Experiments in ecology and management: Their logics, functions and interpretations

A. J. UNDERWOOD
Institute of Marine Ecology, Zoology Building,
University of Sydney, NSW 2006, Australia

Abstract
Experimental tests of clearly articulated hypotheses are an increasingly widespread feature of modern marine ecology. Increased use of experiments has not, however, been accompanied by increased understanding of the logical structure of falsificationist tests. Most observations can be explained by several different models or theories. To distinguish among these requires demonstration of the falsity of the consequences or predictions of incorrect models. This is best achieved by deriving from each model one or more hypotheses (predictions) about the type, form or nature of observations that should occur in some not-yet-examined set of circumstances. Because of logical constraints on the possibility of proving the correctness of such hypotheses, they must be inverted to form logical null hypotheses which comprise all alternative possibilities to those predicted in the hypotheses. Correctness or not of null hypotheses can then be ascertained by an appropriately designed experiment (or test), leading to unambiguous rejection or retention of the null hypotheses. The former corroborates the hypotheses and provides support for the correctness of the explanatory model for the original observations. In contrast, retention of a null hypothesis identifies an incorrect model. The growth of knowledge is thus the elimination of false models, theories and explanations.

Ecological experiments usually require statistical procedures for determining whether or not null hypotheses should be retained. Construction of statistical null hypotheses (i.e. definitions of parameters of frequency distributions of test statistics) sometimes requires that these be identical to logical hypotheses (and not to the logical nulls). This leads to irrational acceptance of hypotheses and the models or theories from which they were derived. It also poses immense problems for determinations of statistical power of experiments.

Ecological experiments are analysed to reveal the nature of, and linkages between, their components in relation to falsificationism, statistical procedures and the logical properties and interpretations of ecological theories.

Introduction
‘Experimentation has a life of its own’ (Hacking 1983)

There has been considerable interest amongst ecologists in recent years to demonstrate their grasp of the fundamental, or sometimes only the surface features, of philosophical debates about the methods of science (Diamond 1986; Loebke 1987; Mentis 1988). Much of this has been of interest, some papers have simply demonstrated ignorance and a lack of common sense and others have attempted to seem and bland dismissal of the need for ecologists to be logical, or at least to be able to discuss coherently the methodologies they use to advance our understanding of natural populations and environments. There has also been some attempt to discuss critically the concepts and approaches to development of general models for understanding, describing or interpreting the complexities of interactive, multi-species systems (e.g. Dayton 1975; Simberloff 1984; Underwood & Denley 1984). Such critical appraisal has not usually met with open acceptance of the need for greater rigour. Finally, there have been attempts to discuss the problems of specific types of experimental designs and the great difficulties involved in construction of experimental analyses for field situations, where the power of statistical tests is limited by logistical and other constraints (Green 1979; Hursher 1984; Stewart-Oaten et al. 1986; Underwood 1986).
Despite this welter of publications, there has been little serious attempt to discuss the philosophical or logical bases of the actual procedures used by experimentalists, in order to bring together the different strands of epistemology and historical analysis that underpin any attempt at rigorous analysis of ecological problems. Serious difficulties beset practical and academic ecologists. An increasing plethora of environmental problems needs solution (including pollution, management of natural resources, and conservation of habitats and rare species). Yet, despite much legal machinery to ensure some consideration of ecological contributions, ecologists have not yet created sufficiently rigorous and agreed methods for solving problems.

In this paper, I consider the intertwined threads of logic, philosophy and statistical inference necessary for the design, completion, analysis and interpretation of ecological experiments. I have deliberately avoided the problems of demarcation (sensu Popper 1968, 1969), that is, those criteria by which scientific activities and arguments are separable from other types of human endeavour. My position on that has been described elsewhere (Underwood 1989a,b; 1989b). Instead, I wish to consider the problems and consequences for those who have already decided that they are going to use experimental procedures (be they manipulative or not) to attempt to distinguish amongst competing potential explanations for ecological phenomena (e.g., Underwood 1983a). In particular, I focus on the central role of logical and statistical null hypotheses, a topic not addressed in most of the mainstream philosophical analysis of experiments. These are complex and there are many interactions; different problems lead to different procedures, which have different logical structures (hence my use of ‘logic’ as a plural). A major problem confronting experimentalists is identified in the context of the logical development of hypotheses, the rigorous design of experiments and the statistical problems of power and errors of interpretation.

The logical components of a falsificationist test

As detailed below, a refutational or falsificationist test has several distinct components with different properties or rules which are related in specific and invariable ways (Fig. 1). At each point along the sequence of components, it is crucial to be able to identify the steps and to ensure that their logical connections are created, preserved or maintained. Different authors routinely use a variety of terminologies for these components. What matters is not the names given to each piece of the procedure, but that each part is identified properly and made explicit. The terms as I shall use them, are defined below.

Observations (puzzles, problems)

Most ecological research programmes (Lakatos 1974) begin with the recognition of some pattern, or departure from it, that is, a sequence of observations in space or time, or some anomaly in such a sequence. It is, of course, possible to start with a major interest in some process and then find situations in which who may be operating, but that has inherent problems of invoking a process in some system or habitat, or for some species, for which there is no evidence that the process actually operates. Here, therefore, I shall assume that experimental components of investigations actually start with observations, however casual or general (Fig. 1).

Some philosophers have called the observations a puzzle to be solved (Wittgenstein 1922) or a problem (Popper 1934, 1962). The terminology does not matter; the important feature is that there is some phenomenon to be explained (which is a scientific task, if it is to be explained in scientific terms, but that leads back to the problem of demarcation and is outside this discussion).

Ecological observations (if they are, in fact, true; see Andrew & Mapstone 1987 for a discussion of identification of patterns in nature) include such statements as ‘there are more of Species X on the tops of boulders than underneath’, ‘more species are found in some (defined) places (e.g. the tropics) than in other (defined) places (e.g. temperate areas), ‘fish are smaller on reeds where people are trying to catch them than on other reeds’. Note that many of the present examples are from marine habitats. No loss of generality will result, however — the conclusions are valid for all areas of ecology.

A basicism of falsificationism is that the observations are not likely to be, or are not at all, free of any prior theory. This point is discussed later and does not matter for the following discussion of the components that come after the observations.

Models (theories, explanations)

Any attempted explanation for existing observations is valid provided that it does actually explain the observations (as best as one can currently tell, given the available technology and the currently available information or which to base decisions about the validity of the explanation). Thus, a model is a series of statements or formulae that provide a reasonable and realistic account of why the observations have been made (Nagel 1961). There are no other requirements for a model (Fig. 1).

Many other authors use the terms ‘theory’ or ‘explanation’ where I am using ‘model’ (see other terms in Trusted 1979). Many authors use the term, ‘hypothesis’, but that has become inextricably confused with the next step in this procedure. Again, the terminology does not matter, only that the model (theory or explanation) provides a realistic possible account of why the observations have been made. The term ‘explanation’ might be better reserved for that property of a model or theory that it does explain. A model can be sophisticated or simplistic, naive or refined (Nagel 1961), but it must be a legitimate possible explanation for the events observed. Thus, the repeated pattern of rise and fall of seawater that we tend to associate with the tides could be explained by the legend that a sea-monster lives off the coast and has a regular pattern of breathing which causes a pattern of rise and fall of the water.

This apparently fanciful model demonstrates two important points. First, it is not the terminology nor the complexity of a model that gives it some status of correctness or scientific validity. Second, it is usually the case that several different models can be invoked that are equally capable of explaining the observations, but differ in fundamental ways. Thus, tidal rise and fall may equally well be explained by some relationship with the pattern of rotation of the earth—moon system and the earth on its own axis (although this usually involves complex and, to the casual observer, quite fanciful notions about forces of gravity; Darwin 1844).

Comparing, but not exclusive, models for the observation that ‘more of X are found on tops of boulders’ include, in turn: ‘species X needs sunlight, which is not available under boulders’ or ‘species X is vulnerable to predators that only forage underneath boulders’ or ‘species X cannot usually settle as larvae underneath boulders’, depending upon whether your favourite explanation for ecological patterns is about physical factors, predation or recruitment. Combinations of these or alternative, more complex models. Articulation of a model is insufficient to demonstrate its validity and some procedure is needed to contrast or compare different models that can be proposed to explain some observation.

