entry of blood constituents, then a lesion forms as a benign tumor derived from smooth-muscle cells transformed by agents admitted from the blood (Ross and Glomset, 1976). Thus the class of atherosclerotic plaques is a single phenomenon, but both their origin and their growth probably have multiple causes.

Second, what would we know today if atherosclerosis had been viewed as having a single necessary and sufficient cause? Each risk factor now emphasized in disease counseling, with the possible exception of smoking, would have been systematically rejected because there are human populations exposed to them in which the incidence of atherosclerosis is low. The interaction of two or more factors is detected only slowly and inefficiently by researchers searching for single causes. The factors influencing rate of cholesterol uptake into plaques, such as high-density lipoproteins, would have been ruled out. It seems likely that some version of the Response-to-Injury hypothesis would have survived repeated attempts at rejection, but that could have resulted in the pessimistic view (*cf.* Texon, 1974) that hemodynamic stress is pervasive and that atherosclerosis is untreatable.

Thus this is a case in which systematic testing of a single hypothesis, coupled with a strong belief in single causes, could have seriously misled the investigator. Schaffner's (1980) penetrating analysis of the immune response suggests a similar conclusion.

8. The gains made in switching to multiple causation

Small mammal cycles are probably caused by the interaction of several necessary factors no one of which is sufficient. It is important to see how perceptions of the causation of such a process changed as a field shifted from assuming single causes to assuming multiple causes. Two well-studied

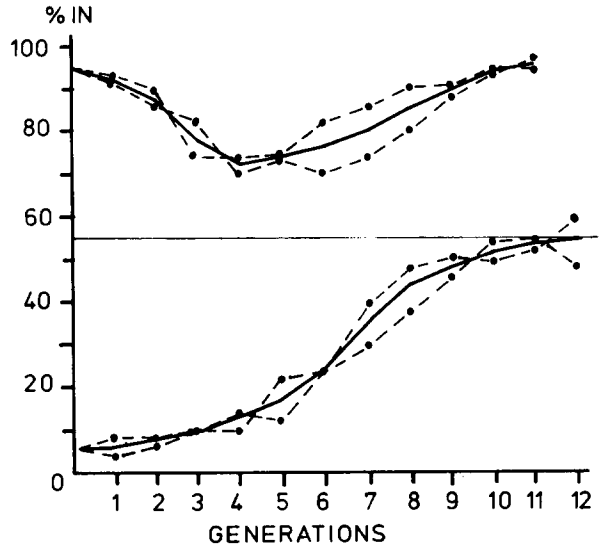

Figure 1. The frequency of a chromosomal inversion, BL, in laboratory populations of the Australian grasshopper, *Moraba scurra*. The points represent individual data; the heavy lines represent average behavior. (From Lewontin, 1974b, p. 279.)

cases are at hand, both from reductionist disciplines in which the understanding of mechanisms has achieved a level of detail not yet available in population dynamics. It is this level of detail that permits statements about changes in the perception of causation through *gedanken* – experiments in which one analyzes the process first from the single-cause viewpoint and discovers contradictions that are eliminated with the introduction of a second cause. The first cause is from population genetics: inversion polymorphisms in the grasshopper *Moraba scurra* (Lewontin, 1974b). The second case is from chemistry: response surfaces for reaction yields (Box, Hunter, and Hunter, 1978).

8.1. Inversion polymorphisms: one or two loci?

Suppose we bring into the laboratory many fertilized female grasshoppers and set up replicated populations with different initial frequencies of the inversion BL on the CD chromosome element. The frequency of BL in nature is about 0.65. Populations started at a low frequency of BL move upward along an S-shaped trajectory to an intermediate equilibrium near 0.65, but populations started at high frequencies start downward and then move back up, apparently toward fixation (Figure 1). Frequency dependent selection cannot account for the results because the populations have passed through the same gene frequencies going in opposite directions.

