COMMENTARY

A CRITICAL ASSESSMENT OF LEVINS'S *THE STRATEGY OF MODEL BUILDING IN POPULATION BIOLOGY* (1966)

STEVEN HECHT ORZACK

Department of Ecology and Evolution, The University of Chicago
1101 East 57th Street
Chicago, Illinois 60637 USA

ELLIOTT SOBER

Philosophy Department, University of Wisconsin
Madison, Wisconsin 53706 USA

ABSTRACT

Richard Levins's 1966 article "The strategy of model building in population biology" is an extremely influential analysis of the nature of scientific model building. His claims that model building involves a necessary trade-off among generality, realism and precision, and that truths about nature can be revealed by finding "robust theorems" are important and deserve careful scrutiny. We clarify the concepts of generality, realism and precision and argue that there is no necessary conflict among them. We also examine the idea of robustness and conclude that it lacks proper definition and that its bearing on the question of whether a proposition is true is highly problematic. Accordingly, we believe that neither of Levins's claims should be accepted.

INTRODUCTION

EVERY SCIENTIFIC DISCIPLINE confronts the problem of coping with nature's complexity. If every scientific theory is selective in the details it chooses to characterize and if each introduces simplifying assumptions, it is only reasonable to wonder how theories can ever hope to describe nature as it really is. Richard Levins addresses this fundamental problem in his well-known and influential 1966 article "The strategy of model building in population biology." His solution to this problem consists of two important claims. The first is that model building involves a necessary trade-off among generality, realism and precision. The second important claim involves the concept of *robustness*. Levins asserts that truths about nature can be revealed by finding "robust theorems." He uses this term to refer to a proposition that is a joint consequence of independent models of the same biological phenomenon. Finding such theorems supplies an access to truths about nature that supplements the more familiar procedure of testing theoretical predictions with data.

Levins does not define any of the model

Editors' Note: In submitting their manuscript, Orzack and Sober proposed that Levins be invited to provide a response for publication in the same issue. We found this suggestion entirely agreeable. Levins's response follows immediately after this Commentary by Orzack and Sober.

The Quarterly Review of Biology, December 1993, Vol. 68, No. 4

0033-5770/93/6804-0002$1.00

properties that he discusses, nor does he provide an argument for why they are mutually antagonistic. He does describe models that he claims exemplify each of the three model types that result from sacrificing one of these characteristics in order to obtain the other two. Levins also does not define the concept of model independence on which his concept of robustness relies. Our goal is to clarify the meaning of Levins's claims and to assess their plausibility.

Our belief that such clarification is needed stems from our examination of almost all of the journal citations of his paper listed in *Science Citation Index* and *Social Sciences Citation Index* (as well as many of the citations in books). Our impression from this survey is that Levins's arguments have been widely misinterpreted.

Although Levins's analysis is framed entirely in terms of population biology, his trichotomy and the concept of robustness can be applied to models about any subject matter. For this reason, Levins's ideas and our analyses should interest scientists in a variety of disciplines.

LEVINS'S TAXONOMY OF MODELS

The first strategy of model building ("Type I") is to sacrifice generality for the sake of realism and precision. Levins's description suggests that he has in mind a "purely" numerical model, in which relevant parameters are taken into account, their values estimated from the data, and a prediction deduced. A Type I model is not general, we surmise, because it can describe only those systems that have the parameters and values entered into the model.

The second strategy ("Type II") is to sacrifice realism for the sake of precision and generality. Levins says this is model building of the frictionless-plane variety. He calls this unrealistic, because some parameters known to have an effect are assumed to play no role. The hope behind this model-building strategy is that idealization will not appreciably diminish the accuracy of the model's predictions.

The third strategy ("Type III") is to sacrifice precision for the sake of generality and realism. Here one makes only qualitative assumptions and deduces only qualitative predictions. This is what Levins means by a model being imprecise.

THE DEFINITION OF MODEL ATTRIBUTES

Given our summary of what Levins says, how might the concepts of generality, realism and precision be defined? We suggest the following:

> If one model applies to more real world systems than another, it is *more general.*
>
> If one model takes account of more independent variables known to have an effect than another model, it is *more realistic.*
>
> If a model generates point predictions for output parameters, it is *precise.*

We have defined generality and realism as comparative notions, but precision and imprecision as a dichotomy. We hope that Levins's claims do not require absolute measures of how general or realistic a model is, since we doubt that such concepts can be characterized meaningfully. It also is worth pointing out that assessments of generality and realism are most straightforward when the set of systems described by the first model properly contains the set of systems described by the second. There is little point in comparing models that relate to entirely different phenomena. Models in population ecology do not compete with models in plate tectonics. We will assume that the models being compared are "about" the same phenomena, that is, they have the same *dependent* variable. However, even when there is this overlap between the models, it is not always obvious how to assess their generality. One possibility is to count occurrences of biological phenomena, but even knowing how to do so in particular instances can be difficult.

It is worth noting that Levins treats generality as a possible characteristic of a biological model. We agree with him, although we believe that many biologists, in their hearts, regard this as a pipe dream. A general model applies to many real world systems. This does not require that those systems have the same values for a given set of parameters. The amount of rainfall in one field may differ from the amount in another, but that is no impediment to a general model that applies to both. Skepticism about generality reflects the suspicion that different real world systems have

different *kinds* of processes at work in them. It would mean that even if the two fields were alike in rainfall, sunlight, and so on, the effect on some dependent variable would be different. We won't try to defend the assumption that general models are possible. It is worth realizing, however, that this assumption is regarded by many as central to the whole enterprise of scientific theorizing.