Hypotheses (predictions, deductions)

One approach that has considerable support among scientists of all sorts, including many who do not actually believe in falsificationism, is the attempt to use the model to make predictions about the world and then to set out to prove the worth of the model by demonstrating that its predictions come true. The problems with this are discussed next, but it suffices to
demonstrate that a model can be evaluated by using it to predict some new, as yet unexamined, set of observations. Such predictions are deduced from the model (Fig. 1) and are hereafter called 'hypotheses'. The only rule about hypotheses is that they must follow the following frame of reference: if the model is correct, then under some (defined) new set of circumstances, I predict that the following observations will be made, because the model is correct (see, particularly, Hempel 1966). Given the example of Species X listed above, the first model leads, among other possibilities, to the following hypothesis: 'Because Species X is limited to the tops of boulders by the lack of light underneath, if I put shades over the tops of boulders to make the intensity of light equal to that under boulders (creating circumstances in which no observations have yet been made) then the number of species of Species X should decline to match those underneath boulders'.

Note that this prediction is not likely to be true if the important process was recruitment. If recruitment arrived, there is no reason under that model for the existing members of Species X to disappear, nor if the model about predation was correct (unless the predators respond to the reduced light, which would form the explanation for why reducing the light intensity was important, if that turned out to be the observation when the test was done).

Idealized hypotheses that might be deduced from any of the models, those that are most different from the hypotheses deduced from possible alternative models should be the ones chosen for further examination. It will generally be easier to attempt to distinguish among alternatives that are very clearly distinct. In this, I disagree with Popper (1968, 1969) who stated that the most unlikely conjectures (the ones with the greatest chance of being wrong because they are complex, or have long chains of linked assumptions, or are very specific or general) are the most rewarding. He did not deal with complex biological scenarios where many alternative models can usually be invoked to explain a pattern of distribution, abundance or so forth of organisms (Platt 1964; Chamberlin 1965). A major problem with many discourses about the philosophy of science is that they only deal with simple sciences like physics. The most prob-

able method of proceeding in ecological research is to devise as many models as possible, and then to deduce from each the most different hypothesis(es), because this provides the maximum possibility of contrasting the various models (see also Underwood 1985A for a set of models and their hypotheses in the distribution of marine algae).

Proof and disproof

Having deduced one or more hypotheses from the one or more models of interest to explain a set of observations, there remains the task of critically evaluating them. Some reason, it is still (despite a long history; Whewell [1837]) widely assumed that scientists set out to prove their hypotheses—a task with no logical possibility of succeeding; among others, Hume (1779, 1939) and Popper (1968) are renowned for demonstrating that proof is not a possible outcome for anything but the most trivial set of propositions in nature.

Proof requires that every possible set of circumstances like that identified in the hypothesis be examined and that every one of these cases conform to the prediction in the hypothesis. Suppose that a hypothesis has been proposed from some model (not specified in detail here) that there are no sharks in a certain bay or estuary. Proof of this hypothesis would require intensive sampling for sharks throughout the bay, and no sharks being found. Of course, to avoid the possibility that the sharks move from areas that have not yet been sampled to other parts of the bay, the whole bay needs to be systematically sampled in such a way that the sharks cannot move into areas already sampled, or, better, that the entire bay be examined simultaneously. The intensity of sampling, the simultaneous requirement that the whole area be sampled with nets of very small mesh-size and the logistic constraints make it impracticable, even if theoretically possible, to demonstrate that there are no sharks present.

Arguing that less intensive sampling is all that is required leaves one open to the problems of Bertrand Russell's (1912) inductivist turkey (which was beheaded on the evening of the day on which he finally convinced himself that he had been fed on enough mornings to prove that breakfast always arrived in the morning). Without being able to examine the entire relevant universe, no number of observations in a subset will prove anything. Inductive generalizations of this sort are doomed to fail because it makes no difference how many confirmatory cases are observed (we have sampled many places and there are still no sharks—so far so good), an untrue hypothesis is still untrue. Confirmatory cases, however numerous, cannot alter that (despite the detailed analyses of Carnap (1962) and Hempel (1965)).

There is, however, as brilliantly argued by Popper (1968, 1969), a huge imbalance between the onerous burden of proof and the relative simplicity of disproof. One contradictory case is, in theory, all that is required to disprove a statement (or hypothesis). Thus, the discovery of one shark in one of the samples, at any stage of the procedure, will certainly disprove the hypothesis that there are no sharks.

An attempt to disprove has two immense advantages. First, it dictates how many contradictory observations will cause one to stop the test. Second, disproof is a compelling tautological logical, known as the modus tollendo tollens (e.g. Lemmon 1971; Hocutt 1979), which provides a rational justification lacking from confirmatory attempts at proof. The logical argument is simple (Table 1). If, as a proposition, some statement (p) implies (is associated with or causes) some other statement (q), the discovery that the consequent statement (q) is false negates the procedent (p). When q is wrong, p must be wrong.

Null hypotheses

Because of the logical impossibility of proof, the alternative strategy, seeking disproof or attempting to falsify, has become the most powerful method for parsing ecological problems. This involves the construction of a null hypothesis, which is carefully identified to include all possibilities except the actual proposition in the hypothesis (Fig. 1). This basis for construction of logical null hypotheses is in complete contrast to the notion that a null hypothesis somehow has logical priority over other alternatives (Strong 1980). The logical null hypothesis as described here is made necessary by the impossibility of proof and stems from the previous (primary) statement of the hypothesis, which itself is derived from a model, originally proposed to explain some observations. Thus, it is called a null hypothesis, fact, secondarily derived and appears fairly late in the sequence of steps in an experimental procedure.

As an example of a null hypothesis, consider the situation where a hypothesis states that the numbers of some species will increase when predators are removed. The null hypothesis is then simply phrased as the numbers will not change, or will, in fact, decrease when predators are removed. Disproof of this null hypothesis leaves the hypothesis as the only alternative, but obviates the need to attempt to prove the hypothesis. Using the modus tollendo tollens (see Table 1), the null hypothesis predicts no null hypothesis, if these data are not forthcoming because something else happens, the null hypothesis is known logically to be false—it has been falsified, or disproven. As another example, the hypothesis that there should be a negative correlation between the diversity of species of meiofauna and the concentration of some pollutant leads to the null hypothesis that there will be positive, or no correlation, between these variables.

These examples are one-tailed hypotheses, but the logic works for two-tailed situations. If the hypothesis is that application of some fertilizer will alter the diversity of litter (by increasing or decreasing the number of spe-
cies), the null hypothesis states that application of the fertilizer will have no effect. In all cases, as defined here, the logical null hypothesis excludes only the stated hypothesis.

Note that the proposed derivation and definition of a logical null hypothesis is not the same as the definition of a statistical null hypothesis, which will be considered later. For experimental procedures where statistical analyses are necessary (most ecological circumstances), there is a great potential for confusion. Statistical tests require the construction of a frequency distribution of the chosen test statistic: under some stated circumstances, this construction requires definition of the mathematical parameters of the frequency distribution (usually called the null hypothesis or the null distribution of the test statistic). In the cases so far considered, the test statistic is derived from the logical null hypothesis. Thus, the last example can be examined by sampling experimental plots with the fertilizer and comparing the mean number of species of weeds with that in a sample of control plots (with no fertilizer). A test statistic that uses information about the mean number of species of weeds (e.g., a t-test) or the difference between modes (e.g., a Wilcoxon test), which should both be zero under the logical null hypothesis, or a test statistic that uses the ratio of the mean in control to experimental plots (which should be 1 under the null, etc.) will suffice as a statistical procedure (Snedecor & Cochran 1967). Evaluation of the statistical null hypothesis (the demonstration of a non-zero mean or mode) provides support for the hypothesis because its predictions have been borne out.

As will be demonstrated later, there are some hypotheses for which the statistical null distribution must be derived from the hypothesis and there is no identifiable logical null hypothesis. This situation is fraught with great difficulty.