The dimension missing from the description is the frequency of another inversion, TD, on a different chromosome element, TF. The two inversions

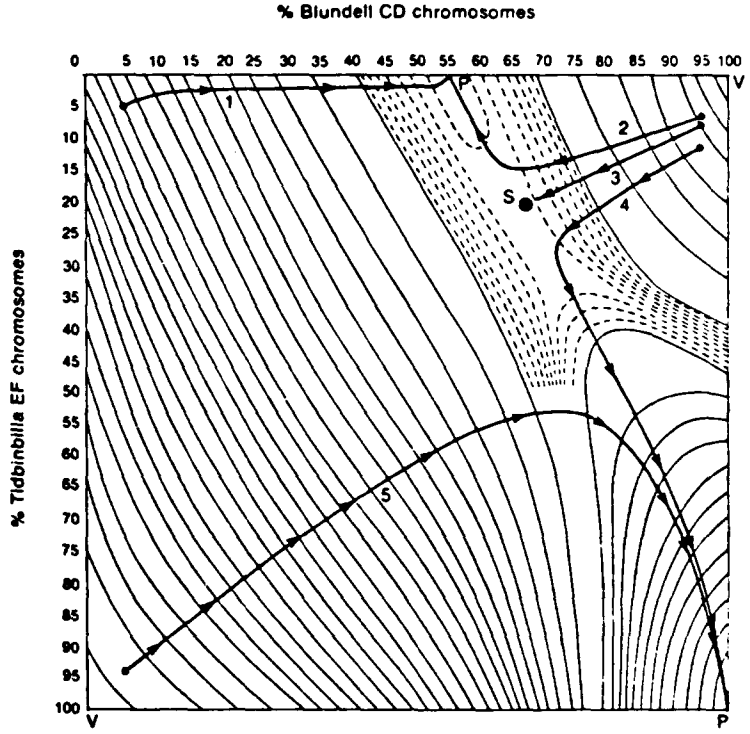

Figure 2. A response surface on which are projected changes in the frequency of two polymorphic inversion systems in *Moraba scurra* from different initial compositions. The arrow-marked trajectories represent solutions to differential equations of gene frequency change, incorporating fitness estimates from nature. (From Lewontin, 1974b, p. 280.)

interact in determining fitness, which is best represented as a response surface (Figure 2). In this representation, the x- and y-axes represent the proportion of each chromosome in the population, and we are asked to imagine a surface emerging from the plane of the page like a hill, whose contours are depicted by both solid lines (large intervals) and dotted lines (small intervals). This surface represents fitness; thus the trajectory of chromosome frequencies should describe a path that always tends to climb the hill. A population started along trajectory 1 or 2 will stabilize at a point where it contains 55% Blundell CD chromosomes and no Tidbinbilla EF chromosomes; a population started along trajectory 4 or 5 will stabilize at a point where it contains 100% of both chromosome types. Trajectory 3 converges to a saddle point at intermediate frequencies of both chromosomes that is only stable if not perturbed. Seeing the inversion dynamics as a two-locus problem removes the contradiction imposed by the single-locus view.

The indeterminacy and the self-contradictory paths of inversion-frequency change in these populations are reflections of the dimensional insufficiency of a single-locus treatment. Even though we may be interested in following only one segregating entity,

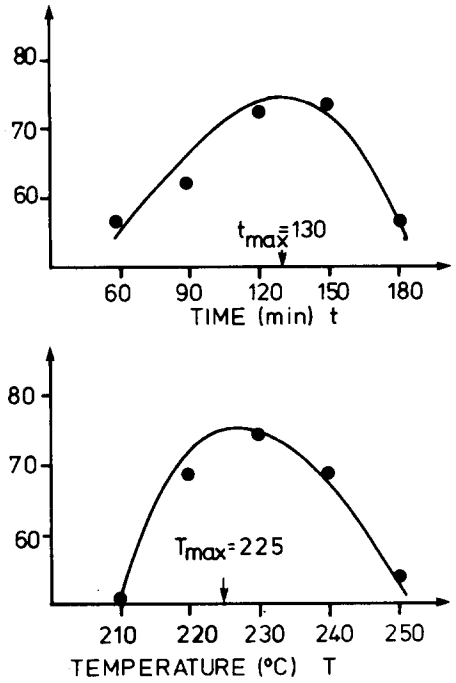

Figure 3. Hypothetical results from the one-variable-at-a-time approach to determining optimum yield from a chemical reaction. (From Box *et al.*, 1978, p. 511.)

say a third chromosome inversion in *D. persimilis*, an understanding of evolution along that one dimension requires *first* a synthetic treatment of the genotype and *then* an abstraction of the single system of interest from the complex mass. We cannot reverse the process, in general, building a theory of a complex system by the addition or aggregation of simple ones (Lewontin, 1974b, p. 281).

