Ann. Zool. Fennici 46: 93–111 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 30 April 2009

Components of design in ecological field experiments

Antony J. Underwood

Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories A11, University of Sydney, NSW 2006, Australia

Received 18 Feb. 2008, revised version received 8 Aug. 2008, accepted 27 Aug. 2008

Underwood, A. J. 2009: Components of design in ecological field experiments. — *Ann. Zool. Fennici* 46: 93–111.

Over the last 50 years, ecological experiments under field conditions have exploded in number, type and scope. They remain complex because of intrinsic variability in ecological measures from place to place and time to time, requiring care in their design and implementation. An experiment and its design can only be sensibly considered after thought and knowledge are used to make clear the logical basis for doing the experiment, so that its results can be interpreted in a robust and reliable manner. There are different approaches to any sequence of components of an experiment. Here, a falsificationist methodology is considered, which relates observations (things we know) to models (what we think explain the observations) to hypotheses (predictions about what will happen under novel circumstances if the model(s) is (are) correct). Experiments are then designed to create the novel circumstances in order to test the predictions. How an explicit framework influences the design of experiments is discussed, including the nature of replication and of controls for artefacts. Improving the match between natural historical and ecological knowledge and the interpretation of results of experiments will always help advance the discipline of ecology.

Introduction

Background

Ecological experiments, particularly manipulative experiments, have, over the last 50 years, revolutionized the ways in which ecological understanding has advanced. It is, however, still the case that many experiments are poorly designed, sampled, analysed or interpreted. It therefore seems appropriate to consider, yet again, some of the fundamental issues about the nature of experiments, in the hope that this will help with their planning, execution and successful outcomes.

For this exercise, an experiment is any test of a logically structured hypothesis or prediction (Ford 2009). The distinction between manipulative and mensurative experiments (*see*, notably, Hurlbert 1984) is irrelevant to this. The former may often be better methods to test the reality of ecological theories, because they involve the direct, controlled alteration of processes to determine whether predicted responses occur (*see* Connell 1974, Hairston 1989, Underwood 1997). Mensurative experiments, in contrast, rely on sampling existing sets of conditions that differ with respect to the processes about which predictions are being made. These are, however, perfectly appropriate for some situations and

are the only possible option where large spatial scales, long time-courses, excess expense, logistical impossibility or ethical considerations make manipulations impossible.

It is also notable that not all ecologists — and possibly not even a majority of ecologists — are supporters of the nature of experiments, or, indeed, the need for experiments, or the logical justification for experiments. Those who prefer to describe or model nature, who find satisfaction in *post hoc* data-mining or reconstructive interpretation and those who use other methods are inevitably entitled to their views. There is no space here to argue the pros and cons of different philosophical approaches to the science of ecology.

There is, however, an issue that cannot be avoided. When ecological theories have been put to the test by direct intervention (or manipulation) or by structured comparisons, on many occasions these theories have been demonstrated to be wrong or to be inadequate. Recognizing this requires identifying that the experiments were not so flawed that the results should not be accepted. That is the reason for this paper — trying to improve understanding of the nature, logic, and requirements of experiments so that their validity can be identified (or so that their deficiencies can be recognized).

There will always be situations where experiments are inappropriate, impractical or impossible. One good and obvious set of circumstances is the field of macro-ecology, which is concerned with large-scale patterns in diversity or other ecological characteristics (Brown 1999, Gaston & Blackburn 1999). Another is processes of natural selection which generally would require enormously long time-courses for any valid experimental test of influences on populations or speciation. These situations are often investigated by smaller scale or shorter-term tests that can successfully demonstrate outcomes consistent with the proposed processes, but cannot unambiguously identify that a proposed process is really operating.

All such circumstances should benefit from recognizing what ecologists can most learn from experiments — that many ideas are shown to be wrong when tested (so, many interesting, but untestable or as-yet-untested ideas are probably also wrong) and that it is easy to be misled about the data and the patterns generated by observation and sampling.

This contribution was requested to be focussed on field experiments. There is, of course, no difference in the attention necessary for experimental design in the field *versus* the laboratory (Campbell *et al*. 2009). Field experiments have numerous advantages in terms of valid tests of hypotheses — notably hypotheses can be tested under much more realistic conditions of weather, predation, recruitment of other relevant species, etc., etc. For many ecological theories, tests under laboratory conditions are, at most, pointless and, at best, misleading. Nevertheless, particular attention is often necessary in field experiments precisely because many conditions are not regulated or controlled (*see* Connell 1974).

Logical components of experiments

There have been many accounts of the ways experiments are organized (for example, Medawar 1969, Heath 1970, Hairston 1989, Underwood, 1990, 1997, Peckarsky 1998 among many others).

It is notable that most of these pay considerable attention to the nature of experiments and their design, with some discussion of the types of hypotheses being tested (which are often described as questions being answered). Some accounts, e.g. Diamond (1986), went to great lengths to describe all sorts of studies as having value, but concluded that ecologists should not be blind "to the possible value […] of some manipulations unconstrained by prior hypotheses". This sort of consideration completely begs the question of what the manipulation was for and how it might be interpreted. Bacon's (1620) methodology is clearly alive and well, despite years of being criticised (e.g. by Medawar 1969). Yet, its modern advocates, such as Diamond (1986), never explain systematically what guided the choices of things to manipulate, how much and in what ways they should be manipulated and what variables should be measured, over what temporal and spatial scales. Usually, despite the claims of practitioners, Baconian tests involving giving some part of the world a shove and

recording the outcomes are *not* done freely of hypotheses. In fact, there are usually quite clear hypotheses guiding the nature, design, analysis and interpretation of such manipulations. It is, however, the case that they are not explicit and magically appear when the results are available. It is very dubious that most ecologists will attempt to manipulate nature without *any* prior ideas about the processes operating and why they should be manipulated in particular ways. It is, however, inevitable that such experimental procedures will mostly have flaws in design, execution, analysis or interpretations because there is no structure (i.e. logical thought) to which they must adhere.

Preconceptions which are not made explicit (in the context discussed here, made explicit as possible models) are most often (probably always) dangerous for making sound conclusions. Any experimental data that are biased or influenced by unrecognized prior expectations cannot allow reliable interpretations. Assuming that certain processes do or do not operate may simplify any study, but the consequences of such assumptions for any conclusions must be made clear. For example, if it assumed that currently observed patterns of distribution of a species are regulated by currently operating processes, all sorts of confusion may occur. There may be significant (statistically and biologically) important effects of processes such as predation or competition. These may, indeed, be important processes *maintaining* a current distribution or abundance of some species. This is, however, not the same as the processes that *caused* the current pattern in the first place. Preconceptions that only current processes are important will preclude designing experiments that actually discriminate amongst a set of perfectly sensible possible reasons for the observed patterns — because many of these processes will never be considered.

Obviously, not all authors agree about the necessity or nature of experiments. It does, nevertheless, still seem very appropriate to try to consider the components of experimentation, in an attempt to identify what they are and how they are related to one another. Without this, I see no point in trying to discuss issues of experimental design, because the design has to be made right to retain the logical structures that necessitate the experiment and its interpretation. Here the following framework is used (and described in detail, with examples in Underwood 1990, 1991). Like everything in ecology, any attempt to define components of an experiment (or even to see the need for experiments) provokes critical comment. This is probably healthy because it promotes thought and discussion. Rather than attempting to deal with the criticisms in the limited space available here, it would be better for readers to examine the alternative methodologies being proposed (although some critics of experiments never seem to produce a detailed account of a preferred methodology). Then, readers can make up their own minds about how to proceed. Some core issues about experimental design transcend most methods and are as relevant to describers of nature and modellers as they are to experimentalists.