Tests (experiments, critical evaluations)

Having decided that an experimental procedure is going to be used in some ecological study, there is little point in debating the merits of different types of experiments. The only reason for doing the experiment is to subject the model proposed to account for some observation to test, i.e., to evaluate whether it is of any worth as an explanation, or, more formally, to attempt to falsify its predictions and thereby demonstrate that it is wrong. Medawar (1969) reviewed the four types of experiments available to scientists (demonstrative, Aristotelian, inductive or Baconian, deductive or Kantian—thought-experiments and critical, or Galilean) tests. Aristotelian tests consist of noticing new observations that fit with the prediction of the hypothesis. Thus, they are not supported by the logic of the modus tollendo tollens and are based on naive inductivism. The illogic of such tests will not be considered further here.

Baconian experiments are also still popular amongst some ecologists (see the argument that 'giving the world a shove may reveal something about the processes operating'). The supposed justification for spending the tax-payers' money on such games is that we do not know enough about the world to propose any models for the observations we already have. Therefore, we shall run around gathering information of all sorts in the hope that it will be revealing. The problem is that the procedure contains no explicit mechanisms to organize what sorts of information to collect, nor to guide the investigator to reach a conclusion about what it is that the collected data have actually revealed. Baconian data-gathering (Bacon 1620) is coupled with requests for ever-increasingly expensive equipment and logistic support (boats, laboratories, etc.) purely because other researchers have access to the sorts of data available with this support, the whole notion becomes not simply irrelevant, but also a matter of the vanishing of scientific resources (Underwood 1988).

In fact, it can usually be demonstrated that these Baconian tests (despite what their practitioners state) are not hypothesis-free, but are designed to test unarticulated predictions from unidentified models. As a result, it should come as no surprise that the results are often difficult to interpret. It is extremely doubtful that anyone actually derives into ecological problems without any notions about the way the world works (see 'theory-laden observations' later). It would be more profitable to identify the steps in the second of the procedure outlined in Fig. 1 before embarking on the actual tests. Then, the appropriate experiment (sampling, manipulation, whatever) can be designed with proper attention to the requirements of the model and its subsequent analysis and interpretation (Connor & Simberloff 1979, 1986; Connell 1980; Underwood 1981, 1986, 1989; Hurlbert 1984; Underwood & Denley 1984).

There is also considerable room for improvement in the methods of statistics that are used in the planning and analysis of experiments (Green 1979; Underwood 1981, 1986; Hurlbert 1984).

Conclusions (interpretations or inferences)

A fair and realistic test of some null hypothesis is one that has only one of two possible interpretations. The first is that the null hypothesis should be rejected, which (by use of the modus tollendo tollens, see earlier and Table 1) allows rational support for the hypothesis and, if the logical construction of the hypothesis has been correct, support for the model. This route leads one back to the model with some faith in its relevance or correctness (Fig. 1). The next step in the research program is to refine the model, either by making it better capable of precise prediction at some smaller or more complex scale (i.e. to make the model more detailed), or to refine it by making it more general and broadly compelling (i.e. by making it capable of predicting patterns and processes at a larger or more widespread scale).

Having discovered that the processes leading to elimination of a species from some part of its range involve competitive interactions for food, the model (i.e. competition explains the pattern of distribution of the species) can be made more precise by reformulating it to explain the particular pattern of abundance of the species relative to the available amounts of food and the numbers of the competitor in different sites. This would lead to predictions about the actual numbers of animals killed by competitive interactions, or about the amount and pattern of population growth. Whatever the difference in the number of species of food and competitors. Alternatively, the model could be expanded to include explanations of other species that have not yet had their distributions explained (competition explains Species X; therefore I propose the more general model that competition is also important for Species Y and Z). In either case, the procedure requires that the newly-expanded model be tested by deducing specific new hypotheses from it and then, testing to test them. This is particularly important where models are being made more general. Without further testing, acceptance of an expanded or more general model is simply inductivist. To use experimental proce-
dutes to test the first model, but not to bother when it is expanded is illegitimate.

The only alternative conclusion from a fair test is retention of the null hypothesis (Fig. 1). Because of the definition of the null hypothesis (above), retention of the null hypothesis means that the model must be rejected (if the model's predictions are properly deduced and are not borne out, the model must be wrong). Notice particularly that the decision to reject the hypothesis (and the model from which it was derived), when the null hypothesis is retained, is also a falsification-test, well-based on the logic of the modus tollens. It is simply the inverse of the previous one (i.e., for any data (a) implied by the null hypothesis, the opposite data (b) are implied by the hypothesis). Therefore, if the prediction of the null hypothesis (a) is borne out, the opposite prediction of the hypothesis is not and the hypothesis (and model) is (are) negated using the modus tollens. Either way, there is a good logical basis for arriving at a conclusion about the truth or falsity of the hypothesis. When the null hypothesis is retained, the original set of observations has been expanded by the new ones gained during the test, and the entire set of observations is now warranting explanation, noting that the previous model is not correct. Progress can therefore be made very rapidly by failing to reject null hypotheses, because that procedure eliminates incorrect models as explanations for observed phenomena. Whatever the outcome, the experimenter embarks back towards the beginning of the procedure to the model (if the null hypothesis has been rejected) or to the observations, right at the start, if the null hypothesis has been retained (Fig. 1).

If any other conclusion is arrived at during the test, that is, if the result is neither rejection nor retention of the null hypothesis, the test was not a test of the null hypothesis at all. For example, consider the model that some species of alga is only found in small abundance in some parts of a shore because grazers consume the plants in those places. In other places, the algea are abundant because grazers are absent. One possible hypothesis is that removal of grazers from such areas will lead to an increase in algal abundance compared with that in controls where grazers occur naturally. The null hypothesis (that there will be no difference in algal abundance when grazers are removed) can only be tested if algea are not all about to die from some other process (so that the original observations are not going to be preserved throughout the experiment). Under these circumstances (where the control and experimental plots do not differ at the end of the test, the null hypothesis is apparently to be retained, but there are no algae present in the controls, nor in the experimental), there was no test of the effects of grazers. To identify processes that create some observed pattern, the processes must be operating during the experiment, and the pattern must be observable during that period. Thus, control observation of natural abundances of grazers are needed and must continue throughout the experiment to look different from other controls where grazers are naturally absent.

Similarly, the model that there are more barnacles in some areas than others because of predation can be tested by proposing the hypothesis that removal of predators from replicate areas will lead to an increase in numbers of barnacles, eventually to match those in areas where there are no predators. The null hypothesis predicts no change (or a decrease in numbers in experimental plots) relative to controls where there are predators. If, however, during the experiment, the predators wander out of the control areas, there will be no difference between the mean numbers of barnacles in experimental and control plots. The null hypothesis will be retained. In this case, however, there was no test, the conditions under which the hypothesis was formulated were not maintained during the experiment. Where a null hypothesis cannot be rejected, nor unambiguously retained, the test was not a test and the procedure must start again.

Problems for interpretation

There are, of course, several possible problems in the interpretation of ecological experiments, even when appropriate detailed attention has been paid to the logical structure of the tests. One very serious issue is that of statistical errors, the incorrect rejection or retention of null hypotheses when probabilistic statistical procedures are used for making the decision but, by chance, non-representative samples are obtained in the experiment. These are well-known, but are briefly reviewed below. Then, two common abuses of procedures that are widespread in ecology are considered in relation to the logic of experimental protocols.

Statistical problems (types of error)

There are two possible statistical errors in ecological experiments (where statistical tests are used, e.g. Winer 1971). First, a perfectly correct null hypothesis may be rejected. This happens, for example, whenever a sample of some population of relevance to the null hypothesis is not representative of the population. Consider a test of the null hypothesis that there will be no difference between the mean number of Species X in areas where predators have been experimentally removed and the mean in areas where predators operate. If the null hypothesis is, in fact, true (and predators make no difference to the abundance of X), the samples of say, quadrats in experimental and control areas should have the same mean, except for sampling error (i.e. each of the samples is estimating the same population mean). In practice, however, it is always possible to get two samples from the same population and discover that the sample means are statistically significantly different. One may be usually large (it overestimates the population's true mean). The other may be too small (it underestimates), or both may be occurring. This type of error is known as Type I or alpha error (Table 2). The probability of this occurring by chance when the null hypothesis is true is determined by the experimenter before the experiment (by convention, this is often chosen to be a probability of 0.05).