8.2. Reaction yields: one or two dimensions?

Suppose our problem is to maximize the yield of a chemical reaction by varying reaction time and temperature. A chemist employing the classical one-variable-at-a-time approach would first determine yield for different reaction times at a fixed temperature (Figure 3), conclude that the best reaction time is, say, 130 minutes, then vary the temperature in a second experiment while holding reaction time fixed. This chemist would conclude that overall yield is maximized at 75 grams with $t = 130$ minutes and $T = 225°C$.

Although Figure 3 shows that yield will decrease if either reaction time or temperature is changed individually, what it does not reveal is that yield will increase if both variables are changed at the same time (Figure 4).

To understand the possible nature of the joint effect of time and temperature on yield we must think in terms of joint functional dependence of mean yield on time and temperature... the one-variable-at-a-time strategy fails in this example because it tacitly assumes that the maximizing value of one variable is independent of the level of the other. Usually this is not true (Box *et al.*, 1978, pp. 512–513).

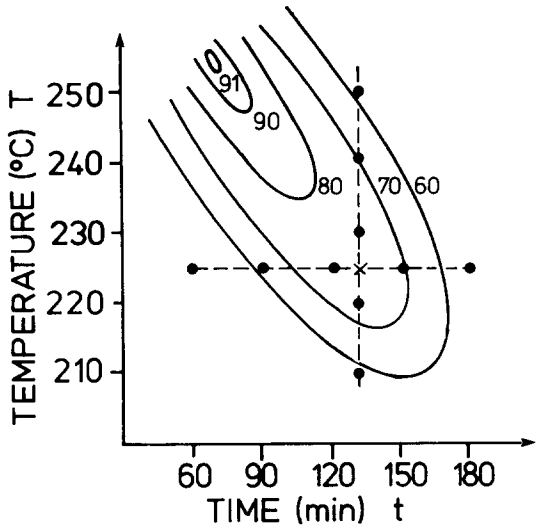

Figure 4. Possible true response surface representing yield versus reaction time and temperature, with the points drawn from the one-variable-at-a-time approach of Figure 3. (From Box *et al.*, 1978, p. 512.) The vertical dashed line represents the projection of the results of Figure 3a onto the plane of the temperature–time surface; the horizontal dashed line represents the similar projection of the results of Figure 3b.

Box *et al.*, go on to discuss the principles and techniques underlying experiments designed to get at precisely these kinds of interaction effects. One of their examples, a pair of second-degree equations representing a mini-max saddle point (*op. cit.* p. 528), is the response surface of Lewontin's two-locus inversion system (Figure 2).

These examples suggest several general conclusions. First, problems of multiple causation extend into disciplines traditionally viewed as precise, mature, and reductionist. They are not diseases peculiar to young, poorly developed fields awaiting the cure of proper methodology. Second, the methods for dealing with such problems are well developed and accessible (*cf.* Box *et al.*, 1978). Third, the shift from a single-cause to a multiple-cause viewpoint brings with it an increase in the dimensionality of the models required but extracts no cost in precision or understanding. It does imply more effort per experiment, but fewer total experiments to achieve a given level of insight.

9. Discussion

The essence of the problem is that the search for a single necessary cause frequently fails. We suggest three mechanisms that produce this failure.

First, one cause may be necessary and sufficient, in that it can account for the presence, but not the entire amplitude, of the effect. Thus it produces only a small quantitative effect which is augmented by each of a number of other factors which may be present or absent, and if present, present in varying intensities. These secondary factors are neither necessary nor sufficient. In trying to dissect the causal structure of such a situation, a single-causationist would reject all factors, including the necessary cause whose effect is masked by unsuspected and uncontrolled factors. Second, each of several causes may be sufficient, but none necessary. Each cause in turn may be rejected by a single-causationist if the effects of each individual cause are small and the controls are not precise. Third, several causes may be necessary, but no single cause sufficient, so that the process appears only as an interaction effect. Again, the single-causationist would systematically reject each causal factor.

At least three methods can be used to investigate multiple-caused processes. One can execute a factorial experiment that isolates the quantitative contribution of each of several causes and measures their interactions (Cochran and Cox, 1957). One can model the process, assuming multiple causes, and predict the quantitative dynamics of the system, then refine the model by testing each component, then the whole system, cyclically (Gilbert *et al.*, 1976). One can also use the response-surface methods of Box *et al.* (1978). Dynamic modelling does not solve the problem, but is does provide an effective means of expressing complex hypotheses clearly. Fractional factorial analysis can considerably reduce the effort involved while retaining an efficient and powerful analysis of interaction effects in multifactorial designs. Porter and Busch (1978) have used it to analyze growth and weaning success in deer mice.