Further clarification of the concept of generality is needed. We have talked about the generality of a model by considering the number of real world situations to which the model applies. But what does it mean for a model to apply to a particular situation? For example, consider the Hardy-Weinberg Law. One way to formulate the law is:

> If no evolutionary forces are present, then the genotypic frequencies are p^2, $2pq$, and q^2.

Here p and q are allelic frequencies at some diallelic locus in the population. This law is not general because every population experiences some evolutionary forces. On the other hand, the law also can be described as saying that:

> If genotypic frequencies depart from p^2, $2pq$, and q^2, then some evolutionary forces are acting.

Construed in this way, the law has considerable generality since genotypic frequencies often depart from Hardy-Weinberg values.

It is unsatisfactory that the generality of the Hardy-Weinberg Law depends on how it is described. One solution would be to regard the first formulation as canonical; for example, one might insist that the point of a model is to generate predictions for given parameters. This would mean that the Hardy-Weinberg Law is understood as predicting genotypic frequencies in the absence of evolutionary forces. Interpreted in this way, the Hardy-Weinberg model has no generality. We do not claim that this is the only way to make the notion of a model's generality unambiguous. At the very least, the conclusion to be drawn is that the concept of generality requires clarification.

MODELS AS MATHEMATICAL AND EMPIRICAL STATEMENTS

We wish to make a pair of distinctions that are important in connection with Levins's claims. Consider any mathematical model

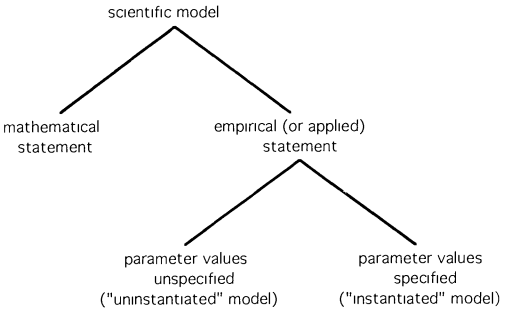

FIG. 1. THE RELATIONSHIP BETWEEN THE MATHEMATICAL STATEMENT OF A SCIENTIFIC MODEL AND TWO TYPES OF EMPIRICAL FORMULATION IT MIGHT RECEIVE

said to describe or explain a particular phenomenon. The first distinction is the one between the model as a mathematical statement and the model as an empirical claim about part of the physical world (Beatty, 1981). Within this second class, we can distinguish models whose parameter values are left unspecified and those whose values are specified. The resulting tree structure of models is represented in Figure 1.

To illustrate these distinctions, consider the standard one-locus, two-allele model of viability selection in a diploid population. The three genotypes *AA*, *Aa*, and *aa* are assigned fitnesses w_{AA}, w_{Aa}, and w_{aa}, respectively. It is a mathematical truth that if selection is the only force acting on the population, then a stable polymorphism will occur when $w_{Aa} > w_{AA}, w_{aa}$. It would be pointless to examine natural populations to see if this statement is true. Merely checking the algebra suffices. This mathematical model is applied when a biologist makes a statement like the following: The one-locus model of heterozygote superiority explains the polymorphism at a particular locus in a given population. The resulting claim is no longer a piece of pure mathematics, but requires empirical testing.

Within the category of applied models, we distinguish two kinds of claims. If the model of heterozygote superiority is uninstantiated, the biologist simply asserts that the locus in question is characterized by heterozygote superiority and notes that this predicts a stable

polymorphism. No measurement of fitnesses or prediction of point values for the frequencies is involved. An inequality among the fitnesses is asserted, and the prediction is made that neither allele will be eliminated. In contrast, the applied model may be instantiated. Here the biologist examines individuals in the population of interest, obtains values for the three fitnesses, derives the genotype frequencies predicted at equilibrium, and compares the observed and predicted frequencies.

This pair of distinctions has several implications in regard to Levins's claims. A given model will be more general when it is uninstantiated than when it is instantiated. This is trivial. There are more loci at which there is heterozygote superiority than there are loci at which $w_{Aa} = 1.0$, $w_{AA} = 0.95$, and $w_{aa} = 0.83$.

The fact that a single model can be applied in an uninstantiated or instantiated form is relevant to the assessment of distinctions between models. Consider Levins's claim that one can choose to construct a general and unrealistic model or an ungeneral and realistic model. Let's assess this claim by considering two familiar models of the instantaneous rate of change of population size N:

$$\frac{dN}{dt} = rN, \tag{1}$$

$$\frac{dN}{dt} = rN + \alpha N^2. \tag{2}$$

Here r is the growth rate of the population (assumed to be constant), and α is a constant. Model (1) is the so-called "density-independent" model, while model (2) is often called the "density-dependent" model. This label, however, is not accurate in the present context. Model (2) is an uninstantiated model; it allows for density independence ($\alpha = 0.0$) and density dependence ($\alpha < 0.0$). Model (1) is a special case of model (2) in that any population described by the uninstantiated model (1) also is described by the uninstantiated model (2). So model (2) is more general than model (1). It is also true that every variable that potentially plays a causal role in model (1) also is a variable in model (2), but not conversely. So model (2) is more realistic than model (1). In this case, the two properties are *necessarily* associated; generality and realism are not model attributes that may be altered independently.

Matters change when the two models are instantiated by assigning values to r and α. Model (1) already assumes that $\alpha = 0.0$, whereas the value of α in model (2) needs to be specified. It is hard to see that there is any difference in generality between saying that $r = 0.001$ and $\alpha = 0.0$ versus saying that $r = 0.001$ and $\alpha = -0.01$. Nor is it obvious that one of the instantiated models is always more realistic than the other. Generality and realism are *not* necessarily associated in this case. Our point is that assessments of the relationship between generality and realism are not fixed, but depend upon the mode of application of the models being compared.