In contrast, there is a possibility that a false null hypothesis will not be rejected because of sampling error. Consider the situation in the previous example, when the null hypothesis is false (predators do influence the abundance of Species X). The mean abundance in a sample of the experimental areas should be greater than that in the sample from the control areas (where predators are active). It is nevertheless still possible for a random sample from the control area to be an underestimate, that from the experimental area to be an underestimate, or both. Under these circumstances, the two samples will not be discovered to be significantly different. This error (failure to reject a false null hypothesis) is known as Type II or beta error (Table 2). The probability of this occurring by chance is often not known for any particular experiment (Siegel 1953; Cohen 1977; Underwood 1981; J & She 1983; Andrew & Mapstone 1987). The power of a test is the probability that the test will correctly cause rejection of a false null hypothesis; thus, the power of the test is (1 - probability of Type II error) (see Fig. 2).

The probability of alpha error can be controlled by the experimenter, and, at least in theory, the probability of beta error can also be controlled by use of a power analysis prior to the experiment (Winer 1971; Cohen 1977; Underwood 1981). For this to be effective, it is
necessary to be able to state with some precision the alternative to the null hypothesis under consideration. Thus, if the hypothesis is that predators have some influence, the precise alternative to the null (no influence) is not known. If the hypothesis can be stated in terms such as 'predators will remove 30% of the prey', then the magnitude of the difference between the two areas, if predators are indeed relevant, is known and the power of the test can be calculated. A test of suitably large power could then be done, provided that logistical constraints do not prevent it. Although there have been attempts to bypass the need to identify the minimal magnitude of an alternative to the null (i.e., to make the hypothesis quantitative rather than qualitative; see Cohen 1977), there is no method to calculate power in advance of an experiment when the magnitude of the alternative has not been specified. It is, however, sometimes useful to determine the power of experiments retrospectively (Underwood 1981; Andrew & Mapstone 1987).

**Interactions of logical and statistical problems**

There are several ways in which statistical and logical problems in experimental procedures can interact (Table 3). Table 3 is arranged to show the theoretical outcome on models, when there is confusion about the construction of logically derived hypotheses, statistical error, or both. In each case, the model is listed as true or false (which, of course, is not known in real life, otherwise the exercise is pointless). In Case 1 in Table 3, the hypothesis is logically sound and reflects the known features of the system under investigation. As a result, it is true, the null hypothesis is therefore false (see above) and the test produces the result that the null hypothesis is false and it is rejected. As a result, the alternative hypothesis is supported. At this point, there is logical validity in rejecting the null and thereby supporting the hypothesis (Table 1 and note the definition of a logical null hypothesis as given above). There remains the problem, however, that any model leading to the same predictive hypothesis has also passed this test (Chamberlin 1965). This is not solved in this procedure (Caswell 1976). Hesitancy and effort must be used to try to formulate all alternative models for the observations at the start of the proceedings (Platt 1964; Chamberlin 1965) so that they can be subjected to test, by construction of conflicting hypotheses. This is also the reason why Newtonian mechanics survived for so long, despite being inadequate as a model to explain the observed patterns of motions that had been observed over many years (Chalmers 1979).

In this scheme of ecological investigation, the experimenter returns to the model to refine it. If it is wrong, and some process explains the observations better or instead, the model will fail in too many iterations. Of course, one naively believes that the truth has been established by one test, the alternative models will never be sought and the first model that produces rejectable null hypotheses will always be taken to be correct. This is not a problem for this scheme that is not matched in any other protocol. The important point is that the procedure is a closed loop. At any point in the future, if someone suggests an alternative model for the observations, the procedure must be used again to attempt to eliminate it. As better explanations are found, the model becomes more secure. The selection of a null hypothesis is a process of elimination, and the process of selection is linear, not circular.

<table>
<thead>
<tr>
<th>Case</th>
<th>Model</th>
<th>Hypothesis</th>
<th>Null Hypothesis</th>
<th>Null Hypothesis appears</th>
<th>Conclusion/error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>Reject H₀, dont H₁</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>Reject H₀</td>
</tr>
<tr>
<td>3b</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>Support incorrect model</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>Reject H₀</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>Support incorrect model</td>
</tr>
<tr>
<td>6b</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>Support incorrect model</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>Reject H₀</td>
</tr>
<tr>
<td>8a</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>Reject incorrect model</td>
</tr>
</tbody>
</table>

*Logically correct inference about the model.
**Logically correct inference about the model; two wrongs make a right!
* T = true, F = false.

In two of the logical parts of the protocol were correct but statistical errors (types I and II respectively) caused the wrong conclusion to be made about the null hypothesis. Case 2 is unfortunate because it causes the experimenter to return to the original observations with the new ones obtained during the test and to start again. This will tend to obfuscate the worth of the original model and seriously slow down its eventual resurrection into the thinking, and concomitant questioning of the worth of the previous experimental conclusions.

In Case 7, an incorrect model was supported because of erroneous rejection of a true null hypothesis. The probability of this type of error is always under the control of the experimenter. Repeated experimentation is unlikely to repeat the error under most normal circumstances. This type of situation warns against over-reliance on single sets of experimental data.

There are also two cases where logical mistakes occur (Table 3). In Case 4, the hypothesis is incorrectly deduced from the model. An example is shown in Table 3. In this case, consider the model that competition between two species (A and B) explains why one is more abundant in some parts of the habitat and the other more abundant elsewhere. The model requires that the species compete (for some resource in short supply) and that the chances of 'winning' are different in the two parts of the habitat. The hypothesis can be proposed that removal of one species (A) from some parts of the habitat where the other is not abundant will cause numbers of the second species (B) to increase by migration into the area. The null hypothesis is that removal of one species will not affect (or will decrease) numbers of the other species. The experiment is done (with suitable controls and replication) and the null hypothesis is retained; there was no migration of B into the experimental areas, and thus no increase in abundance of B as predicted in the hypothesis. This could have been because the animals do not migrate in response to changed circumstances. The pattern of distribution is set by competitive interactions operating whilst they are young and they do not then move around sufficiently when older to discover that the competitor has been removed. The logical flaw here is that the experimenter is trying to make too specific a hypothesis from a general model. The hypothesis that should
have priority is that removal of A will lead eventually to an increase in B (by whatever process, such as migration, increased arrival of propagules etc). The general model of competition would then have been sustained in an experiment that continued for long enough. Of course, if the pattern originally observed was created by competition that no longer operates (Connel’s (1980) ‘ghost of competition past’), the model will continue to be rejected (an example of how evolutionary timescales and unique historical events are not amenable to logically justifiable, unambiguous interpretations in ecology). During the experiment on the more general hypothesis, if abundance of B increased in experimental areas because of increased recruitment of young rather than migration, the next round of refinement of the model can be quick and efficient. More detailed explanation is possible in the light of the result. Of course, it might always be better to assume that the general model ‘competition caused it’ is too general, and to propose a series of possible competitive models (competition on propagules, competition for pre-emption of resources, competition through interference between adults of the two species etc.). These could be used to make different predictions about the process leading to changes in numbers, and to its time-course and scale before the same sort of experimentation was started. Clarity of thought about the proposed model will always aid in avoiding this sort of logical error.