All these approaches are used in biology, but they are not widely used in ecology and evolutionary biology. One reason is that they require very careful quantitative examination of the competing hypotheses and careful experimental designs. The reaction to non-experimental population biology has been an emphasis on experimentation. The next step, we hope, will be an emphasis on more carefully planned experiments in which the alternative of proceeding directly to multifactorial designs is considered. All three methods can be difficult and can require more effort and expense than a single test of a single cause. However, none of the three is as likely to mislead, and we believe that in the long term our knowledge will increase more rapidly, more reliably, and less expensively if researchers adopted one or more of these approaches when multiple causes are suspected.

The sociology of progress in scientific careers may bias us toward the single-cause viewpoint. We learn to emphasize in our papers lists of hypotheses to be tested, for it is harder to publish papers that do not present clean tests of simple hypotheses. Students in ecology and evolutionary biology are told early in their graduate careers to test some hypothesis and to keep

it simple, because they only have a limited time to complete their program. Scientists working on research grants with two, three, or five year periods will work hard to produce experimental designs that will fit those periods, and in the process may explicitly avoid multifactorial desgins if they feel those designs will require an effort that will not produce results rapidly enough to justify the next grant renewal. Thus not only do our assumptions about causes bias our view of nature; so do our social institutions. Of course, the actual causes of the processes we study have no necessary relationship either to our assumptions or to our social institutions. We note that we cannot change the causation of the processes of interest, but we certainly can change our assumptions, our social institutions, and our responses to existing institutions.

Acknowledgements

Comments by Stephen Arch, George Bealer, Scot Carley, Dennis Chitty, William Clark, Piet den Boer, Peter Larkin, Polly Ann McClure, Frank Pitelka, David Reeve, William Schaffner, Carl Walters, William Wiest, William Wimsatt, and Ralf Yorque improved this paper. We are especially grateful to Will Neuhauser for his clarification of our ideas. Stearns was supported by NSF DEB78-22812.

References

Andersson, M. & L. Hansson (1974). Population regulation in small rodents – some hypotheses (original title in Swedish). – Fauna och Flora, *69*, p. 113–126.

Ashby, W.R. (1956). An introduction to cybernetics. – London, Chapman & Hall.

Astrup, P. & K. Kjeldsen (1974). Carbon monoxide, smoking, and atherosclerosis. – Med. Clin. N. Amer. *48*, p. 325–350.

Beaumont, J.L. & U. Beaumont (1978). Immunological aspects of atherosclerosis. – Atherosclerosis Rev. *3*, p. 133–146.

Benditt, E.P. (1978). The monoclonal theory of atherogenesis. – Atherosclerosis Rev. *3*, p. 77–86.

Benditt, E.P. & J.M. Benditt (1973). Evidence for a monoclonal origin of human atherosclerotic plaques. – Proc. nat. Acad. Sci. USA *70*, p. 1753–1756.

Box, G.E.P., W.G. Hunter & J.S. Hunter (1978). Statistics for experimenters. – New York, J. Wiley & Sons.

Chitty, D. (1960). Population processes in the vole and their relevance to general theory. – Can. J. Zool. *38*, p. 99–113.

Chitty, D. (1967). The natural selection of self-regulatory behavior in animal populations. – Proc. ecol. Soc. Aust. *2*, p. 51–78.

Cochran, W.G. & G.M. Cox (1957). Experimental designs. 2nd Ed. – New York, J. Wiley & Sons.

Cohen, M.R. & E. Nagel (1934). An introduction to logic and scientific method. – London, Routledge.

Critchley, M., ed. (1978) Butterworths medical dictionary. – London, Butterworths.

Duguid, J.B. (1949). Pathogenesis of atherosclerosis. – Lancet *2*, p. 925–927.

French, J.E. (1966). Atherosclerosis in relation to the structure and function of the arterial intima, with special reference to the endothelium. – Int. Rev. exp. Pathol. *5*, p. 253–353.