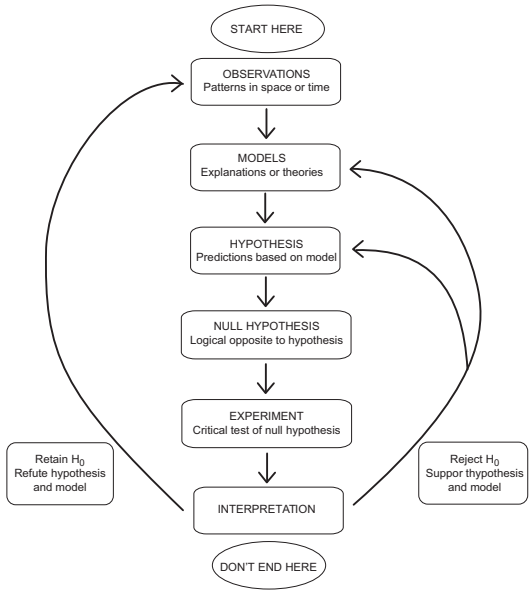

Fig. 1. A framework for components of ecological field experiments. Note that experiments are at the end of a chain of thought processes and that there are no exits. *See* text for detailed explanation (from Underwood 1997).

Consider a framework in which an experiment is done (Fig. 1); more complete accounts are in Underwood (1990, 1991, 1997) and considered by Peckarsky (1998).

In this framework, a research project starts with observations — information about nature, ecological patterns in space or time (*see* Andrew

& Mapstone 1987 for identification of patterns). These are the substance of the science and the research is to determine why these observations have been made (and, indeed, whether they are correct). So, some philosophers have called the initial observations a puzzle (Wittgenstein 1921) or a problem (Popper 1982) which the researcher has decided to attempt to solve.

One major objection to research being motivated by observations is that they may be incorrect because of biases on the part of the observer, because no-one observes the world without being influenced by prior knowledge (so that the observations are "theory-laden"; Chalmers 1979). How to deal with this issue has been discussed in detail elsewhere (Underwood 1990, 1997).

The obvious next step (Fig. 1) is to propose explanations of why these observations or patterns exist. These explanations or theories or models (for other terms *see* Trusted 1979) are usually about processes that could account for the observations. Many authors call these explanations "hypotheses", which is no problem as long as the term does not get confused with the next step (hypotheses or predictions). Such models can be simple or complex (*see* particularly Nagel 1961). The only requirement is that the models should provide a potentially realistic explanation of the observations. The models proposed are not considered to be true or false for the particular observations under consideration (which is why there need to be experiments) — only plausibly realistic.

As an example, suppose it has been described that there are more beetles of a particular species per 100 cm^2 of surface of the ground under the canopy of trees than in the open, where there is no canopy. Plausible models to explain this include: the habitat under a canopy is less harsh, allowing greater densities; or the beetles are more vulnerable to predators foraging in the open; or there is more food under a canopy; etc., etc. There can be many possible explanations, including interactive combinations of any of these three (and any others). Your knowledge of the relevant literature (in passing, the literature about distributions of organisms, not just beetles!) will provide you with lots of explanatory models that have been proposed by other ecologists (*see* the early, but reprinted paper by Chamberlin (1965) for a good discussion of alternative models). Proposing a model does not demonstrate that it is the appropriate explanation (even if it seems to be the only possibility). Often, proposed models turn out to be incorrect or, at best, very poorly supported by subsequent evidence (examples from marine ecology were reviewed by Underwood & Denley 1984). So, experimental procedures are needed to eliminate (falsify) models that are incorrect, as explained below.

Therefore, instead of formulating the prediction as an hypothesis to be corroborated, we create a null hypothesis, consisting of all possible logical alternatives to the hypothesis. If the null hypothesis is disproven (falsified) by the experimental data, the only alternative is the hypothesis, so the experiment supports the hypothesis by falsifying its alternatives. As the other possibility, the experiment generates data that conform to the predictions in the null hypothesis — as opposed to what was predicted by the hypothesis. Consequently, the hypothesis is falsified by the experiment. Either way, in principle, one or other of the only two alternatives (the hypothesis *versus* the null hypothesis of all other possibilities) will be falsified.

So, in this framework, when (and only when) the hypothesis (hypotheses) have been carefully constructed (Ford 2009) is it sensible to begin to design (to plan) the experiment (or experiments if more than one is appropriate). If a logically structured hypothesis(es) is (are) clear, the experiment(s) can be designed to be interpretable. This requires, as considered below, care in ensuring that the outcome — the results — of the experiment can be interpreted carefully in relation to its purpose, which is to test the explicit hypothesis.

Falsification, rather than attempting to confirm models, is a necessity because of logic. Formal logical structures to support this statement have been described many times over many years (going back to Hume 1779, 1941, Russell 1912, Popper 1968, 1969). The argument, in a simple form, is that corroborating or confirmatory evidence is never complete. It does not matter how many new observations are made, they are never a complete set. They are the observations made from sampling and experiments and are limited in scope by resources,

time, and logistics — the sheer impossibility of being everywhere, simultaneously to observe all possible cases of some phenomenon. Proof or corroboration of an hypothesis requires inductive reasoning. However many confirmatory observations are made (Cannap 1962, Hempel 1965), does not validate the inductive conclusion that a statement has been proven (notably *see* Russell 1912).

Attempting to corroborate an hypothesis is an attempt to prove that it is correct. The logical structure of this notion is called by logicians "affirming the consequent" (e.g. Lemmon 1971, Hocutt 1979). If some hypothesis p predicts some particular data or observations, q and, in an experiment to test the prediction q occurs, the temptation is then to conclude that p must be correct. Formally, ((p implies q) and q) implies p. In fact, all that has happened is that q has occurred, in association with or caused by all sorts of possibilities, which include p, but p has *not* been demonstrated to be the only possibility.

In contrast, disproof of an hypothesis is much more safe in logic. If p predicts q, but q does not occur in the experimental test, p cannot have been happening. Formally, ((p implies q) and not-q) implies not-p.

Why is it important to keep observations, models and hypotheses separate and explicit?

Observations and models

It is extremely important to keep observations, models and hypotheses separate and clear. Observations are things that have been seen or are known (even if they are confused, incorrect or atypical). Models are derived from knowledge, past experience, inductive reasoning or, sometimes, inspired guesswork to explain why the observations were made. As elements of any study, they are not particular to scientific endeavours — there are many interesting and useful models in history, economics, origins of grammar, etc. Hypotheses, in contrast, are logically derived, by deduction, from models. They are predictions that can only be correct if the models from which they are derived are correct.

So, what happens if observations and models become confused? A simple, but realistic, example will illustrate the difficulties. Off the east coast of Australia lies Lord Howe Island, a volcanic island with a tropical coral reef and lagoon on its western side and temperate subtidal kelp-bed on its eastern coast. On it lives the endemic rail, the Lord Howe Island woodhen.

The relevant information summarized here was described in full by Caughley and Gunn (1996). Woodhens, *Oncydromus sylvestris*, were numerous over the island, but, by the 1970s were confined to the southernmost mountainous areas of the island and there were only 20–25 birds in the population. By 1980, there were only 3 to 6 breeding pairs of birds in territories on the mountain. Much research ensued, based on trying to determine what was going wrong with the areas occupied by the birds, interpreted to be their preferred habitat. None of this revealed any reasons for the demise of the birds. Only when the observations about where they lived were separated from the model that they live there as a result of some preference did alternatives get explored. Miller, an ecologist, proposed that feral pigs, which were widespread over the lower parts of the island were preventing woodhens from breeding or surviving outside their territories (Miller & Mullette 1985). From this model, he hypothesized that removal of pigs would lead to increased numbers and an expanded occupation of the island. All but one pig were killed and woodhens immediately began to increase in numbers. At present, with help from a captive breeding programme, the birds are numerous over most of the island.

Clearly, the problems of trying to resolve how best to manage the habitat so that the woodhen's population could be conserved in a viable state were made much more difficult by the notion that the birds were in a restricted part of the island because that was their preferred habitat. By recognizing that the observation (the birds were only in a restricted area of the island) was being explained by a model (the birds were only in that part of the island because it represented the appropriate habitat), it is much more straightforward to identify other possible models (in this case, the birds were restricted to that part of the island by inimical features of the rest of

the island). If evidence then rejected hypotheses derived from the first of these, the second could be elaborated to consider components of the environment in the rest of the island that were causing birds to stay where they were observed. Thus, shortage of resources, predators, diseases, extreme harshness of climate, etc., etc. would form the basis of alternative models about the negative nature of the rest of the island.

Note the vast difference between:

i. the birds are only found in habitat A,
ii. the birds prefer to be in habitat A.