In Case 3 in Table 3, the opposite logical error occurs and an incorrect model is apparently supported because a wrongly deduced hypothesis is supported by rejection of its null hypothesis. Consider the same observations, species A is more abundant in areas where B is sparse and vice versa. Competitive interactions between the two species are proposed as the model to explain this pattern, and the specific hypothesis is that removal of A from appropriate areas will lead to increases in B compared with controls. The null hypothesis (removal of A makes no difference) is rejected, B increases in areas where A was experimentally removed. The correct mechanism was, however, that a disease carried by Species A reduced the density of B where A was present. Death due to disease is not the result of a competitive interaction. When A was removed, B increased because it was released from the disease, and not because of any competitive mechanism at all. Again, the specific hypothesis raised from the model was thus not sustained in an experiment that continued for long enough. Of course, if the pattern originally observed was created by competition that no longer operates (Connel’s (1980) ‘ghost of competition past’), the model will continue to be rejected (an example of how evolutionary timescales and unique historical events are not amenable to logically justifiable, unambiguous interpretations in ecology). During the experiment on the more general hypothesis, if abundance of B increased in experimental areas because of increased recruitment of young rather than migration, the next round of refinement of the model can be quick and efficient. More detailed explanation is possible in the light of the result. Of course, it might always be better to assume that the general model ‘competition caused it’ is too general, and to propose a series of possible competitive models (competition on propagules, competition for pre-emption of resources, competition through interference between adults of the two species etc.). These could be used to make different predictions about the process leading to changes in numbers, and to its time-course and scale before the same sort of experimentation was started. Clarity of thought about the proposed model will always aid in avoiding this sort of logical error.

Confounding (ecologically, pseudoreplication)

Hurlbert (1984) has drawn attention to the serious problems of lack of replication in many types of ecological experiments (see also Underwood 1986). The class of problems that he defined as pseudoreplication has, in fact, long been well-known as confounding. The difficulty in experiments without appropriate levels of replication is that the experimenter believes that replicates (and associated degrees of freedom for statistical comparisons) are available when, in fact, they are not. As an example, consider the observation that some species of intertidal mussel is less abundant in some areas of the shore than others, and negatively correlated in distribution with that of a major predator. One useful model to account for this lack of correlation is that the predation rate is responsible (e.g. Paine 1974). One testable hypothesis from this model is that removal of the predators from experimental plots will lead to an increase in the abundance of mussels compared with control plots in which predators are naturally abundant. If the null hypothesis (there will be no difference in numbers of mussels in the experimental and control areas) is tested by removing predators from one experimental area and sampling the numbers of mussels in replicated quadrats within that and a nearby control area, the experiment is completely spatially confounded (or pseudoreplicated). There may be many differences between one area of a shore and another, regardless of the presence or absence of predators. The appropriate spatial scale for replication is that of the experimental plots; there need to be several experimental and control areas, each of which may be sub-sampled to estimate numbers. To avoid any differences between randomly-chosen areas of a shore being approximately equal in experimental and control sets of areas. If removal of predators has some effect consistent with the hypothesis, there will be a greater difference at the end of the experiment between the number of mussels averaged over the experimental plots and that averaged over the control plots than the sort of differences that would be expected by chance among the plots within either set. This sort of design requires a nested, or hierarchical, analysis which clearly takes care of the two levels of spatial replication (Groen & Hobson 1970; Underwood 1981).

The difficulty in ecological experiments seems to be a result of the fact that there is nothing wrong with the logical components of an experimental test to examine the null hypothesis that there is no difference between two areas, one experimental and one control. This null hypothesis can, if false, properly be rejected by having the confounded, pseudoreplicated design (i.e. replicated samples are taken in each of one control and one experimental area) and the data are amenable to appropriate statistical considerations: if these are not necessary (e.g. a t-test would suffice). As a result, it is possible to consult a statistician and be informed that such a design is valid, what, for the purposes of the ecologist, it is not. The problem lies not in the experiment as done, nor in the statistical procedures, nor in the conclusion that the null hypothesis (no difference in numbers of mussels in the two areas) should be rejected. The problem is in the logical relationship between discovering a difference and then stating that such (correct) rejection of the null hypothesis allows one to conclude that removal of predators caused the difference (as stated in the hypothesis) and thus to support the model (that predation is responsible for the observed pattern). Of course, many other possibilities exist — any innate difference between the two areas in their previous history, or that exists at the start of the experiment, or that occurs during the experiment, could have caused the observed rejection of the null hypothesis. It is completely irrational to conclude that the one preferred, or chosen difference (removal of predators) is responsible.

Because of lack of thought about the relationship between the hypothesis, the null hypothesis and the test, ecologists manage to come up with some considerable percentage of all experiments in such confounded fashion (see Hurlbert (1984) for numerous examples and a description of other types of spatial and temporal confounding).

'Natural' experiments: natural 'experiments'

A major problem confronts those ecologists who wish to invoke evolutionary explanations for a given set of ecological characteristics of organisms (see Peters 1976; Connell 1980). These problems have such long timescales that no realistic critical experiment is possible. No current test can eliminate alternative models about events that occurred millennia ago. Similarly, there are many types of investigation where ecological, rather than evolutionary, timescales are excessive and therefore rigorous experiments are not yet possible, or are prevented by prohibitive logistical constraints. As a result, it has sometimes been considered desirable to do what is called a natural experiment. In fact, these are sometimes recommended by less structured thinkers (Diamond 1986).
Consider the observation that two species tend to be found at different positions along a gradient (e.g., up a mountain, at different heights on a shore). One model to account for the difference is that the species compete and one (A) wins in one habitat (say the lower one) and the other (B) wins in the other locality (the higher one). One logically valid hypothesis derived from this model is that in areas where, say, Species A is absent, Species B should be able to extend its range along the gradient (be found farther down the mountain or shore), because the cause of its restricted range is absent. The null hypothesis is that the distribution of B will be the same in areas where A is naturally present or absent on the gradient. A natural experiment would therefore consist of finding places where A was naturally absent and examining the range of distribution of B, in comparison to the distribution of B in "control" areas where A was present (e.g., Brown 1971).

Obviously, this is not a logical procedure for determining the value of the proposed competitive model. It is simply a confirmatory or Aristotlean experiment (see earlier; Medawar 1969). Even if the null hypothesis of no difference in distribution of B in areas where A is present or absent are rejected, there are many possible differences between the two types of area examined that are only one of the absence of Species A (Connell 1980; Underwood 1986). In fact, the areas are always chosen because they are known to be different (some do not have A), despite Diamond's (1978) inability to conceive of differences between habitats. Nevertheless, there are no controls to determine that the absence of Species A is the only important difference. Such experiments invalid and can, at best, provide only necessary support for the model of competition, but cannot provide sufficient evidence.

The use of natural experiments, deliberately or by accident (Underwood 1986), is widespread in some areas of ecology. The usual way of describing the procedure is to use quotation marks around 'natural' to draw attention to the fact that there are problems with the study. This is an interesting example of the use of success grammar (grammatical tricks to alter the meaning of words) to divert attention from the real meaning of the words (Stove 1980). It would be better if these illogical procedures were abandoned. If not, they should be described as natural 'experiments' to make it clear that it is the experimental part of the procedure, and not its naturalism, that is suspect.

**Theory-laden observations**

One major criticism that has been levelled at falsificationist research programmes is that the observations from which they start are not free from some theoretical or paradigmatic framework (Hanson 1958; Koyré 1968; Popper 1968; Feyerabend 1975). Thus, previous training, experience, or current biases about what is important lead the experimenter to make the particular observations. In ecological research, this has two consequences. First, the generality of the observations is suspect — we tend to focus on particular phenomena that appear interesting, but may not be representative of the world at large (this is the basis for some of the just-so stories that permeate some areas of ecology and evolution). Also, if we persist in focusing on particular processes (such as competitive interactions) because that is what we think we keep observing, the richness of ecological models and explanations is lost. Choosing what to observe and where to observe it may also explain some of the differences of opinion about generalities in ecology (Underwood & Denley 1984; Underwood & Fairweather 1986).

Second, it is true that observations are only made in response to pre-existing theories, then the whole procedure outlined here is backwards and it would be profitable to try to identify these theories so that the processes of choice are also understood. It is not a compelling criticism of falsificationists that observations are functions of previous theory ('theory-laden'). All scientific enquiry shares this criticism, which may be devastating for ecologists who do not recognize it. We are all already in the framework outlined in Fig. 1. As a result, it is immensely easy to confuse observation with explanation in ways that cannot be unravelled. For example, observing that a species of gorilla is found only on certain mountain tops and describing this as the 'gorillas preferring to live there' implies not only that they are found there, but also that they find resources there they need, are safe from enemies there, and have behavioral mechanisms that allow them to express their preference to be there. Of course, gorillas may prefer to live on open lowland woodlands, but changes to habitat (probably induced by man) may have made such preferences impossible. It is, of course, possible that gorillas that do move to these preferred habitats after it has ceased to be a viable place to live will gradually be selected against and, at some point in the future, may evolve a preference for mountain-tops. This is not simply a semantic difference. Research into preservation of gorillas might advance faster if the features of the 'preferred' habitat were not the focus of attention, and the incidental features of other will gradually be explored. If, however, gorillas do prefer mountain-tops, the latter is irrelevant. This is an example of why the observation-statements should be free of explanatory models.