The contrast between instantiated and uninstantiated models is relevant in another respect. Surely it is inappropriate to compare the generality or realism of two models when one model is treated in its instantiated form and the other in its uninstantiated form. If we wish to compare a one-locus model of selection with a two-locus model, it is wrong to assign point values to the parameters of one, but to leave the other uninstantiated. Comparing apples with apples and oranges with oranges requires that the two both be treated consistently as uninstantiated or instantiated.

We showed in the comparison of the uninstantiated models (1) and (2) that generality and realism must go together. This will always be true when one model entails another, but not conversely. In this case, how are generality and realism related to Levins's third model attribute, precision? Levins characterizes precision as a dichotomous variable. Accordingly, we must ask whether *increasing* generality and realism means *switching* from precision to imprecision.

It is easy to see that this need not occur. Take any precise (and uninstantiated) model stated in terms of some number of independent variables. Now add a new independent variable. The model now gains in realism and generality because the model allows for the possibility that the new independent variable influences the dependent variable. However, the new model is still precise, just as the old one is. Both the new and the old models gener-

ate quantitative predictions from specified parameter values.

An alternative way to understand what precision means would be to regard an uninstantiated model as precise to the degree that it generates an *accurate* prediction of the value of the dependent variable. Precision is now a matter of degree, not a dichotomous property. Is there a trade-off among generality, realism and precision? Once again, there need not be. To see why, consider an uninstantiated model that describes some of the processes determining the value of some dependent variable. An example is a population genetics model in which the allele frequency at a locus is a function of evolutionary "forces" such as mutation, migration, natural selection, genetic drift, linkage, and recombination. Compare this model to some other uninstantiated model in which some of these forces are omitted. For example, suppose genetic drift and mutation are omitted because it is thought that they have a small effect on the allele frequency. These omissions make this model *less* general, *less* realistic, and *less* precise (in the sense of predictive accuracy) than the first one. In moving from the first model to the second, there will be no trade-off among generality, realism and precision; all have decreased. Levins's general claim turns out to be mistaken, once the relationship among model attributes is properly understood.

THE NATURE OF TYPE I, TYPE II, AND TYPE III MODELS

Levins describes the fishery models reviewed by Watt (1956) as being Type I models. In Watt's words these models are used for "1) predicting the catch in subsequent years. 2) Predicting how to maximize the catch in subsequent years" (p. 614). The input variables for the various models differ but include, for example, age structure, previous catch size, mortality rate(s), population size, and population growth rate. These are models about *population* dynamics. The same can be said about the two models cited by Levins as being Type II models. For example, Leigh (1965) analyses a model of interacting species in order to understand the relationship between the number of population crashes and explosions and the size of the community's food web. Watt's models and this model do not have the same focus, and each includes unique assumptions. But this is all that separates their uninstantiated versions in an abstract sense. All of these models contain (to use Levins's phrase) "general equations from which precise results may be obtained" (p. 422). It is telling that Levins uses this phrase to *distinguish* Type II models from the others. It is hard to see why it does not apply to the uninstantiated versions of *all* mathematical models.

Levins says that the models he classifies as Type I have little generality, though they are realistic and precise. This claim makes the apples and oranges mistake we described in the previous section. Type I models are *not* less general than other models, when all are viewed as uninstantiated. Similarly, if a Type I model is instantiated and then compared with some instantiated version of a model from Types II or III, again it is unclear that the Type I model will be less general. What makes Type I models look less general is that one compares instantiated versions of them with uninstantiated versions of models from other categories. Nor is it correct to say that Type I models are a distinct kind of model because their uninstantiated versions consist merely of the claim that a particular system is describable by some unspecified set of equations. The uninstantiated versions of what Levins calls Type I models (e.g., those in Watt, 1956) contain *biology* just as do those he calls Type II and Type III models.

We have just expressed doubt about Levins's contrast between Type I and Type II models. The same skepticism is in order, we believe, with respect to the contrast between Type II and Type III models. The reason is that most models in population and evolutionary biology make both qualitative and quantitative predictions. When uninstantiated, a standard model of viability selection at a single diallelic locus is Type III-like because it makes qualitative predictions (e.g., the fitness ordering $w_{Aa} > w_{AA}$, w_{aa} results in a stable polymorphism). When instantiated, this model is of Type II because it makes quantitative predictions (e.g., it predicts the equilibrium frequencies of the genotypes given particular fitness values). The fact that the *same* model can exhibit Type II or Type III characteristics, depending upon the mode of application, shows that Levins's trichotomy is not a division of types of models at all.

Type II and Type III would distinguish different kinds of models if Type III models entailed no quantitative predictions at all. Levins offers his own work on evolution in variable environments as an example of Type III modeling. It is important to notice that in such models, precise values for the parameters *can* be specified and precise predictions *can* be generated. What is interesting about Levins's models is not that they are purely qualitative (they are not), but that qualitative inputs entail qualitative outputs. In this respect, however, they are identical with what Levins calls Type II models.