The former is an observation. The second is an observation (about where the birds are) *and* a model (a statement about why they are where they are observed).

This confusion of observations and models is not uncommon, particularly where ecological patterns are being described (*see* also discussion in Underwood 1990). Thus, a casual, unquantified observation that more of some species are found in the open than under trees (or in rockpools, or …) is an observation. If not considered formally, such an observation can be taken to be an accurate statement about nature. If taken to be a valid description of nature, it is assumed that alternative models to explain the observation have been eliminated by falsifying their predictions. For this example (more species in the open), one obvious alternative model is that there are, in fact, equal numbers of the species in the open and under trees, but they are easier to see because there is more light or they are less camouflaged when in the open or they move about more in the open, or other possibilities. For the original observation to be considered appropriate or correct, it must, at least, be demonstrable that these types of alternatives are wrong. They can all be expressed in a short-hand way by stating that the original observations are wrong and, in fact, there are equal numbers of animals in the open and under trees. There might even be fewer in the open.

So, the first model leads to the prediction that careful sampling that is not confused by light, camouflage, movement, etc., will reveal greater numbers in the open. The second model leads to the contrary prediction that such sampling will reveal similar numbers in the two habitats (or more under the trees). These hypotheses can then be contrasted by the appropriate experiment, i.e. by doing the careful sampling and examining the outcome. Both hypotheses cannot be correct; the latter forms the basis for a statistical null hypothesis (there is no difference in numbers between the habitats).

This statement is chosen as the null hypothesis because this particular statement defines to be zero the amount of difference that, if it were true, should occur between the numbers of animals in the two habitats. Thus, it defines the value (zero) for the parameter of average differences between the two sorts of samples (those in the open *versus* those in the trees). In essence, except for sampling error, there should be zero difference between the mean numbers of animals in the two samples. The other two possibilities (more in the open or more under the trees) state whether you expect the difference between the two samples to be non-zero in a particular direction. Neither of these statements actually defines how big the difference should be. For construction of any statistical test of the hypothesis being tested, it is necessary to define parameters of any statistical distribution that will be used. So, the distribution of a test statistic can be defined where the difference between the two samples should (except for sampling error) be zero. In other words, the parameter of difference is zero. The statement that defines the relevant parameter(s) is declared to be the null hypothesis.

If the null hypothesis is rejected, there is support for the model that more were seen in the open because there really were more in the open. If the null hypothesis is not rejected, there is no evidence to support the model that there were more animals in the open. Misinterpretation of either outcome can occur because of Type I and Type II errors (*see* later). Despite this, only when the predictions have been tested does it become sensible to interpret observations as being realistic reflections of the distribution of the animals.

Explicit hypotheses

Another reason why it is important to make hypotheses as explicit as possible is because any confusion in hypotheses can lead to confu-

sion in analyses and interpretation of data and, ultimately, any conclusions reached about the models being examined, i.e. the ecological processes thought to be operating.

As a specific example of how confusion is likely unless everything is very clear, consider experiments involving transplantation of organisms to test hypotheses about processes influencing the growth and sizes of individuals. These have been discussed in detail elsewhere (Underwood *et al*. 2004).

The original observations are that mature individuals of a species of grass are taller in one habitat (A), near to the bottoms of hills than are individuals in another habitat (B), further up the hills. Of several possibilities, consider two simple, but general models that could explain this difference. Both include the notion that plants reach maturity and cease to grow taller at about the same age in the two habitats (to keep the models under consideration for this example as simple as possible). So, their different sizes are the result of faster growth in habitat A. The two models are then:

i. the individuals in habitat A have intrinsic, genetically controlled faster growth than those in B — the difference is due to the plants being of two types, each associated with a different habitat; or
ii. the plants are intrinsically (genetically) similar and the difference in growth is entirely due to environmental influences (water, soil, nutrients, disturbance, diseases, etc.) which cause faster growth of individuals that happen to be in habitat A.

There are, of course, all sorts of other possibilities, including combinations of these two (for example, there are genetical *and* environmental influences operating which combine to create the different rates of growth). The point here is to keep the example simple.

From model (i), the hypothesis can be proposed that very young individuals from habitat A transplanted to habitat B will continue to grow as fast as similar juveniles left in A and will therefore reach taller mature sizes (like those in A) than similar juveniles in B. Conversely, small, young individuals transplanted from B to A will grow more slowly and reach smaller final sizes than those in A (and will be like those left in B). These hypotheses are illustrated in Fig. 2. For the moment ignore the other experimental treatments (discussed later, *see* section on controls).

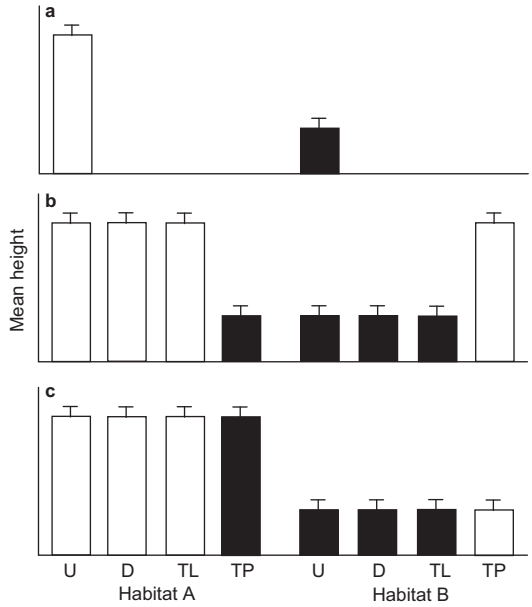

Fig. 2. Observed and predicted patterns of mean (+ SE) heights of plants in two habitats (A and B). — **a**: Observed difference in height; — **b** and **c**: Predicted heights in experimental transplants; U are undisturbed plants remaining, untouched, in the original habitat; TP are transplanted to the other habitat. White originate in habitat A, black in habitat B. D are disturbed (control) plants; TL are translocated (control) plants, as explained in the text. **b** are hypotheses from the model that growth/height differ because of intrinsic differences between plants from the two habitats. **c** are hypotheses from the model that growth/height are determined by the habitats themselves.

In contrast, if model (ii) is correct and environmental factors influence growth and sizes, juveniles transplanted from habitat A to B will now grow more slowly than those remaining in A and will reach shorter final sizes matching individuals remaining in B throughout. Conversely, juveniles transplanted from B to A will now grow as fast as those originating in A and therefore will reach taller mature sizes than those remaining in B. (These hypotheses are shown as treatments U and TP in Fig. 2c.)

Making the hypotheses as explicit as possible (and, where possible, indicating precisely what

data each predicts) clarifies the expectations of results from the experiments and greatly aids the interpretation of results in relation to the models being contrasted.

Explicit hypotheses in this case are also extremely important because there are choices of how data can be analysed. Without worrying about details, the data here could be analysed by general linear modelling or analysis of variance (O'Hara 2009). You could choose to distinguish between the two sets of hypotheses (from the two models being discussed) to analyse data considering their *origin* or their final *destination*. The two models predict very different patterns in the data (as in Fig. 2b *versus* 2c), but also different patterns in analyses by origin or by destination (Table 1).

If analysis is by origin, there are two experimental factors. "Origin" contrasts all plants from habitat A with all plants from B, wherever they finished up in the experiment (so compares the data in white columns with those in shaded columns). The second factor is "Experimental treatment" which contrasts undisturbed plants (those in their original habitat; U in Fig. 2b or 2c) with transplanted individuals (TP in Fig. 2b or 2c). As shown in Table 1, if model (i) is correct, such an analysis will reveal a significant difference between origins because all individuals from A will grow at similar rates and faster than those from B, whichever habitat they are in. In contrast (as in Table 1b), if model (ii) is correct, there will be a statistical interaction (Origin × Treatment) because the individuals transplanted from A to B now grow more slowly than undisturbed individuals from the same origin (i.e. A; U in Fig. 2c), but the transplants from B to A will grow more quickly than the undisturbed individuals from the same origin. Thus, the difference between experimental treatments U and TP depends whether individuals came from A or B (i.e. depends on their origin). This is straightforward and the two different outcomes are easy to identify and associate with the two different models.