**Refutationists tests of patterns and observations**

The falsificationist experimental procedure outlined so far will aid in uncoupling observations from embedded explanatory models. Consider a simple case: Two ecologists walk along a rocky shore and, at that point, one observes that a certain species of small blue snail was only seen at the top of the shore; there was none of this species low down. If this observation were 'theory-free', it might still be wrong. If it were dictated by the ecologist's extensive knowledge of the relevant literature (which is extremely unlikely; most ecologists apparently (or at least) and therefore 'theory-laden', it might still be correct. What is needed is a mechanism to explore various models to explain why the observer claims to have made the observations.

One model (the one advanced by the observer) is that this pattern was seen because it is there; there is the case that small blue snails are found higher up on the shore and not low down. Alternative models are that the observer is deluded, that the snails lower down are camouflaged or cryptic or somehow different in appearance, or that the observer is lying or foolishly misrepresenting the truth (perhaps to propagate some hidden previous theory), and so on.

At this point, different hypotheses can be deduced from the models. The most useful one for progress in ecological (instead of, perhaps, psychological or legal) research is that, if the pattern were seen because there really were more snails higher up (this is the model), then, if carefully defined and appropriate sampling procedures are used equivalently at low and high levels on the shore, we should predict greater numbers of blue snails in samples from higher than from lower areas (taking into account taxonomic identifications, careful searching for cryptic specimens and delving into cracks and other shelters and ensuring that camouflaging is not deceiving, etc.) All the other models suggested above (including that these observations were made because of personal blinkers or other aspects of theory-laden selective behaviour by the observer) require that there be no systematic difference between the two heights on the shore in mean numbers. These latter hypotheses are, in fact, the null hypothesis for the chosen model. The test can then proceed. If the null hypothesis is rejected, the model can be supported; more blue snails are seen higher up because there are more higher up and we now know that the observer was not deluded, nor wilful, and that there was no problem with camouflage, crypsis and so on. (Note that this leads to the rejection of non-deluded, non-wilful observers; this suggests that the original observer should not necessarily be the one to do the test, but the issues of scientific morality and honesty are not the topic here.)

Thus, careful use of correctly identified models and hypotheses can lead to unbiased and well-supported results. If there is a problem with observations because they may not represent the world 'because they are theory-laden', this model (that the observations were made because of some aspect of theory-ladeness) can be examined. It can be tested as a contrast to the model that the observations were made because the world is really like that identified in the observations. Each leads to different hypotheses and a test can eliminate the incorrect one.

**Logical and statistical null hypotheses: A serious problem**

Unfortunately for falsificationist procedures in ecological research, there are many cases
where it has not proven possible to make falsifiable null hypothesis work for the hypotheses of interest. This is a serious problem that causes many interpretations from such tests to rest on illogical thought-processes (and to depend on the fallacious procedure outlined in the second part of Table 1). It is worth considering such cases in detail so that the types of hypotheses generated can be thought about very carefully. Models that are only supportable by fallacious logic may be compelling, but they may also be incorrect. There is, however, no easy way to prevent them from being widely elevated to the status of paradigms — even though they cannot be supported by sound logical procedures, although paradigms do not have to be logical — only widely believed (Faireweather, pers. comm.).

Logical hypothesis equals statistical null hypothesis

It is often the case that the only possible statistical null hypothesis, the only definition of a possible distribution of a test statistic (Mather 1967), is in identical terms to those used for the logically derived hypothesis (Hegh 1970). Whenever this occurs, the model and its derived hypothesis can be rejected (falsified) if wrong, but cannot receive logical support. This occurs because of the use of the hypothesis as the statistical null hypothesis. An example will make this clear (with reference to Fig. 3).

Consider the observation that animals are not scattered uniformly over an area of habitat. One model to explain this is that propagules of the species have arrived in some randomized way and the process results in random dispersal of the animals throughout the habitat in accordance with, say, a Poisson distribution. From this one could hypothesize that careful quantitative sampling with appropriately sized quadrats (forget for a minute the well-known problems of quadrat size and this test; e.g. Elliott 1977) will generate a frequency distribution of numbers per quadrat consistent with a Poisson distribution with a mean equal to that estimated from the sampled quadrats. There are several statistical procedures for examining such a proposition. One would be to fit the observed (sampled) distribution of numbers per quadrat to that expected under the hypothesis (usually using a Chi-squared test), but in that case the null hypothesis for the statistical test is identical to the hypothesis proposed from the model. The hypothesis is that the frequency distribution of numbers per quadrat will conform to a Poisson distribution. The null is that the data will conform to the particular Poisson distribution that has the mean estimated from that set of data. This null hypothesis predicts that, in a goodness-of-fit test, the value of Chi-squared will be less than the critical value chosen for the appropriate number of degrees of freedom (i.e. a Chi-squared value that falls in the range of values expected often by chance if the null hypothesis were true).

Note also that no alternative to the proposed hypothesis has been suggested and none is warranted so far because the observations are being explained by the particular model under test. Any attempt to construct a statistical test based on some other statement of the null distribution of the test statistic would require guessing what all other possible outcomes there might be were the Poisson distribution to be incorrect (i.e. all other distributions except that specified in the null hypothesis) and, for each of these infinite number of possibilities, attempting to calculate the distribution of an appropriate test statistic (Mather 1967). This, of course, is equal to trying to prove that the distribution is a Poisson one. The model proposed provides absolutely no help for determining what might be appropriate alternatives if the null hypothesis is wrong; the model simply predicts the null.

If the test were done and the null hypothesis were rejected, the conclusion must be (harking Type I error, see below) that the model was wrong, the organisms are not distributed at random according to a Poisson distribution and their pattern of over- or under-dispersion with respect to that sort of process can be added to the original observations and a new model invoked to explain both sets of data. This is a well-founded logical procedure; it conforms to the paradigm (Hegh 1970) shown in Table 1 and is rational. It is also the opposite pathway and conclusion to that used in the previous examples of the protocol where logical and statistical null hypotheses were different (compare Fig. 1 & 3). This is identified as Case 7 in Table 4, where all the possibilities are indicated. A false model leads to a false hypothesis and therefore a false null hypothesis (these being the same thing). Rejection of the null hypothesis leads to rejection of the model, rather than to its support as in the previous case (Fig. 1; Table 3).

As with the previous procedures (Table 3), there can still be problems with conclusions when the null hypothesis is falsified. First, there may be statistical Type I error (Case 1 in Table 4), leading to erroneous rejection of the null. Or there can be an incorrect hypothesis (and therefore an incorrect null hypothesis) derived illogically from a correct model (as in Case 3 in Table 4). Finally, there can be a combination of faulty logic and Type I error (Case 5 in Table 4) leading to a correct rejection of a
false model. These last two situations were discussed for the previous protocol and the comments apply equally here.

**Corroboration and paradigms**

The situation when a statistical null hypothesis is equal to the logical hypothesis, and is not rejected, is more complex. Under these circumstances the model is supported because its hypothesis has been corroborated. Yet this decision rests on no logically firm foundation (Table 1; Hocutt 1979). There are the possible problems of Type II error (Cases 4 and 8 in Table 4) and the previously identified possible problems of incorrectly deducing a hypothesis from the model (Cases 4 and 6 in Table 4), all of which were discussed for the previous protocol.

In addition, there is the problem of correct corroboration (Case 2 in Table 4), where the decision is made to accept the truth of the model because its predictions are apparently borne out (the null hypothesis is not rejected). This conclusion is based on the fallacy of affirming the consequent (described in Table 1). It occurs when a statement (p) implies a consequent (q), the consequent occurs and it is concluded that p must be true. This fallacy has been well-known for centuries (Lemmon 1971; Hocutt 1979). As a result of the fallacy, corroboration of a model even when it is correct, is based on no logically sound procedure (Casswell 1976). No falsificationist test has been done.