We do not claim that Type III models do not exist in biology. For example, it is reasonable to regard some models of the mechanism of recombination as being of Type III; they are qualitative. Such hypotheses merely assert the *existence* of a process or relationship (e.g., whether single-strand or double-strand nicks in DNA duplexes initiate recombination; see Szostak et al., 1983, for further details). In this sense, these models lack a quantitative dimension. Some models in population and evolutionary biology are of this kind. Of well-known models, the allopatric and sympatric models of speciation come to mind as valid examples. Each makes only a qualitative prediction in that reproductive isolation is said to arise with or without geographic isolation. In this sense, these models are not mathematical. Accordingly, it is not as though qualitative models of population biological phenomena cannot exist. Yet the statement "reproductive isolation arises allopatrically" differs from the statement "heterozygote superiority produces a stable polymorphism" in that the latter stems from a mathematical model that allows one to make quantitative predictions. At present, the same is not true for statements about allopatric speciation. Qualitative models must say nothing of a quantitative nature. So we agree with Levins that Type III models exist in population and evolutionary biology. Nonetheless, Levins's trichotomy is unable to distinguish among *mathematical* models, that is, among almost *all* of the models that he and others have identified in the literature as being of Types I, II and III.

THE CONCEPT OF ROBUSTNESS

Although Levins's claim about the trade-off among generality, realism and precision is specifically formulated as a claim about mathematical models, his claim about robustness applies to mathematical and nonmathematical models alike. The concept of robustness is introduced to address this important question: If every model contains false assumptions, how can we ever hope to discover what is true? Levins's answer is that we should look for "robust theorems." These are propositions that are the joint consequences of many different models. "Our truth," Levins writes (p. 423), "is the intersection of independent lies." When he further writes (p. 427) that a particular "non-robust" theorem "cannot be asserted as a biological fact" it becomes clear that Levins means that a statement's *robustness*, as distinct from its *observational confirmation*, can be evidence for its truth.

There is a special case in which the connection between robustness and truth is clear. Suppose we know that one of a set of models $M_1, M_2, \ldots, M_n$ is true, but we do not know which. If R is a robust theorem with respect to this set, then R must be true. That is, the following argument is deductively valid:

M_1 or M_2 or . . . or M_n is true.

For each i, M_i implies that R is true.

R is true.

What is much less obvious is why robust theorems should be regarded as true if we drop or change the first premise. Two alternatives need to be considered: The first is that we know that each of the models is false; the second is that we do not know whether one of them is true.

If we know that each of the models is false (each is a "lie"), then it is unclear why the fact that R is implied by all of them is evidence that R is true. Consider, for example, all models in which natural selection is said to be the only force acting on a population. This assumption has as a consequence that population size is infinite. Accordingly, this is a robust prediction for this set of models. This gives us no reason, however, to think that populations in nature really are infinite. Here the robustness of a theorem reflects the fact that the assumption is convenient, not that it is true (see also Sober, 1993).

If we do not know that one of the models is true, then it is again unclear why a joint

prediction should be regarded as true. Consider, for example, different classifications of the same set of species. Constancy of taxonomic relationships across classifications is often regarded as evidence of the truth of the joint classification. The following passage from Dobzhansky (1937) is quite typical of this line of reasoning:

> The fact is that the classification of organisms that existed before the advent of evolutionary theories has undergone surprisingly little change in the times following it, . . . The phylogenetic interpretation has been simply superimposed on the existing classification; a rejection of the former fails to do any violence to the latter. . . . [This fact's] connotations are worth considering. For the only inference that can be drawn from it is that the classification now adopted is not an arbitrary but a natural one, reflecting the objective state of things (p. 305).

Dobzhansky cannot be faulted for failing to know that cladistic analysis would reject many traditional taxa. Nonetheless, we still may ask whether the robustness of a classification really is evidence of its truth. Should the fact that a given group is recognized within a variety of frameworks be grounds for increased confidence in its reality? It is worth considering the possibility that robustness simply reflects something common among the frameworks and not something about the world those frameworks seek to describe.

When Levins describes robustness, he mentions that the models must be independent. In the example just described, the basis for a claim of independence *or* nonindependence is obscure. In another example, the nature of the dependency between models is clearer. An optimality model (MacArthur, 1965), shows that 1:1 is the sex ratio that will evolve if individuals maximize the product of the fitness "returns" from the two sexes. In this model the genetic basis of the trait is unspecified, although it is reasonable to assume in this model that phenotypes beget like phenotypes. Analysis of a model with an explicit genetic mechanism underlying the sex ratio trait (Uyenoyama and Bengtsson, 1979) indicates that 1:1 is the sex ratio that occurs when the population attains certain equilibria. These authors do not rely upon any maximization principle to derive this result. This model and MacArthur's are clearly different. On the other hand, they share many assumptions (e.g., that the population exhibits random mating, that the individuals live in a constant environment). Why then should the "robustness" of the 1:1 sex ratio make us more confident that 1:1 will evolve than if we had only one of the models to consult? A similar question has been raised by Wade (1978) in regard to the robustness of the conclusion that group selection is unlikely to occur (see also Wimsatt, 1980).

This example underscores the fact that the importance of robustness depends on the models' being independent. But how should the concept of independence be understood? Two possibilities are worth considering. The first makes use of the concept of *logical* independence. Two models are logically independent when neither implies the truth or falsity of the other. For example, a model with the assumption of random mating is not logically independent of a model with the assumption that mating is assortative; the reason is that the truth of one entails the falsity of the other. Generally speaking, competing models are not logically independent. If competing models are supposed to generate robust theorems, then logical independence does *not* describe what it means for models to be independent.

A second way to understand the concept of independence is *statistical* independence. We know that draws from an urn are independent because we understand the sampling process. The draws are independent because the probability of obtaining a green ball on the first draw and a red ball on the second is the product of the probabilities of obtaining each result. We have no comparable way to specify "the space of all possible models" in a particular case, nor to describe how the models involved are drawn from that population. To talk about two models being statistically independent, one must be able to talk about the probability each model has of being true. Standard statistical practice is to discuss the probability of the *data* conditional on a specified model. We agree with the assessment that assigning probabilities to models is not a coherent notion, the efforts of Bayesians notwithstanding.