In contrast, if the data are analysed according to where the plants finished up (their destinations), the patterns expected in analyses are quite different (as shown in Table 1b). Now, if model (i) is correct and differences are due to whether plants originally come from habitat A or B, there will be an interaction between factor destination (i.e. all plants grown in A *versus* all plants grown in B) and Treatments. The plants taken from A to B (white TP in Fig. 2b) will now grow more quickly than those left in B (black U in Fig. 2b), but the transplants from B to A (black TP in Fig. 2c) will grow more slowly than those remaining in habitat A (white U in Fig. 2c). Clearly, the difference between treatments U and TP depends upon the habitat in which the plants grew.

If model (ii) is correct, there will now be no interaction. Because growth and final size are determined by the habitat, all plants in treatments U and TP in Habitat A (i.e. white U and black TP in Fig. 2c) will grow quickly and all plants in treatments U and TP (black U, white TP in Fig. 2c) will grow slowly. There is no dif-

Table 1. Illustration of analyses of experimental transplants (*see* also Fig. 2). "Origin" indicates whether experimental plants come from habitat A (white columns in Fig. 2) or habitat B (black in Fig. 2). "Destination" indicates whether plants grew in habitat A or in habitat B. Experimental treatments are U, or undisturbed (i.e. remaining in the original habitat or not being moved to the destination habitat) and TP, or transplanted (from the original to the destination habitat). * indicates components of analyses expected to be large under different hypotheses. For detailed explanation, *see* text.

	Model (i) is correct	Model (ii) is correct
a: Analyses considering *origin* of plants		
Origin = O [A *versus* B]	*	
Treatments = T [U *versus* TP]		
Interaction: O × T		*
b: Analyses considering *destination* of plants		
Destination = D [A *versus* B]		*
Treatments = T [U *versus* TP]		
Interaction: D × T	*	

ference between treatments, a major difference between destination habitats and no interaction (Table 1b).

Note that this experiment is vastly oversimplified (there are other models and other treatments, controls D and TL in Fig. 2). Nevertheless, it illustrates that only by being very clear about the hypotheses is it going to be possible to be clear about the analyses and interpretations with respect to the different models about ecological processes.

Some issues in experimental design

Estimating appropriate variation

Whatever framework of analysis is going to be used (including hypothesis-testing, Bayesian methods, likelihood ratios, modelling), no sensible inference is going to be made from data that are inappropriate. The most common issue is to design the sampling or experimentation to include replication — so that the estimates of parameters from samples can be compared against the appropriate natural and sampling variation. One of the best discussions of this issue for ecologists is Hurlbert (1984). Despite recent attempts to pretend that the notion of replication is unnecessary (Oksanen 2001), Hurlbert (1984, 2004) is correct about the necessity to measure variation.

There are, in fact, two different issues — replication to measure variability and to ensure that conclusions are not confounded. The first is straightforward. Every ecological feature that needs to be measured (the numbers of species in areas of forest, the sizes of frogs in a study area, the rate of feeding or movement, etc., etc.) is intrinsically variable. This occurs either because of genetic variation among the organisms, so that they intrinsically vary in rates of feeding, digestion, processing of energy, etc. and therefore vary in rate of growth, final size, speed of movement and so forth. Even if there were no intrinsic variability, the habitat and environment are not constant in time and space. As a result, individuals in different places encounter food of different quality, in different amounts and have different amounts of time to feed. Even if a stretch of habitat were somehow constant, many animals and plants can only occupy it as a result of processes of dispersal (as juveniles or adults). But dispersal is variable (the distances travelled and rates of movement vary because of genetical and environmental variation), so the numbers arriving, times of arrival, etc., are variable. Finally, even if the biological property were not variable, there can be considerable variation in its measurement, as a result of the methods used. So, for example, there may be a constant concentration of chlorophyll per gram of tissue in a species of plant, but measuring chlorophyll and weight are each subject to error in the methods and equipment used.

So, a very important issue is to identify the appropriate error to measure, so that results can be reliably compared.

A simple case is a comparison of the numbers of fish in areas of seagrass where there are adjacent mangrove trees with those in seagrass without adjacent mangroves. The observations were the numbers of fish vary from one area of seagrass to another and the model has been proposed that a major contribution to this variation is the presence or absence of mangroves which provide nutrients for this fish (e.g. Pittman *et al*. 2004, Skilleter *et al*. 2005). The hypothesis to be tested is that the numbers of fish will, on average, be greater in seagrass with adjacent mangroves than in seagrass without. The null hypothesis is that there will be no difference, on average, between these two habitats, or there will, in fact, be more fish where there are no mangroves. Note, this would have been a 2-tailed proposition if the alternative model was also considered that fish spend more time in mangroves and would therefore be less abundant in seagrass near mangroves. Suppose that fish can reliably be sampled by some appropriate nets and that it is decided to sample 10 areas of seagrass with and 10 areas without mangroves. Ignore for now how these 20 areas are chosen to be sampled and assume that they are all sampled independently (which probably means that they are well-spaced, so that the numbers in any area are not correlated with those in adjacent areas). It is, however, not realistic to assume that any particular area can be well sampled by one net. In other words, if several

nets were used in any area, the numbers of fish would be different from net to net. So 6 nets are scattered in each of the 20 areas.

There are now two types or scales of replication, net-to-net in any area and area-to-area in each type of habitat. The net-to-net variation is the natural variability from place to place within a patch of habitat. The area-to-area variation estimates differences among patches of the same type of habitat. It is this variability that must be examined in any test of differences between the two types of habitat. The rationale for this is extremely simple (Hurlbert 1984), even though it has been missed by many ecologists. If you calculate the average number of fish in the 6 nets in one area of seagrass without mangroves (call it A1–), it is an estimate of the average number of fish per net over the entire area. The equivalent estimate for one area of seagrass with mangroves (A1+) is the average from the 6 nets in that area. It is, however, expected that these two averages will not be the same because of natural variation in numbers of fish from area to area, regardless of any influence of mangroves. The average numbers of fish in areas without mangroves will differ because of variability from area to area in all sorts of ecological processes. There will be similar variation among areas with mangroves. In addition, even if the true average number of fish over each of the two areas really happened to be identical, there is sampling error — the average measured by sampling 6 nets is not exactly the true average. The sampled average is affected by how many nets are used (the more nets, the closer the sampled estimate will be, on average, to the real number) and how the numbers of fish vary across an area (the variance in numbers).

So, the hypothesis being tested is that the average number of fish in areas of seagrass with mangroves is larger than the average number in areas without mangroves and the difference is larger than would be expected between any areas of seagrass of the same type, each sampled with 6 nets. In other words, areas of seagrass with or without mangroves will have different average numbers of fish simply because they are different areas. If mangroves exert or are associated with any difference, there has to be a systematic and larger difference between the two habitats than occurs due to spatial variability in numbers of fish from area to area. The relevant null hypothesis is not then that the two types of habitat are equal, because no-one expects them to be. The null hypothesis is that the average numbers of fish differ between the two types of habitat by an amount consistent with natural variation among areas of the same type.

To determine how much variation in mean numbers of fish would be expected from one area to another requires replication of areas of each type. Thus, there are three potential differences: (i) between areas with and areas without mangroves (as predicted by the hypothesis, but absent under the null hypothesis); (ii) from area to area with mangroves and from area to area without mangroves (due to differences from area to area in various ecological processes); (iii) from net to net in each area (due to spatial variation in the number of fish in each area and due to sampling error because each net can have different efficiency or effectiveness). To test the hypothesis statistically and to estimate how large the difference is between types of habitats (to determine whether it seems large enough to be ecologically important) requires that (i) be larger than (ii). The third difference is irrelevant to this procedure.

An alternative way to understand this is to note that the comparison of the two types of habitat (to test the hypothesis) could be done using the average number of fish in each area, i.e. averaging the numbers of fish from the 6 nets in each area. Then, there *are* only two sorts of differences: (i) and (ii) above. Both include the variation from net to net ((iii) above) because both are samples from nets (and *see* Winer *et al*. 1991, Underwood 1997 for discussion of how and why the various differences combine in samples).