Gilbert, N., A.P. Gutierrez, B.D. Fraser & R.E. Jones (1976). Ecological relationships. – Reading, W.H. Freeman.
Gotto, A.M., Jr. (1979). Status report: plasma lipids, lipoproteins, and coronary heart disease. – Atherosclerosis Rev. *4*, p. 17–28.
Gries, F.A., T. Koschinsky & P. Buchtold (1979). Obesity, diabetes, and hyperlipoproteinemia. – Atherosclerosis Rev. *4*, p. 71–96.
Gresham, G.A. (1976). Atherosclerosis: its causes and potential reversibility. – Triangle *15*, p. 39–43.
Kannel, W.B. & T. Gordon, eds. (1971). The Framingham Study. An epidemiological investigation of cardiovascular disease. Section *27*. – Washington, U.S. Govt. Printing Office.
Keith, L.B. (1974). Some features of population dynamics in mammals. – Proc. int. Congr. Game Biol. *11*, p. 17–58.
Krebs, C.J. (1978a). A review of the Chitty Hypothesis of population regulation. – Can. J. Zool. *56*, p. 2463–2480.
Krebs, C.J. (1978b). Dispersal, spacing behavior and genetics in relation to population fluctuations in the vole *Microtus townsendii*. – Forsch. Zool. *25*, p. 61–77.
Krebs, C.J. & J.H. Myers (1974). Population cycles in small mammals. – Adv. ecol. Res. *8*, p. 267–399.
Lewontin, R.C. (1974a). The analysis of variance and the analysis of causes. – Am. J. hum. Genet. *26*, p. 400–411.
Lewontin, R.C. (1974b). The genetic basis of evolutionary change. – New York, Columbia University Press.
Lidicker, W.Z., Jr. (1978). Regulation of numbers in small mammal populations – historical reflections and a synthesis. – In: D.P. Snyder, ed., Populations of small mammals under natural conditions.
Mackie, J.L. (1965). Causes and conditions. – Amer. phil. Quart. *2*, p. 245–255, p. 261–264.
McGill, H.C., Jr. (1977). Atherosclerosis: problems in pathogenesis. – Atherosclerosis Rev. *2*, p. 27–65.
McGill, H.C., Jr. & M.P. Stern (1979). Sex and atherosclerosis. – Atherosclerosis Rev. *4*, p. 157–242.
Mustard, J.F. & M.A. Packingham (1975). The role of blood and platelets in atherosclerosis. – Thromb. Diath. Haemorrh. *33*, p. 444–456.
Nagel, E. (1961). The structure of science. – New York, Harcourt, Brace, & World, Inc.
Pearson, O.P. (1966). The prey of carnivores during one cycle of mouse abundance. – J. anim. Ecol. *35*, p. 217–233.
Pitelka, F.A. (1973). Cyclic pattern in lemming populations near Barrow, Alaska. – In: M.E. Britton, ed., Alaskan arctic tundra. – AINA Techn. Paper *25*, p. 199–215.
Popper, K.R. (1959). The logic of scientific discovery. – London, Hutchinson.
Ross, R. & J.A. Glomset (1976). The pathogenesis of atherosclerosis. – New Eng. J. Med. *295*, p. 369–377.
Simon, H.A. (1962). The architecture of complexity. – In: H.A. Simon (1969). The sciences of the artifical. – Cambridge, MIT Press.
Southwood, T.R.E. (1980). Ecology – a mixture of pattern and probabilism. – Synthese *43*, p. 111–122.
Stearns, S.C. (1976). Life-history tactics: a review of the ideas. – Quart. Rev. Biol. *51*, p. 3–47.
Stearns, S.C. & R.E. Crandall (1981, in press), Quantitative predictions of delayed maturity. – Evolution *35*.
Stebbing, L.S. (1931). A modern introduction to logic. – London, Methuen & Co.
Strong, D.R., Jr. (1980). Null hypotheses in ecology. – Synthese *43*, p. 271–285.
Taitt, M. (1978). Population dynamics of *Peromyscus maniculatus auterus* and *Microtus townsendii* with supplementary food. – Ph.D. thesis, Dept. of Zoology, Univ. of British Columbia.
Texon, M. (1974). Atherosclerosis: its hemodynamic basis and implications. – Med. Clin. N. Amer. *58*, p. 257–268.
Wilson, E.O. (1975) Sociobiology: The new synthesis. – Cambridge, Belknap Press.
Wimsatt, W.C. (1974). Complexity and organization. – In: K.F. Schaffner & R.S. Cohen, eds., Boston Studies in the Philosophy of Science, *35*, p. 67–86. – Dordrecht (Holland), Reidel.