Levins provides no help with regard to understanding what it takes for two models to be independent. Indeed, it would seem that his "protocol" for the discovery of robust predictions guarantees that the models under consideration are *not* independent. He writes (p. 423), ". . . We [should] attempt to treat the same problem with several alternative models each with different simplifications *but with a common biological assumption*" [emphasis added]. This procedure guarantees that one will find robust theorems in the very assumptions one holds constant.

Although we are skeptical that model independence can be assessed in a reasonable and unambiguous way, we do not think that considerations of robustness are without value (see also Wimsatt, 1980). Robust theorems based on "independent" models would be desirable if we could get them. Perhaps, for example, a fair meiosis explanation (Leigh, 1986) and Fisher's (1930) explanation for a 1:1 sex ratio are "independent" enough that they can be said to provide a robust truth. We have no magic test procedure that investigators can use here, only the caveat that the value of robustness depends on the models' degree of independence. This latter quantity, unfortunately, is elusive.

Since model independence is a problematic concept, it is worth considering whether robustness can be defined without invoking the concept of independence. Perhaps robust theorems are worthwhile, regardless of whether the models are independent and whether the concept of model independence even makes sense. The problem with this attitude is that robustness then comes too cheaply. *Every* statement about nature is robust in this sense because every statement is entailed by more than one model.

So far we have considered robustness to be a property that a proposition has in virtue of its invariance *across* models. It is also worth considering the concept as it applies *within* a single model. A numerical prediction of a model is said to be robust if its value does not depend much (or at all) on variation in the value of the input parameters. For example, consider a mixture of two polymorphic populations, each at Hardy-Weinberg equilibrium. Wahlund's principle tells us that the genotype frequencies in the mixture will be exactly at the Hardy-Weinberg equilibrium frequencies predicted by the allele frequencies in the mixture only if the two original populations had identical allele frequencies. However, only when the allele frequencies in the two populations differ radically will the genotype frequencies in the mixture show large deviations from the Hardy-Weinberg values. In other words, the Hardy-Weinberg prediction is robust. This type of "internal robustness" is meaningful and can be very useful if only because it allows one to distinguish between assumptions that are strong determinants of a prediction and those that are weak ones. Still it is worth bearing in mind that internal robustness is no sure sign of truth. Everything depends upon the correctness of the model within which the robustness is determined.

ROBUSTNESS ACROSS AND WITHIN DATA SETS

Levins's concept of robustness is supposed to apply to a proposition that is invariant across variation in models; data play no role in the definition of this concept. However, there are other concepts of robustness, distinct from Levins's, which are also worth considering.

One is the idea of a proposition that is well supported by each of several data sets. Such a proposition is robust across variation in data. If more evidence is preferable to less, then it is hard to see how robustness across data sets can fail to be a virtue. Suppose that two data sets favor model X over model Y. Shouldn't our confidence that X is superior to Y be stronger in the light of the two data sets than it was when we possessed only one of them? If this were always the case, then robustness across data sets would always be evidence of truth.

Clearly, the amount that our confidence increases when the second data set is taken into account will depend on the degree of independence between the two data sets. For example, suppose the two models are competing phylogenetic hypotheses and that the first data set consists of morphological measurements from the left forelimbs of the taxa under scrutiny. If the second data set consists of measurements on the right forelimbs, the data sets are probably not independent, and our degree of

confidence in the favored phylogenetic hypothesis will not be much enhanced.

It is worth noticing that the concept of independence used here is much less problematic than the one discussed earlier in connection with Levins's proposal. Here it is not models that are said to be independent, but data sets. In this case, independence can be assessed by statistical criteria alone.

The concept of robustness across data sets, however, is not without problems. There are cases in which a proposition X can be robust across data sets and fail to be well supported by the combined data set. An example occurs in the use of consensus methods in phylogenetic inference. Suppose we use cladistic parsimony (or some other method of phylogenetic inference) to identify the best tree for one data set and do the same thing for another data set. We then identify the intersection of the best trees constructed for each data set; this will include the monophyletic groups on which the two data sets agree. It is possible that this "consensus tree" can contradict the best tree constructed for the combined data set (Barrett et al., 1991). In this case, robustness across data sets is not a sign of truth. Nor is there anything peculiar about the phylogenetic inference problem or about parsimony that generates this result. The same result can occur in problems of maximum likelihood estimation (Barrett et al., 1991). Robustness across data sets may sound like a virtue, but it is important to realize that it can clash with a principle of total evidence (according to which competing hypotheses must be evaluated relative to all of the available data).

We just considered the robustness of a proposition that is common to two models, each of which is well supported by its own data set. A different concept is also worth considering, namely that of robustness across models, all supported by the same data set. Suppose that each of two competing models is reasonably well supported by the data. If R is a robust theorem that they share, should we conclude that the data support the common element in the two models? Presumably, if the data had been different, we would not have regarded the models as well supported. The question is whether we would be prepared to doubt R in this circumstance as well. If not, then this robust theorem is not tested by the data and consequently is not well supported by them. Notice that in this argument the robustness of R is not by itself a reason to believe it. Whether R is plausible depends on the data and not on the fact that R is robust. It is relevant in this context to note that testability of predictions (an obvious goal of models) depends upon having *nonrobust* theorems to test, that is, those that are not entailed by all of the models under test.