An appropriate test then compares the difference between the two types of areas (i.e. areas with mangroves minus areas without mangroves) with the differences among replicate areas of the same type. If the former is large relative to the latter (and, in statistical tests, if the former is sufficiently large relative to the latter that it is unlikely given the variability among areas), the null hypothesis is rejected. This is then usually interpreted to mean that the hypothesis is supported and the presence of mangroves is associated with more fish in adjacent areas of seagrass.

Controlling for artefacts

Any experimental manipulation or any mensurative comparison between different types of habitats is potentially confounded by uncontrolled causes of ecological differences. Trying to *control* these influences is extremely important. The need for controls has a long history in thinking about experiments (at least as far back as J. S. Mill's (1865) canons of similarity and difference). One of the satisfying things about controls is that thinking carefully about them requires knowledge and careful thought about the sorts of processes that may cause confusion in experiments. It also requires skill with the ecology and biology of the processes, habitats and organisms under investigation, so that meaningful controls can be attempted.

There is no space here to consider this in detail, but there are useful discussions in Mead (1988), Hairston (1989), Paine (1994) and Underwood (1997). One example will suffice to illustrate the issues. Consider again the experimental transplants discussed earlier and shown in Table 1 and Fig. 2. Small, juvenile plants are experimentally transplanted between habitats A and B and their growth assessed (as their final, mature sizes) and compared with the sizes of similar individuals left undisturbed in their original habitats (i.e. treatments TP and U, respectively).

Suppose at the end of the experiment, there are differences between plants in different habitats and treatments, leading to a refutation of one of the two models investigated (and experimental support for the other model). The problem is that the comparison of transplanted (TP) and undisturbed (U) plants does not just involve a difference between their origins and their destinations, as proposed by the hypotheses.

To transplant the individuals, two very different types of disturbance are inevitable, apart from the actual shift between habitats. First, to be transplanted, plants must be disturbed, i.e. taken from where they germinated and put into a different piece of ground. This involves numerous potential influences on growth and final size due to damage to roots, possible damage to leaves and stems, etc. So, transplanted individuals may differ from undisturbed individuals because of disturbance (nothing to do with the various hypotheses) and not because of transplantation (the possible effects of which were being predicted).

The second issue is that there are all sorts of local influences on growth due to variation in water, nutrients, soil, shade, etc., from one place to another in a habitat. These will lead to differences in growth and sizes among the individuals in different areas of habitat A and among those in different areas of B, even though they have not been disturbed or transplanted. So, moving individuals from A to B or vice versa may change their growth and size because they are now in a different spot and not because they are in a different habitat.

Controls to unconfound these influences consist of disturbing plants without moving them. Thus, plants are disturbed by being dug up and are then replanted, each in the same spot, using the same techniques as used for the transplanted plants. These are treatment D in Fig. 2. The second control is individuals that are disturbed by being dug up in the same way, but they are then planted in a new randomly-chosen position in the original habitat, rather than in a randomly-chosen position in the other (destination) habitat. These are often called a "translocation" control (TL in Fig. 2). As discussed in detail in Chapman (1986, 2000), TP differ from U plants by being disturbed, being moved to a new position and being transplanted to a different habitat. Demonstrating effects of transplantation, as predicted by one or other hypothesis, requires demonstration that treatments U, D and TL (undisturbed, disturbed and translocated) are similar, but differ from TP according to one set of predictions. This is the situation illustrated in Fig. 2, where there were no artefactual influences on growth and sizes due to disturbance or translocation. In this case, analyses should be for similarities of treatments, by tests for bio-equivalence (McDonald & Erickson 1994), but that is not considered here.

In many experiments, controls can be quite numerous and complex (e.g. Underwood 1988 for experiments with densities of limpets). This is particularly the case where animals are moved among habitats because there are not only potential artefacts affecting their physiology, but also many potentially serious influences on their

behaviour. Allowing logically coherent interpretations of results of experimental manipulations requires that confounding due to artefacts of the procedures and treatments does not confuse outcomes. So, carefully constructed controls are necessary.

As a final point on this issue, some ecologists (e.g. Connell 1974, Diamond 1986) draw distinctions between experiments done in the field and laboratory. They indicate that the latter are more carefully controlled, by controlling unwanted influences (temperature, moisture, light, etc.). The distinction is largely illusory. What laboratories (or mesocosms or any other artificial habitat) allow is better regulation of the variability of such influences, compared with what is possible in the field (Campbell *et al*. 2009). Thus, the variances in levels of environmental factors can be made much smaller in the regulated environments of laboratories. It is not usually zero (for example, there are inevitably temporal variations of temperature in temperature-controlled incubators), but small. It is still important to have appropriate replication and control treatments to ensure that these factors are not confusing the interpretations from experiments.

Analyses and interpretations

Because ecological measures are variable, data usually require statistical analyses. These impose considerations about assumptions required for tests, in addition to all the caveats about logic and maintaining consistency of the procedures leading to the experiments. Here, frequentist tests are considered, because of personal preference and because of conceptual problems about constructing prior probabilities for Bayesian procedures (*see* particularly Dennis 1996, Mayo 1996). If Bayesian analyses are favoured (*see* for example Ellison 1996 and Läärä 2009 for arguments in favour of these), different considerations will require attention.

Statistical tests require some model about the nature of the data. In principle, a relevant and appropriate test must be calculated from the data. The choice of appropriate statistic can be complex and ecologists are well advised to consult with statisticians before considering the final experimental design to be used. Once statistical procedures have been chosen, it is necessary to know what is the distribution of the test statistic if the null hypothesis were true. In other words, if the null hypothesis were true and there were many repeats of the experiment, each would generate data and therefore each would provide a calculated test statistic. The frequency distribution of these test statistics is their distribution when the null hypothesis is true (because it was true for every experiment). If the distribution of the test statistic can be known or assumed, it is not necessary to consider numerous repeats of the experiment and the observed value of the test statistic (using your single experiment) can be compared with the frequency distribution of the statistic if the null hypothesis were true.

So, it is usually necessary to make assumptions about the data (some of which can be tested using the experimental data), in order to assume that the test statistic has the particular frequency distribution that is assumed to apply. There is no space here to consider this topic in detail (get help from statisticians). As an example, however, student's *t*-test is widely used in ecological analyses, because it is versatile and because biological data often fit well to its assumptions. A common use is to compare two sampled means (as in the earlier example about fish in different areas of seagrass). The assumptions underpinning the use of a *t*-test are that the data in each sample are independently sampled, that the samples from the two habitats (plus *versus* minus adjacent mangroves) are approximately normally distributed (i.e. that the average numbers of fish across many areas of the two types of habitats are distributed with a particular shape of frequency distribution, called a normal distribution; see any statistical text for details, e.g. Sokal & Rohlf 1981, Winer *et al*. 1991, Quinn & Keough 2002). Finally, it is also assumed that the distributions of sample means in the two types of habitat have the same variances (i.e. are scattered variably to the same extent in each type of habitat). Many biological and ecological data do have sample means that fit reasonably closely to normal distributions. If the data do not, it may be possible to transform data to some other scale of measurement in which they are approximately normally distributed (*see*, for example, Tukey

1957, Box & Cox 1964, Legendre & Gallagher 2001). Preliminary tests on the data can be done to determine whether the variances are sufficiently similar. Care with sampling can help ensure that data are likely to be independent (*see* discussion in Underwood 1997).

Note that these types of assumptions are not unique to *t*-tests. The equivalent rank-order or so-called nonparametric test, sometimes called "distribution-free" test is the Mann-Whitney test (Hollander & Wolfe 1973, Siegel 1953). Use of this test also assumes that data are independently sampled and that the distributions of mean numbers of fish in the two habitats have the same variance. It does not require that the data are normally distributed, but it does require the distributions of data to be generally similar, except for their means (which will be similar if the null hypothesis is true and not if the hypothesis is correct).

If the data generally conform to the requirements of the test, the test statistic calculated from the data can be compared with the frequency distribution of the test statistic if the null hypothesis were true. This allows determination of the probability of getting the observed value of the test statistic if the null hypothesis were true. If that probability is large, the most likely reason is that the null hypothesis is true, i.e. the experimental data do not give any reason to consider that the null hypothesis is incorrect, because the results are consistent with what is expected from experiments where the null hypothesis is correct. If that probability is small, the interpretation is made that such a result is improbable (unlikely) if the null hypothesis is true and therefore the test rejects the null hypothesis (because it is unlikely to have led to that result) in favour of the hypothesis (the only alternative, given the way hypotheses and null hypotheses are constructed — *see* earlier).