By this sort of reasoning, powerful paradigms (i.e. generally applicable and widely believed models; Kuhn 1970) are commonly and often ‘corroborated’. Experimental tests of models for which derived hypotheses simultaneously serve as statistical null hypotheses are very similar to ‘Aristotelian’ experiments (Medawar 1969; see earlier). The chance of illogically erroneous corroboration is great but there has been no critical test. Thus, the dominance of prevailing ideas tends to blind the experimenter to any alternative explanations of the observations and the lack of disproof en route to a conclusion provides apparent support for the prevailing paradigm. Ideas that are already fashionable and well-established (Kuhn 1970; Lakatos 1974) become very general. New examples are found when predications from the model are supported by experimental tests in novel situations.

There is also the problem that prevailing generalizations or paradigms prevent novel and contradictory observations because of theorety-ladenness in the acquisition of observations or data about patterns in previously unexamined systems. Observations in new systems are then readily interpreted as conforming to the predictions from existing generalizations when no falsificationist procedures have been used and no alternative explanations have been sought (see Underwood & Denley 1984 for examples in marine habitats). The use of logically sound falsificationists tests of models about the reality of observations would greatly reduce this problem (see earlier discussion of this topic).

**Power of statistical tests**

The other enormous problem with this type of experiment, where the logical hypothesis is identical to the statistical null hypothesis, is that no estimate can be made of the statistical experiment (the capacity of the experiment to reject a null hypothesis when it is, in fact, incorrect). Because no quantitative alternative situation to the null hypothesis can be specified, the probability of Type II error (see Fig. 3, Table 2 and the earlier discussion of power) cannot be calculated. This serious difficulty lies in the fact that the null hypothesis is the logical hypothesis. There is therefore no theory, model or logical guide to what its alternative might be. Where hypotheses are deduced from models to have the following form: ‘frequencies of phenotypes will be in the ratio of 1 recessive homozygote to 2 heterozygotes to 1 dominant homozygote’, or ‘animals are distributed as a Poisson distribution with a mean of 3’ or ‘the mean distance moved per minute will be 8 m’ (as possible examples from Mendelian population genetics, quantitative descriptions of an ecological pattern in space and, perhaps, optimal foraging theory, respectively), their quantitative alternatives are unspecifiable (or are infinitely various). Neither way of considering them (no alternative or an infinite number of alternatives) allows calculation of the power of a test, because there is no (or an infinite number of) frequency distribution(s) of any test statistic when the null hypothesis is false (Mather 1967).

This is very marked contrast to the situation where logical hypotheses are inverted to form statistical null hypotheses (see earlier and Fig. 1; Table 3). In those cases, when the hypothesis predicts, for example, that predators will remove some number of prey, the null specifies that there will be no difference in mean number of prey in areas with and without predators, or that more prey will be found where predators are present. Although the precise quantitative alternative to this null has not been specified, the hypothesis at least indicates the direction of the possible difference between the areas. The model and associated observations about the natural history of the system will usually provide a good basis for determining the magnitude of the effect of predators on abundance of prey as a quantitative alternative to the null if predators are, in fact, relevant (as specified by the model).

Where the set of observations being explained is itself quantitative, a very precise alternative is specifiable and can be calculated from the model. Consider the following example. You have observed, on average, twice as many of species X in areas where there are predators as in areas where predators are absent. You propose the model that predation causes this pattern, from which you deduce the hypothesis that removal of predators will cause the abundance of Species X to double. This is a quantitative alternative to the null hypothesis that removal of predators will make no difference. These sorts of examples demonstrate the precision with which an alternative to the null hypothesis can be calculated and therefore used to estimate the power of the test or to design an experiment of suitably large power.

Where the statistical null hypothesis is identical to the hypothesis derived from the model, no alternative can be considered and, in fact, no alternative becomes necessary until the null has been disproven and the hypothesis and model have been rejected. Situations where Type II errors of unknown probability cause failure to reject null hypotheses add to the potential expansion of generalities and paradigms in ecology.

**Conclusions**

There are several consequences to ecological investigations when the components of experimental protocols and their relationships and interpretations are not clearly identified and explained. Examples of these are considered below for different parts of the procedures.

**Recognition and differentiation of components**

It is crucial to be able to recognize the separate steps of the experimental protocol before starting on an experiment. For example, considerable heat, and not a lot of light, has been generated in discussions of the role of competition in the structure of bird 'communities' (Gilpin & Diamond 1984; Connor & Simberloff 1984). The problem in this altercation is that one side of the debate was discussing whether or not there are observations which might need to be explained by processes of competition (Connor & Simberloff 1986) and the other side was attempting to demonstrate that competition is necessary to explain the observations already available. Recognition of the different steps in the procedure and that each has different properties would help a lot to make such debate more progressive. This is a point well made by Strong (1980). Once both sides agreed to debate the existence of observations about patterns in the distributions of species of birds on islands, the discussion could at least progress. There is little point to the construction of models (i.e. explanations of observations) if there are no clearly identifiable and agreed observations which need to be explained. Other ecological examples of debate about the definition of the problem may be found in Hubbell (1981) and Murray (1982).

It is also important for interpretation of tests about ecological patterns and processes to recognize that the model is not the same as the hypothesis. Thus, the intermediate Disturbance Model (Connell 1978) is an explanation for the observed differences in numbers of species in different habitats. The model explains that the number of species is small in areas where there are severe, large or frequent disturbances which eliminate species. In areas
where disturbances are small and/or rare, populations are not thinned out and competition for limiting resources eliminates inferior competitive species. More species are therefore found where there are intermediate levels of disturbance. Note that the model is about processes; the pattern of relationship between the number of species and the regime of disturbance is an observation. Discovery that such a pattern exists in some habitat requires an explanation (which might be the Intermediate Disturbance Model or some other explanatory model). The model’s hypotheses should be cast in terms of the processes operating where disturbances are not intermediate (see Sousa (1979) for experimental analyses of this). Alternatively, attempts to generalize the model to new habitats require that it predict the relationship of a greater number of species at intermediate levels of disturbance (McGinnness 1984). If the model applies to some new habitat it is necessary that the predicted pattern be found. Discovery of the pattern is, however, insufficient to demonstrate the truth of the model as an explanation of the new system, that requires experimental tests about the processes the hypotheses about competitive interactions and the role of density-independent mortality eliminating species at the two ends of the gradient of disturbance.

Finally, there are cases where models have been proposed that cannot actually explain the observations. For example, where two species (A and B) are negatively correlated in abundance from place to place, the observations can be explained by competitive interactions between the two species. This is often all that is stated, but on its own, cannot explain why either species does not dominate in all places. The model has to include phraseology that A wins the competitive interactions in some parts of the joint habitat and B wins in others. Otherwise, the model cannot account for some places having large numbers of A and small numbers of B, and other places having more B than A (without which there would not be a distribution between the two densities, which was the starting observation).

Relationships among components

Recognition that there are important relationships among the separate components of these protocols may help to prevent some of the erroneous conclusions that are identified in Tables 3 and 4. Several of these are straightforward. For example, lack of appropriate controls in experimental manipulations introduces the possibility that some artefact of the experimental procedure explains why the null hypothesis is rejected. That artefact confounds the process being examined in the hypothesis. Examples of situations in experimental transplants where the behaviour of intertidal animals is being investigated may be found in Chapman (1986) and Underwood (1988). Other examples are in Dayton and Oliver (1980), Hurlbert (1984) and Underwood (1986). More careful thought about appropriate and necessary controls will help remove other processes that lead to the same outcome as the hypothesis under investigation (see the early discussion by Mill (1965, 1961)).

Consideration of the appropriate sorts of spatial and temporal replication will also assist in reducing potentially confounding alternative processes causing the corroboration of models. A good discussion of this is Hurlbert (1984) for many types of field experiments. Any sort of confounding built into experimental and sampling designs will lead inevitably to erroneous and illogical conclusions from the experimental results.