THE MEANING AND RECEPTION OF LEVINS'S CLAIMS

We believe that our arguments raise considerable doubt about the validity of Levins's claims concerning models and the concept of robustness. In particular, we doubt the existence of a trichotomy of models based upon a necessary trade-off among the three desiderata of generality, realism and precision. In addition, we believe that there has been considerable arbitrariness in how biologists locate specific models within these three dimensions. Our basis for saying this is simply that almost all claims about particular models simply assert whether the model of interest is general or realistic, and do not cite *another* model that serves as a benchmark. As argued before, it makes little sense to say without qualification that a model is general or realistic. What is required is a *comparison*; one should argue that the model of interest is *more* (or *less*) realistic or general than another.

Consider, for example, Clark and Terrell's (1978) statement that, ". . . we could abandon realism in favor of generality and precision: the Hardy-Weinberg Law is a mathematical model of this sort" (p. 302). We find it difficult to assess this claim about the Hardy-Weinberg Law without some clarification of the meaning of generality and realism. We argued earlier that these two properties are *necessarily* associated for uninstantiated models that are nested, that is, you cannot abandon just one of them. In addition, it is hard to see how the Hardy-Weinberg Law is general; surely, the point is not that the same false assumptions can be applied to many loci. Finally, we want to emphasize that readers should not be left guessing as to the set of models being compared. There should be as much "rigor" and

"repeatability" when biologists assess the characteristics and relationships of models as there is when they assess the characteristics and relationships of species.

Our belief that there has been no meaningful assessment of how Levins's trichotomy applies to models in population and evolutionary biology is reinforced by the spectrum of interpretations in the literature regarding Levins's claim of a necessary trade-off among generality, realism and precision. Some authors write as though Levins *demonstrated* that there must be such a trade-off. Others appear to accept Levins's claim in that they have made a specific judgment as to which pair of the three desiderata their model achieves (see, e.g., Cooperrider and Behrend, 1980, for a "precise and realistic" model; Wassersug and Hoff, 1979, for a "general and precise" model; and Wiens and Innis, 1974, for a "general and realistic" model). Some investigators refer to the "general impossibility" of constructing a general, realistic and precise model (e.g., see Mueller and Ayala, 1982, and Smith, 1988). Still others accompany model assessments with a specific affirmation of Levins's claim of a *necessary* trade-off (see, e.g., Leviten, 1976; Costanza et al., 1990; and Hanski and Gilpin, 1991). Standing in contrast to these investigators who either implicitly or explicitly endorse Levins's claim (or a weaker form of it) are those who cite Levins, but assert they have produced a general, realistic and precise model (Armstrong, 1988), or state or imply that such models are possible (Innis, 1975; Keefer et al., 1991). Of course, misinterpretations of Levins's claim are not his responsibility. Yet they do serve as a reminder that Levins's view of models requires more detailed scrutiny.

Even if it is wrong to think of models as *necessarily* constrained in the way Levins claimed, one could still maintain that his trichotomy describes distinctions among "most" models. To that extent, it would serve as a guide to the results one should expect when constructing a biological model. We cannot rule out the possibility that this less ambitious claim is correct. We wish to emphasize, however, that the claim can be assessed only after the critical concepts of generality, realism and precision are further clarified (see above).

IMPLICATIONS FOR RESEARCHERS

If Type III models existed—models that are general and realistic, but imprecise—then it would be appropriate to subject these models to a qualitative, rather than quantitative test. We believe biologists often assume that qualitative mathematical models should be tested qualitatively (although Levins does not specifically advocate this procedure; see Levins, 1966: 422).

A typical example of such testing is a comparison of the "trend" in data with the "trend" of predictions. For example, a model might predict both the *direction* of bias (female or male) in a sex ratio and the *degree* of this bias. A qualitative test involves determining whether the direction of bias in the population's sex ratio matches the model's prediction. A quantitative test of this model involves determining whether the degree of bias in the population matches the model's prediction given standard assumptions about the nature and extent of sampling error. The attitude that only qualitative tests are appropriate often stems from the conviction that nature is complex and biology is inherently an inexact discipline.

While we agree that qualitative testing can be scientifically useful, we think this approach is problematic for methodological and conceptual reasons. Grounds for accepting qualitative predictions are often left unstated. One consequence is that investigators sometimes use contradictory criteria to judge the same model (see examples shown in Orzack, 1990). Although this is also a potential problem in quantitative testing, the latter approach usually leads biologists to state test criteria explicitly. Qualitative assessment of fit can also be highly dependent upon the manner of graphical presentation.

The most important defect in qualitative testing, however, is that it fails to allow one to answer the most important question about a particular model: How well does the model explain the data? The significance of this question and the implications it has for model testing can best be seen by looking at a particular group of models. Optimality models (which identify the evolutionarily best behavior given a set of alternatives) are often conceived of as being qualitative models. These mathematical models play an important role in modern

efforts to understand adaptation (e.g., see Parker and Maynard Smith, 1990). In the context of clarifying the meaning of such models, it is essential to distinguish between the proposition that natural selection played an important role in the evolution of a trait of an individual, and the proposition that natural selection is a *sufficient explanation* of the trait. The latter idea means that natural selection has been so important in a trait's evolution that nonselective forces may safely be ignored. This is a stronger claim than the thesis that natural selection has been *an* important force in a trait's evolution. If the latter were true, though we would not be able to ignore natural selection, it might also be true that some nonselective forces were also important. When "adaptationists" claim that details of underlying genetics, mutation, migration, and genetic drift are of little or no importance in explaining some trait, they are saying that natural selection is, in our sense, a sufficient explanation.