Considerable argument has been raised about the logic of such decision-making procedures. As with all other frameworks about statistical procedures, choices have to be made, in advance of doing the experiment and therefore before the data have been collected, about how to arrive at a conclusion. For frequentist tests (those that use a frequency distribution of the test statistic) of the sort described above, the core issue is how improbable should the test statistic be so that the conclusion should be to reject the null hypothesis?

In other words, any result may occur, but some are unlikely if a null hypothesis is a correct description of events. If a null hypothesis were true, some outcomes (in terms of observed data) are quite likely. Others are relatively unlikely and yet others are very unlikely. What any experimenter would consider to be so unlikely that it would be more realistic to reject the null hypothesis is a matter of individual choice. This is considered below.

Here, note that the assumptions of the statistical procedure often require data to be collected in particular ways, so that sampling for an experiment may have to be carefully designed to ensure that assumptions are met. In the above example of a *t*-test, sampling must be designed to ensure that the data in each area of seagrass sampled are independent.

Design and analyses

There is obviously no room here to consider much about the way data should be analysed and how conclusions should be reached. It should be clear that organizing the framework and setting of the experiment, by due care to its logical requirements will make these tasks simpler and, generally, more reliable. Assuming that due care has been taken, it should be straightforward to ensure that the experimental procedures are at the appropriate scales of space and time to match the original observations and the processes being proposed to account for them. Thus, if it has been observed that densities of beetles vary over spatial scales of 10s of metres, experiments at smaller scales are unlikely to be testing relevant hypotheses. Variation, or lack of it, among experimental plots at scales of 10s of cm apart would not be a very convincing evaluation of some purported explanatory model. If observations are about the patterns of dispersion of juvenile animals just after breeding in spring, testing hypotheses about processes affecting dispersion in autumn is not obviously relevant.

It is also necessary, as briefly discussed earlier, to be very careful about the controls that are necessary in any experimental manipulation. As with

scales above, this is all about trying to understand the relevant biology/ecology of the organisms, habitats and processes being investigated.

There are, however, other concerns to do with the amount of replication of experimental and control procedures that is necessary to be able to reach reliable conclusions. This can best be thought about in terms of frequentist procedures, but important elements are equally necessary for other ways of dealing with data. Whenever ecological measures are being sampled, the data are used to estimate the true values of the measure. The true values cannot be known, because that would require making measurements of every possible relevant unit under study. For example, if some hypothesis predicts that plants grow more quickly under some conditions than under others, the growth of plants under each of these conditions must be measured or the hypothesis cannot possibly be tested. But the measurements are only made on the relatively few plants in the experiment, not on all possible plants to which the hypothesis might apply. The growth of plants varies from individual to individual, so the average or range of measurements will not be the same as the real average or range about which predictions are made.

For statistical procedures using frequentist tests (the familiar χ^2, *t*-test, analysis of variance, etc.), the extent to which the measures made in the experiment are precise (i.e. close to the unknown, true value) can be measured, provided the data conform to the assumptions of the statistical procedures used to measure precision (as mentioned earlier, all statistical procedures impose assumptions). Detailed descriptions of the data from the samples measured in each experimental treatment should therefore include not just the sampled measures themselves (their means, medians, ranges, frequency distributions, variances, etc., as required by the hypotheses), but also indications of their precision (Läärä 2009). This is often done by providing estimates of variance, standard deviation, standard error, or their alternatives. For complex experiments, calculation of these can be complex (*see* discussion in Quinn & Keough 2002).

Discussion of imprecision, because of intrinsic variation in the things being measured and because sampling inevitably does not allow calculation of the real quantities, has led to some general principles that help with designing experiments. Although these are almost always considered with respect to frequentist statistical procedures, in the context of what is called power analysis (Cohen 1977, Winer *et al*. 1991, Underwood 1997, Ford 2009), they are much more general. In many ways, they reflect common sense, which is some sort of criterion that may or may not be useful. In this case, it seems to be, because common sense produces the same conclusions as mathematical analyses. The points are that if a hypothesis, as is generally the case, predicts that some quantity is of different magnitude under different controlled experimental conditions, it is more likely that the difference will be seen in the experimental data if:

i. measures are made on more, rather than fewer, experimental units (i.e. samples are large);
ii. the intrinsic variability in the measures from one experimental unit to another (i.e. the variance of the measures being made) is relatively small; and
iii. the predicted difference(s) between (among) the treatments are relatively large (these differences are often referred to as the "effect-size").

Obviously, considering (iii) first, if you predict that the difference in densities of animals between two experimentally constructed habitats will be about 10 individuals per m^2 you have less chance of detecting it than if you predict it to be 50 individuals per m^2. Depending on the intrinsic variability of densities from place to place and on how precise the samples are, a difference of 10 might be unimportant because of the fuzziness of the measures in each habitat. A difference of 50 might also be difficult to discern, but is larger and therefore more likely to be detectable against the background variation.

The other two points [(i) and (ii) above] are clearly related to the same issue. If the measures being made are intrinsically very variable (there is large variance among replicates), there will be lots of imprecision and differences among treatments will be blurred [as in (ii)]. If samples are small, the measures being estimated are more

imprecise than when the same measures are taken with larger samples [point (i) above].

So, designing experiments includes being very thoughtful about how many replicate measures are appropriate. In general, the more you can afford or can handle logistically, the better will be the experimental results. The design should also take note of the intrinsic variance in the measures. Where this is large, it is generally necessary to have more replicate measures than where this is small (*see* Winer *et al*. 1991, Cohen 1977).

There is no space here to consider details of how to determine an appropriate size of experiment (i.e. how to determine an appropriate number of replicates). Remember that the correct units of study must be replicated (*see* the previous consideration about patches of seagrass, rather than individual nets being the correct components of the experiment to replicate). In general, however, any test which has more than about 30 degrees of freedom (df) associated with the statistics used will be large. What this means in practice is that having enough replicates to achieve 30 df will be nearly as likely to reject an incorrect null hypothesis as an experiment with more replicates (and therefore more df). As an example, consider an experiment to test the hypothesis that removing predators will increase the numbers of a prey species, in each of two types of habitat (say, under trees as opposed to open grassland). There will be 3 experimental treatments: untouched (i.e. intact predators), predators removed and a control for the disturbances caused by removing predators and keeping them out. The experiment will be done in two areas of the study. So, there are 2 areas, 2 habitats, 3 treatments and therefore 12 combinations of all of the requirements of the experiment. If 4 replicate plots of each combination are established in each area (i.e. there are 4 replicates of each of the 12 combinations of treatments) and data are to be examined by analysis of variance, there would be 36 df for some of the tests (*see* Winer *et al*. 1991, Underwood 1997 for how to calculate df). If resources were available to have 5 replicates instead, there would be 48 df instead of 36. This increase in effort and costs would not be worth it because it would not make much difference to the capacity of the experiment to cause the null hypotheses to be rejected.

If, as is usually the case, resources to do the experiment are very limited, having the minimal number of 2 replicates creates only 12 df. Using all efforts to get 3 replicates (24 df) would be a very substantial improvement.

The appropriate number of replicates is a large and important issue. In many cases, increasing replication to make large experiments is not the best strategy, compared with doing several small experiments and combining the results (perhaps in a meta-analysis; *see* Gurevitch *et al*. 1992). Reference to discussions of this topic will help (e.g. Cohen 1977, Underwood 1997).

It is also worth remembering that too much introspection about designing experiments generally results in them being too difficult to do. Suppose the original observations were made about patterns found from few replicate measures. Models about processes causing the observed patterns are then made and predicted patterns are constructed for the data that will be collected from the experiment. If some model or combination of models does actually cause the observed pattern, it is likely that it will appear in the experiment, even if only few replicates are used. Thus, the pattern is evident in observations from a few replicates and under experimental conditions from a few replicates. In contrast, if it took large numbers of replicated observations to discern the original pattern, it is not reasonable to expect it to be easy to see in experiments with few replicates.