Another situation where the pattern of recognition of the relationships among components in the protocol would aid in making conclusions more logical is where the observations and models are constructed with respect to some specified spatial scale, but the hypotheses and tests are at some very different scale. Thus, observations about the distribution and abundance of fish on small patches of habitat on the Great Barrier Reef can be explained by models about processes influencing the colonization and subsequent survival of the fish (Sale & Dybdal 1975). Tests of these models at spatial scales of hundreds of metres or kilometres will not necessarily be relevant (Gradlhofer et al. 1980). The observations are still at some smaller scale and only by some inductive process (or perhaps simply by arm-waving) is it possible to state that results from one scale will automatically apply at larger or smaller scales (Allen & Starr 1982). The model may be quite different from those specified in an explanatory model may be operating at other spatial scales. Tests of a model at one scale should be at that scale (which makes it logically much more difficult to test models that operate at very large spatial and temporal scales).

Use of these protocols

Use of these protocols in many situations where they are not explicitly identified (or identifiable) at present would advance our understanding of marine ecological processes. For example, argumentation about the validity, generality or other aspect of observations and patterns could be obviated if there were general recognition that many observation statements contain embedded explanatory editorialization. Many observations are made casually or are not well-documented (see also Andrew & Mapstone 1987). It is assumed that the observations are ‘real’, in the sense that the observer saw those things because they existed as patterns. That assumption is one of a possible set of different models about why the observations are made (one is satisfied). To separate the description of observation from the model to explain it is fairly straightforward. It is also often the case that quite straightforward sampling under carefully and a priori defined conditions will corroborate the model (the observations were made because the pattern really exists) by a sound falsificationist procedure.

These protocols, for similar reasons, should also be used whenever a general model is about to be invoked in a new situation (to explain the distribution of some other species or assemblage of species, or to explain patterns in some different habitat or geographical region). A cautionary tale can be read in Connell’s (1961, 1970) accounts of the vertical distribution of two sets, each of two species, of barnacles in Scotland and the United States. In each case, there was a similar pattern of overlapping distribution of the two species in the two places. The processes that caused the patterns were different. The same patterns (observations) can be caused by very different ecological processes in different parts of the world. Every new case will require some tests of explanatory models before their relevance and validity can be assumed (see also Underwood & Denley 1984; Underwood & Fairweather 1986).

Wherever a model has been corroborated, the very problems inherent in these procedures make it mandatory that the model be subjected to more rigorous scrutiny than was previously done with more specifically constructed hypotheses. Complacency and acceptance of models because they have worked so far does not guarantee that they are correct. Any criticism of a model, or the evidence that has been used to support it, should be met by renewed experimentation, not by argument about the legitimacy of the criticism (or the critic), the role of criticism in science, or attempts to prop up the model by ancillary and possibly irrelevant observations. When a model is criticized, or its experimental basis is suspect, or some alternative model has been proposed, the onus is upon the proponents of all sides of the debate to deduce new and, where possible, contrasting predictions from their versions of the explanation for the original observations. These can then form the basis for renewed testing. It is a mockery to examine some model critically by experimental procedures and then cease to continue to examine it, simply because it has been ‘proven’ once and for all. The whole notion of experimental evaluation requires that models are never assumed to be true — they are assumed to be true so far, because they have not yet been falsified (Popper 1968, 1969).

Where possible, situations where no falsificationist testing is practicable should be avoided (Table 4. Fig. 3 and earlier discussion). In the case of goodness-of-fit tests on Poisson distributions, for example, it may be possible to construct other hypotheses that do not lead directly to the null hypothesis being also the statistical null hypotheses. Thus, if a non-uniform pattern is explained as being due to randomizing Poisson processes acting on the pattern of distribution of the early stages of the life-history, hypotheses can be constructed such as: if the propagules are experimentally distributed in some uniform or clumped manner, their ultimate distribution will be different from (i.e., less or more variant than) that in natural control areas. The logical null hypothesis has now become that there will be no difference between control (random) and experimental (non-random) areas. Any non-random or density-dependent process acting on the organisms after their early random arrival can then resurface itself in the experimental area, making
the distributions there similar to controls (which are subject to the same processes). Now the logical and statistical null hypotheses are different.

Experimental and so-called applied ecology

Experimental ecology is sometimes considered to be the preserve of "academic" ecologists and has little to no place in the "real world" of management of natural resources, conservation of rare species or environmental issues (Underwood 1983b, 1983a). This detracts from a more useful focus on whether the work is good or bad science, and whether it was intended to be applied. There is, in fact, every reason why the protocols discussed here should form the centerpiece in any attempt at management of ecological issues. Any statement about management of habitats or pollutants, or any prediction about harvesting or conservation of populations can easily be seen to be a predictive hypothesis, as in Figs 1 and 3. Assessments of potential environmental impacts are as yet untested hypotheses about what will happen to the particular species in the particular habitat when the impact occurs. Management of fisheries or other populations requires stating what will be the possible consequences of allowing or forbidding certain components, or numbers, of populations to be harvested, or of allowing harvesting in some areas or at some times, but not in (or at) others.

Usually, these statements (hypotheses) are not explained against a background of the sort of models that are their basis. Sometimes the models are explicit, but not clearly articulated. Often, the models are implicit. It would be an important advance for our understanding of ecological processes if all statements about what will happen to populations or habitats under some recommended course of future action could be presented in relation to the components of the experimental protocols discussed here. Thus, the observations from which these predictions are made should be identifiable. The model(s) to account for these observations and which are the basis for making the predictions should be specified explicitly and conclusively. The relationship between these two components can then subject to critical appraisal before the predictions get out of hand. The hypotheses will clearly be based on the models. Alternative scenarios can be derived from different models about the system and their consequences considered. All of this requires no change in the time and energy currently used to gather information and to construct practices for management or regulation of populations and habitats. It does, however, require more effort in the relationships between different parts of the procedure, which will require more thought than has often been the case.

If this were done, there would be a clearer understanding of the relationship between the different problems of ecology and regulation of populations and habitats. Of course, there would also be the realization that all intrusive management or environmental interventions is a form of experiment. Construction of well-argued models and hypotheses leading to these experiments would also make it more likely that we could build in some controls, some evaluation after the practice of management (or development) has occurred (Hilborn & Walters 1981, Walters 1986) and, perhaps, use these situations to learn about spatial and temporal scales beyond the budgets of most "academic" ecologists. We could also, without doubt, learn about the processes operating in these environmental communities and those operating in the minds of environmental and resource managers which led to the particular models being employed. Finally, beyond any shadow of doubt, we could learn from our previous mistakes, because the experimental perturbations of the environment would be in large part those that are most likely to come out could be used to evaluate the models that underpin the policies and practices.

To this end, it will benefit all parts of ecology to be more sure and careful about the different components of experimental procedures and to use every possible opportunity to avoid the pitfalls of logic that are so abundant in these protocols. Searching for help from philosophers and logicians may (Simberloff 1980; LeUse 1987) or may not (Edson et al. 1981; Roughgarden 1983) aid us in these tasks. Nevertheless, critical examination of our own problems, in the context of many variables complex ecological systems, is increasingly necessary. Having an explicit framework in which the structure, advantages, disadvantages and needs for clarity of thought are identifiable should lead to increased understanding of the logical basis of ecological experimentation.

Acknowledgements

Financial support was provided by the Australian Research Council and the Institute of Marine Ecology and the Research Grant of the University of Sydney, and a Distinguished Research Fellowship at the Bodega Marine Laboratory of the University of California, Davis. Throughout several years, the preparation of this paper benefited from discussions with and insights of Professors D. T. Anderson, L. C. Birch, J. H. Connell, C. H. Peterson and D. S. Simberloff, Drs A. J. Butler, E. F. Conner, P. G. Fairweather, R. T. Hinde, K. A. McGuinness, P. J. Myrschoug, P. S. Perritais, W. P. Sousa, P. A. Underwood, Ms M. G. Chapman and many friends and students. A. J. Butler, M. C. Chapman, J. H. Connell, P. G. Fairweather and D. Simberloff made helpful suggestions to improve the manuscript.

References

Connell J. H. (1961) Effects of competition, predation by Thys is lapilinus and other factors on natural popu-

(Received manuscript accepted April 1990)