Of course, an optimality model will specify "constraints" that make the optimization problem well motivated. So, for example, many life history models consider variation in the optimal timing of reproduction, while retaining a constraint on the total number of offspring produced by each of the competing phenotypes. The retention of such constraints in the model merely reflects the "local" nature of all optimization analyses.

This distinction between the importance of natural selection and the sufficiency of natural selection is at the heart of the debate about adaptationism. We believe that most adaptationists regard natural selection as a sufficient explanation for most traits. "Pluralists," we believe, admit the important role that natural selection can play in trait evolution, but regard it as just one of a number of important determinants.

Assessing the sufficiency of *specific* optimality models for explaining particular traits is one way to resolve the debate between optimalists and pluralists over the truth of adaptationism (Orzack and Sober, in press). Accordingly, it is clear why qualitative testing of models is inadequate by itself (even if test criteria are clearly delineated). Such testing fails to allow one to discriminate between the claim that natural selection is an important cause of what we observe and the claim that natural selection suffices as an explanation for the trait. For example, if an optimality model predicts a sex ratio of 0.95 (proportion female) and an individual produces a sex ratio of 0.6, a qualitative test will lead to the conclusion that the model is adequate; the quantitative discrepancy will go unexamined. However, this quantitative disagreement is precisely what should be examined; it could be due to a real lack of optimality or to ignorance or misunderstanding of the biology, such that the fact of optimality is not detected. This point remains true even if the optimality model is qualitatively correct. Assuming that quantitative agreement with an optimality model would necessarily result if model structure or details were "adjusted" is to resolve the question about the truth of adaptationism by fiat. Prior conceptions may quite naturally and acceptably play a role in determining which of these alternatives an investigator regards as likely. Nonetheless, if the sufficiency of natural selection is to be assessed, demonstrating qualitative *and* quantitative fit of the data to the predictions of *any* model regarded as explanatory cannot be avoided. It is for this reason that qualitative testing of optimality models is not sufficient if the truth of adaptationism is to be determined. Yet appropriate quantitative testing of optimality models has been done very rarely (see Orzack and Sober, in press, for further details), and consequently the truth of adaptationism remains unresolved.

This failure is not Levins's responsibility in that the concept of a "qualitative" model predates his paper. Yet there is no more compelling reason to reassess the view of models endorsed in Levins's 1966 paper than the fact that the idea of qualitative modeling has hindered the development of an unbiased assessment of the truth of one of the most important and influential hypotheses in evolutionary biology.

GENERAL IMPLICATIONS

Our analysis has several implications for research practices. The first is that qualitative and quantitative testing are both essential if one wishes to determine how well any mathe-

matical model explains a set of data. Even if a mathematical model generates qualitative predictions from qualitative assumptions (and thereby appears to be of Type III), its quantitative predictions should not go unexamined.

The second implication is that biologists should develop and apply models without being self-conscious about where particular models fit into Levins's scheme. In particular, Levins's paper has been regarded as endorsing a particular approach to scientific model building, one in which Type III models are judged superior to the other two types (Palladino, 1991). Palladino suggests that Levins presents Type III models as *the* means for developing a general body of population biological theory. Our survey of the citations of Levins's paper provides possible evidence that scientists often understand Levins's message in this way. Very few scientists have opted for a Type I characterization of their work; most have opted for Type III. Perhaps this pattern stems from the belief that a model implicitly labeled as ungeneral or unrealistic will be regarded as being of lesser scientific importance. In any case, biologists should remain open to gaining insights from many different kinds of models. At the very least, only by exploring a variety of models can we develop an understanding of how models relate to one another. In this regard, we can think of no reason why biological phenomena should be any less amenable to scientific modeling (with its implicit faith in the potential for generality) than are physical and chemical phenomena. It is of relevance that claims about trade-offs similar to Levins's have not, to our knowledge, arisen in physics and chemistry.

Another implication of our analysis is that the concept of robustness should be regarded as having heuristic value in some situations but little substance beyond this. As noted above, one might reasonably refer to the robustness of a qualitative or quantitative *prediction* if it is shown not to depend strongly upon particular values of input parameters. Yet such robustness is not guaranteed to reveal truths about nature since it may be specific to the assumptions of the model. Of course, a prediction that is robust in this sense might also be a "biological fact," but such a judgment must be based on a comparison of the prediction with data.

HOW *SHOULD* MODELS BE CLASSIFIED?

As noted at the outset, Levins's three model attributes are formal, not substantive. They could be applied to models in economics, in hydrodynamics, or in any subject. It would be fascinating to find that different disciplines exhibit different distributions of models in this three-dimensional space. It is therefore disappointing to have to conclude that Levins's three characteristics do not underwrite an analysis of this kind.

Although we see no conflict among generality, realism and precision, there is another trade-off that confronts model builders in a variety of disciplines. It is the familiar trade-off between simplicity and goodness-of-fit in curve-fitting problems. By increasing the number of adjustable parameters in one's model, one can improve the model's goodness-of-fit when values for those parameters are estimated. With respect to nested uninstantiated models, increasing the number of adjustable parameters seems to increase generality, realism and precision (at least if precision is measured by goodness-of-fit to a single data set). Simplicity is therefore a dimension additional to the ones that Levins considers. Exploring its significance, however, is beyond the scope of this paper (but see Forster and Sober, in press).