Interpretation of experiments

Given the logical framework briefly described here (Fig. 1), at the end of the experiment a decision has to be made about how the results are to be interpreted. There are only two possibilities. The experimental data lead to the conclusion that the hypothesis is correct. This conclusion is reached when it is decided to reject the null hypothesis, or an appropriate combination of null hypotheses, leaving the predicted pattern of data supporting the hypothesis(es) because that is the only other possibility. It is the only other possibility because of the way null hypotheses are defined and constructed (*see* earlier).

When statistical tests are used to help arrive at a decision, it is necessary to define, in advance, a probability to use as the criterion to make a decision. Briefly, a probability (called alpha, α) is chosen to define unlikely results, i.e. data that are unlikely to occur in the experiment, if the null hypothesis were true. The proposition for frequentist tests is that any result that is actually possible may occur, whether or not the hypothesis or the null hypothesis is actually true. Some results are, however, relatively unlikely if the null hypothesis is occurring. Alpha is chosen so that these results, collectively, have a small probability of occurring when the null hypothesis is true. Suppose, as is often done by convention, α is chosen to be 0.05, i.e. there is at most a 1 in 20 chance of certain results occurring if the null hypothesis is true. If such results do, in fact, occur in the experiment, the decision is made to reject H_0 (because that is what you decided, in advance, to do). The conclusion is then that the hypothesis and its underlying model are supported by the experimental data (as in Fig. 1).

In contrast, if other results occur, the decision must be to retain the null hypothesis, simply because no other decision can be made. The outcome can only be that H_0 is rejected or it is not. Critics of this sort of procedure argue that this is equivalent to saying that the size of the probability of the statistical outcome of the test is used to demonstrate that H_0 is true. This criticism is incorrect, because if H_0 is retained, it is retained because the experiment has not provided evidence to falsify it. The hypothesis has, however, been falsified under these experimental procedures (*see* also the discussion of this point in Dennis 1996).

Note that this procedure does not attempt to measure the extent to which the data are supporting the hypothesis. The probability calculated is the likelihood of getting the observed magnitude of data if the null hypothesis were true. This cannot shed any light on the likelihood that the hypothesis is true given the observed data (i.e. in contrast to the goals of Bayesian analyses).

In either case (and as is in common with other frameworks of decision-masking about experimental data) mistakes can occur. The simplest to understand is that H_0 will be rejected even if it was operating during the experiment. This is *always* possible because α is chosen to represent unlikely results (not impossible results). So, if the data and statistical procedures lead to H_0 being rejected, there is a maximal chance of the chosen value of α (chosen above to be 0.05) that the results would actually be consistent with H_0. There is a 0.05 probability that the null hypothesis should *not* have been rejected.

It is also the case that outcomes that do not cause rejection of H_0, i.e. results that cause H_0 to be retained, can be consistent with (produced as a result of) the hypothesis, not the null hypothesis. If such results have occurred, the hypothesis is being rejected and H_0 retained erroneously. It is not easy to keep the probabilities of both of those types of error small, i.e. to decrease the chances of making either of the two types of mistake. How to attempt this by careful design of experiments has been discussed by many authors (Cohen 1977, Sokal & Rohlf 1981, Underwood 1997 among many others). The issues are complex and cannot be elaborated here.

What is clear is that firm conclusions about ecological processes can never be based on single experiments at one set of times or places, however well designed, done and analysed they are. All experiments are in a programme of experimental projects. So, the conclusions reached are either to reject H_0 and support the hypotheses and models that contradicted it or to keep H_0 and reject the hypotheses and models. Either conclusion can be wrong for purely statistical reasons (as above), because unusual ecological circumstances were prevailing when the experiment was done or because other models, not yet thought of, caused the experimental data. Such confounding models have not been falsified (they have not even been proposed), but are always possible.

Therefore, the outcome of the experiment is to start again, with the new observations gained from the experiment, added to the original observation. Note my agreement with Peckarsky (1998) who considered that experiments can be used to make observations, which is not at all contrary to the framework being considered here. The only point of disagreement is that she did not identify the reasons for doing such experiments in terms of underlying processes and predictions, which are discernible, but not

explicit in her examples.

Starting again requires new models (if H_0 was retained) or different predictions and more severe tests (i.e. predictions that are more precise or more general) if the hypothesis and model were supported. New models can, of course, include the idea that a null hypothesis was retained or rejected because of statistical (probabilistic) errors as described above. Or new models can be about completely different processes now that it is considered that some previous process is not operating.

Because experiments really cannot be interpreted in isolation, ecologists have developed quite sophisticated methods to interpret results from several (ideally many) different experiments. One methodology is to review and collate results from experiments about the same process, but done in different ways on different organisms at different times in different places. An example was the reviews on field experimental tests of hypotheses about competition amongst consumers of resources done by Connell (1983) and, independently, by Schoener (1983). By bringing together numerous experimental studies on competition, each author found several coherent patterns (and used the reviews to test some predictions about where and when competition should be important). Needless to say, the two reviews did not result in identical conclusions, partly because different methods and criteria were used (as discussed in a review of these reviews in Hairston 1985).

A sound methodology uses statistical procedures to bring together results of different experiments to test similar or the same hypotheses. Full details of appropriate methods, with examples, of such meta-analyses are given by Gurevitch *et al*. (1992). There are also procedures [e.g. Fisher's (1935) combinatorial statistic] to bring together results of repeated experiments, each testing the same hypothesis, where samples are small in each experiment.

Clearly, understanding ecological processes is achieved better by treating experiments as parts of an ongoing research programme, rather than considering each experimental conclusion as being valid in its own right. Each experiment can also be used as a pilot or preliminary study to help improve the design of the next one.

Conclusions

This is inevitably a brief and foreshortened account of relevant issues underlying field experiments as tools to help understand ecology. It is also inevitable that many readers will disagree with components of it (or, indeed, with all of it). So, readers should consider these issues and form a view and then practice their ecological research consistently according to that view. That requires being able to contrast what is here with proposed alternatives before forming judgements.

What is important is the fact that ecological field experiments do not exist unconnected to the relevant ecology of the organisms and habitats and the underlying technical issues, assumptions and procedures of analyses of data. They are also firmly enmeshed in the structures, the thought-processes by which they are justified and interpreted. Where describers of experiments have not made clear the logical underpinnings for their work and its outcomes, it is always more difficult to know how to understand and absorb their findings.

Designing experiments requires detailed knowledge of the fauna or flora and the habitats investigated. It requires help with technical requirements and assumptions of analytical procedures. It depends on intimate understanding of the logical frameworks by which decisions about results will be made. Ecological understanding increases as a result of field experiments. These are of limited value when the components underlying their success are not explicit and well-used. Ecological experiments should never be tragic accidents of poor thought. Instead, ecology should be advanced by design.

Acknowledgements

Preparation of this paper was funded by an ARC Professorial Fellowship and ARC Discovery Project grants. I have benefitted greatly from discussions over many years with numerous students and colleagues, most notably Gee Chapman.

References

Andrew, N. L. & Mapstone, B. D. 1987: Sampling and the description of spatial pattern in marine ecology. —

Annual Review of Oceanography and Marine Biology 25: 39–90.

Bacon, F. 1620: *Novum organum, 1889 Edition*. — Clarendon Press, Oxford.

Box, G. E. P. & Cox, D. R. 1964: An analysis of transformations. — *Journal of the Royal Statistical Society, Series B* 26: 211–243.

Brown, J. H. 1999: Macroecology: progress and prospect. — *Oikos* 87: 3–14.

Carnap, R. 1962: *Logical foundations of probability*. — University of Chicago Press, Chicago.

Campbell, D. L. M., Weiner, S. A., Starks, P. T. & Hauber, M. E. 2009: Context and control: behavioural ecology experiments in the laboratory. — *Annales Zoologici Fennici* 46: 112–123.

Caughley, G. & Gunn, A. 1996: *Conservation biology in theory and practice*. — Blackwell Science, Cambridge, Massachusetts.

Chalmers, A. F. 1979: *What is this thing called science?* — Queensland University Press, Brisbane.