Rather than proposing another set of formal characteristics that might have the generality that Levins's taxonomy aspires to, we propose a quite different and more modest approach. Population biology models may be distinguished by their assumptions. This is not a new proposal, but simply recognizes what biologists are doing when they talk about "null" models in community ecology or "neutral" models in population genetics or "optimality" models in behavioral ecology. The same point applies to models in other disciplines. Conflicts between models should be understood in terms of their substance, not their styles. Levins's thesis suggests that it is a matter of taste which two of the three desiderata a scientist chooses to pursue. If different models really did have different aspirations, it would be a mistake to attempt to decide which model is best. Models in different categories would then "go their own ways," an-

swerable to their own standards. The perennial difficulties of understanding nature make this pluralism a temptation. It is, however, a temptation to be resisted. One way to do this is by separating models according to their assumptions and by their degree of fit to data. The resulting classification, though it abandons the quest for descriptions of models that are subject-matter independent, has the merit that it classifies according to features that matter to the goals of science itself.

ACKNOWLEDGMENTS

We thank John Beatty, Marty Kreitman, Egbert Leigh, Richard Levins, Richard Lewontin, Paolo Palladino, Peter Taylor, Bill Wimsatt, and two anonymous reviewers for advice and comments. S. H. O. wishes to thank Dana Wrensch and Jerry Downhower for supporting a research visit to the Department of Entomology, Ohio State University, during which work on this paper was carried out. S. H. O. also thanks Brian Charlesworth, Marty Kreitman, and the Department of Ecology and Evolution at The University of Chicago for support.

REFERENCES

Armstrong, R. A. 1988. The effects of disturbance patch size on species coexistence. *J. Theor. Biol.*, 133:169–184.

Barrett, M., M. Donoghue, and E. Sober. 1991. Against consensus. *Syst. Biol.*, 40:486–493.

Beatty, J. 1981. What's wrong with the received view of evolutionary theory? In P. Asquith and R. Giere (eds.), *PSA 1980*, Vol. 2, pp. 397–426. Philosophy of Science Association, East Lansing.

Clark, J. T., and J. Terrell. 1978. Archaeology in Oceania. *Ann. Rev. Anthropol.*, 7:293–319.

Cooperrider, A. Y., and D. F. Behrend. 1980. Simulation of forest dynamics and deer browse production. *J. For.*, 78:85–88.

Costanza, R., F. H. Sklar, and M. L. White. 1990. Modeling coastal landscape dynamics. *BioScience*, 40:91–107.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. Columbia University Press, New York.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*. Oxford University Press, New York.

Forster, M., and E. Sober. In press. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *Brit. J. Philos. Sci.*

Hanski, I., and M. Gilpin. 1991. Metapopulation dynamics: brief history and conceptual domain. *Biol. J. Linn. Soc.*, 42:3–16.

Innis, G. 1975. One direction for improving ecosystem modeling. *Behav. Sci.*, 20:68–74.

Keefer, B. J., J. L. Smith, and T. G. Gregoire. 1991. Modeling and evaluating the effects of stream mode digitizing errors on map variables. *Photo. Eng. and Rem. Sens.*, 57:957–963.

Leigh, E. G., Jr. 1965. On the relation between the productivity, biomass, diversity, and stability of a community. *Proc. Natl. Acad. Sci. USA*, 53:777–783.

———. 1986. Ronald Fisher and the development of evolutionary theory. I. The role of selection. *Oxf. Surv. Evol. Biol.*, 3:187–223.

Levins, R. 1966. The strategy of model building in population biology. *Am. Sci.*, 54:421–431.

Leviten, P. J. 1976. The foraging strategy of vermivorous conid gastropods. *Ecol. Monogr.*, 46: 157–178.

MacArthur, R. H. 1965. Ecological consequences of natural selection. In T. H. Waterman and H. J. Morowitz (eds.), *Theoretical and Mathematical Biology*, pp. 388–397. Blaisdell Publishing, New York.

Mueller, L. D., and F. J. Ayala. 1982. Population dynamics in the serial transfer system: comments on Haddon's model. *Am. Nat.*, 120:548–550.

Orzack, S. H. 1990. The comparative biology of second sex ratio evolution within a natural population of a parasitic wasp, *Nasonia vitripennis*. *Genetics*, 124:385–396.

Orzack, S. H., and E. Sober. In press. Optimality models and the test of adaptationism. *Am. Nat.*

Palladino, P. 1991. Defining ecology: ecological theories, mathematical models, and applied biology in the 1960s and 1970s. *J. Hist. Biol.*, 24:223–243.

Parker, G. A., and J. Maynard Smith. 1990. Optimality theory in evolutionary biology. *Nature*, 348:27–33.

Smith, E. A. 1988. Realism, generality, or testability: the ecological modeller's dilemma. *Behav. Brain Sci.*, 11:149–150.

Sober, E. 1993. Mathematics and indispensability. *Philos. Rev.*, 102:35–58.

Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. 1983. The double-strand-break repair model for recombination. *Cell*, 33:25–35.

Uyenoyama, M. K., and B. O. Bengtsson. 1979. Towards a genetic theory for the evolution of the sex ratio. *Genetics*, 93:721–736.

Wade, M. J. 1978. A critical review of the models of group selection. *Q. Rev. Biol.*, 53:101–114.

Wassersug, R. J., and K. Hoff. 1979. A comparative study of the buccal pumping mechanism of tadpoles. *Biol. J. Linn. Soc.*, 12:225–259.

Watt, K. E. F. 1956. The choice and solution of mathematical models for predicting and maximizing the yield of a fishery. *J. Fish. Res. Board Can.*, 13:613–645.

Wiens, J. A., and G. S. Innis. 1974. Estimation of energy flow in bird communities: a population bioenergetics model. *Ecology*, 55:730–746.

Wimsatt, W. C. 1980. Randomness and perceived-randomness in evolutionary biology. *Synthese*, 43:287–329.