Chamberlin, T. C. 1965: The method of multiple working hypotheses. — *Science* 148: 754–759.

Chapman, M. G. 1986: Assessment of some controls in experimental transplants of intertidal gastropods. — *Journal of Experimental Marine Biology and Ecology* 103: 181–201.

Chapman, M. G. 2000: Poor design of behavioural experiments gets poor results: examples from intertidal habitats. — *Journal of Experimental Marine Biology and Ecology* 250: 77–95.

Cohen, J. 1977: *Statistical power analysis for the behavioural sciences*. — Academic Press, New York.

Connell, J. H. 1974: Ecology: field experiments in marine ecology. — In: Mariscal, R. (ed.), *Experimental marine biology*: 21–54. Academic Press, New York.

Connell, J. H. 1983: On the prevalence and relative importance of interspecific competition: evidence from field experiments. — *American Naturalist* 122: 661–696.

Dennis, B. 1996: Should ecologists become Bayesians? — *Ecological Applications* 6: 1095–1103.

Diamond, J. M. 1986: Overview: laboratory experiments, field experiments and natural experiments. — In: Diamond, J. M. & Case, T. J. (eds.), *Community ecology*: 3–22. Harper & Row, New York.

Ellison, A. M. 1996: An introduction to Bayesian inference for ecological research and environmental decision-making. — *Ecological Applications* 6: 1036–1046.

Fisher, R. A. 1935: *The design of experiments*. — Oliver and Boyd, Edinburgh.

Ford, E. D. 2009: The importance of a research data statement and how to develop one. — *Annales Zoologici Fennici* 46: 82–92.

Gaston, K. J. & Blackburn, T. M. 1999: A critique for macroecology. — *Oikos* 84: 353–368.

Gurevitch, J., Morrow, L. L., Wallace, A. & Walsh, J. S. 1992: A meta-analysis of competition in field experiments. — *American Naturalist* 140: 539–572.

Hairston, N. G. 1985: The interpretation of experiments on interspecific competition. — *American Naturalist* 125: 321–325.

Hairston, N. G. 1989: *Ecological experiments: purpose, design and execution*. — Cambridge University Press, Cambridge.

Heath, O. V. S. 1970: *Investigation by experiment*. — Edward Arnold, London.

Hempel, C. G. 1965: *Aspects of scientific explanation*. — Free Press, New York.

Hocutt, M. 1979: *The elements of logical analysis and inference*. — Winthrop, Cambridge.

Hollander, M. & Wolfe, D. A. 1973: *Nonparametric statistical methods*. — Wiley, New York.

Hume, D. 1779: *Dialogues concerning natural religion* (2nd ed. 1947). — Nelson, London.

Hume, D. 1941: *Treatise on human nature*. — Clarendon Press, Oxford.

Hurlbert, S. J. 1984: Pseudoreplication and the design of ecological field experiments. — *Ecological Monographs* 54: 187–211.

Hurlbert, S. H. 2004: On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. — *Oikos* 104: 591–597.

Läärä, E. 2009: Statistics: reasoning on uncertainty, and the insignificance of testing null. — *Annales Zoologici Fennici* 46: 138–157.

Legendre, P. & Gallagher, E. D. 2001: Ecologically meaningful transformations for ordination of species data. — *Oecologia* 129: 271–280.

Lemmon, E. J. 1971: *Beginning logic*. — Nelson, Surrey.

Mayo, D. G. 1996: *Error and the growth of experimental knowledge*. — University of Chicago Press, Chicago.

McDonald, L. L. & Erickson, W. P. 1994: Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed? — In: Fletcher, D. J. & Manly, B. F. J. (eds.), *Statistics and environmental monitoring*: 183–197. University of Otago Press, Dunedin, New Zealand.

Mead, R. 1988: *The design of experiments: statistical principles for practical applications*. — Cambridge University Press, Cambridge.

Medawar, P. 1969: *Induction and intuition in scientific thought*. — Methuen, London.

Mill, J. S. 1865: *A system of logic*, vol. 2, 6th ed. — Longmans, Green & Co., London.

Miller, B. & Mullette, K. J. 1985: Rehabilitation of an endangered Australian bird: the Lord Howe Island woodhen *Tricholimnas sylvestris* (Sclater) — *Biological Conservation* 34: 55–95.

Nagel, E. 1961: *The structure of science*. — Harcourt, Brace, World, London.

O'Hara, R. B. 2009: How to make models add up — a primer on GLMMs. — *Annales Zoologici Fennici* 46: 124–137.

Oksanen, L. 2001: Logic of experiments in ecology: is pseudoreplication a pseudoissue? — *Oikos* 94: 27–38.

Paine, R. T. 1994: *Marine rocky shores and community ecology: an experimentalist's perspective*. — Ecology Institute, Oldendorf-Luhe.

Peckarsky, B. L. 1998: The dual role of experiments in complex and dynamic natural systems. — In: Resetarits, W. J. & Bernardo, J. (eds.), *Experimental ecology: issues and perspectives*: 311–324. Oxford University Press,

Oxford.

Pittman, S. J., McAlpine, C. A. & Pittman, K. M. 2004: Linking fish and prawns to their environment: a hierarchical landscape approach. — *Marine Ecology Progress Series* 283: 233–254.

Popper, K. R. 1968: *The logic of scientific discovery*. — Hutchinson, London.

Popper, K. R. 1969: *Conjectures and refutations*. — Routledge & Kegan Paul, London.

Popper, K. R. 1982: *Unended quest: an intellectual autobiography*. — Fontana-Collins, Glasgow.

Quinn, G. P. & Keough, M. J. 2002: *Experimental design and data analysis for biologists*. — Cambridge University Press, Cambridge.

Russell, B. 1912: *Problems of philosophy*. — Willian & Norgate, London.

Schoener, T. W. 1983: Field experiments on intraspecific competition. — *American Naturalist* 122: 240–285.

Siegel, S. 1953: *Nonparametric statistics for the behavioral sciences*. — McGraw-Hill, New York.

Skilleter, G. A., Olds, A. R., Loneragan, N. R. & Zharikov, Y. 2005: The value of patches of intertidal seagrass to prawns depends on their proximity to mangroves. — *Marine Biology* 147: 353–365.

Sokal, R. R. & Rohlf, F. J. 1981: *Biometry: the principles and practice of statistics in biological research*. — W. H. Freeman, San Francisco.

Trusted, J. 1979: *The logic of scientific inference*. — Macmillan, London.

Tukey, J. W. 1957: The comparative anatomy of transformations. — *Annals of Mathematical Statistics* 33: 1–67.

Underwood, A. J. 1988: Design and analysis of field experiments on competitive interactions affecting behaviour of intertidal animals. — In: Chelazzi, G. & Vannini, M. (eds.), *Behavioural adaptation to intertidal life*: 333–358. Plenum Press, New York.

Underwood, A. J. 1990: Experiments in ecology and management: their logics, functions and interpretations. — *Australian Journal of Ecology* 15: 365–389.

Underwood, A. J. 1991: The logic of ecological experiments: a case history from studies of the distribution of macro-algae on rocky intertidal shores. — *Journal of the Marine Biological Association of the United Kingdom* 71: 841–866.

Underwood, A. J. 1997: *Experiments in ecology: their logical design and interpretation using analysis of variance*. — Cambridge University Press, Cambridge.

Underwood, A. J., Chapman, M. G. & Crowe, T. P. 2004: Identifying and understanding ecological preferences for habitat or prey. — *Journal of Experimental Marine Biology and Ecology* 300: 161–187.

Underwood, A. J. & Denley, E. J. 1984: Paradigms, explanations and generalizations in models for the structure of intertidal communities on rocky shores. — In: Strong, D. R., Simberloff, D., Abele, L. G. & Thistle, A. (eds.), *Ecological communities: conceptual issues and the evidence*: 151–180. Princeton University Press, New Jersey.

Winer, B. J., Brown, D. R. & Michels, K. M. 1991: *Statistical principles in experimental design, third edition*. — McGraw-Hill, New York.

Wittgenstein, L. 1921: *Tractacus logico-philosophicus, translated by D. F. Pears and B. F. McGuinness, 1961*. — Routledge & Kegan-Paul, London.