Ecological Monographs, 54(2), 1984, pp. 187–211

PSEUDOREPLICATION AND THE DESIGN OF ECOLOGICAL FIELD EXPERIMENTS[1]

STUART H. HURLBERT
Department of Biology, San Diego State University, San Diego, California 92182 USA

Abstract. Pseudoreplication is defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent. In ANOVA terminology, it is the testing for treatment effects with an error term inappropriate to the hypothesis being considered. Scrutiny of 176 experimental studies published between 1960 and the present revealed that pseudoreplication occurred in 27% of them, or 48% of all such studies that applied inferential statistics. The incidence of pseudoreplication is especially high in studies of marine benthos and small mammals. The critical features of controlled experimentation are reviewed. Nondemonic intrusion is defined as the impingement of chance events on an experiment in progress. As a safeguard against both it and preexisting gradients, interspersion of treatments is argued to be an obligatory feature of good design. Especially in small experiments, adequate interspersion can sometimes be assured only by dispensing with strict randomization procedures. Comprehension of this conflict between interspersion and randomization is aided by distinguishing pre-layout (or conventional) and layout-specific alpha (probability of type I error). Suggestions are offered to statisticians and editors of ecological journals as to how ecologists' understanding of experimental design and statistics might be improved.

Key words: experimental design; chi-square; R. A. Fisher; W. S. Gossett; interspersion of treatments; nondemonic intrusion; randomization; replicability; type I error.

No one would now dream of testing the response to a treatment by comparing two plots, one treated and the other untreated.

—R. A. Fisher and J. Wishart (1930)

. . . field experiments in ecology [usually] either have no replication, or have so few replicates as to have very little sensitivity . . . .

—L. L. Eberhardt (1978)

I don't know how anyone can advocate an unpopular cause unless one is either irritating or ineffective.

—Bertrand Russell (in Clark 1976:290)

INTRODUCTION

The following review is a critique of how ecologists are designing and analyzing their field experiments. It is also intended as an exploration of the fundamentals of experimental design. My approach will be: (1) to discuss some common ways in which experiments are misdesigned and statistics misapplied, (2) to cite a large number of studies exemplifying these problems, (3) to propose a few new terms for concepts now lacking convenient, specific labels, (4) to advocate treatment interspersion as an obligatory feature of good design, and (5) to suggest ways in which editors quickly can improve matters.

[1] Manuscript received 25 February 1983; revised 21 June 1983; accepted 25 June 1983.

Most books on experimental design or statistics cover the fundamentals I am concerned with either not at all or only briefly, with few examples of misdesigned experiments, and few examples representing experimentation at the population, community or ecosystem levels of organization. The technical mathematical and mechanical aspects of the subject occupy the bulk of these books, which is proper, but which is also distracting to those seeking only the basic principles. I omit all mathematical discussions here.

The citing of particular studies is critical to the hoped-for effectiveness of this essay. To forego mention of specific negative examples would be to forego a powerful pedagogic technique. Past reviews have been too polite and even apologetic, as the following quotations illustrate:

There is much room for improvement in field experimentation. Rather than criticize particular instances, I will outline my views on the proper methods . . . . (Connell 1974)

In this review, the writer has generally refrained from criticizing the designs, or lack thereof, of the studies cited and the consequent statistical weakness of their conclusions; it is enough to say that the majority of the studies are defective in these respects. (Hurlbert 1975)

. . . as I write my comments, I seem to produce only a carping at details that is bound to have the total effect of an ill-tempered scolding . . . . I hope those whose work I have referenced as examples will

forgive me. I sincerely admire the quality of these papers . . . . (Hayne (1978)

Among the 151 papers investigated, a number of common problems were encountered . . . . It would be a profitless, and probably alienating, chore to discuss these with respect to individual papers. (Underwood 1981)

But while I here offer neither anonymity nor blanket admiration, let me state an obvious fact—the quality of an investigation depends on more than good experimental design, so good experimental design by itself is no guarantee of the value of a study. This review does not evaluate the overall quality of any of the works discussed. Most of them, despite errors of design or statistics, nevertheless contain useful information.

On the other hand, when reviewers have tried to emphasize the positive by pointing to particular field studies as being exemplary, their choices sometimes have seemed inappropriate. For example, Connell (1974) cites Boaden (1962) as being "one of the best examples of a controlled field experiment"; and Chew (1978) cites Spitz (1968) as "the best example I have of the responses of plants to grazing by small mammals." Yet neither of the cited studies replicated their treatments, and both are therefore uncontrolled for the stochastic factor. Spitz (1968), moreover, misapplies statistics, treating replicate samples as if they represented replicate experimental units.

The new terms offered have been carefully chosen. Perhaps mathematical statisticians will find them inelegant, but I feel they will be helpful at least to ecologists and perhaps to other persons concerned with experimental design. Statistics and experimental design are disciplines with an impoverished vocabulary. Most of this essay concerns what a statistician might term "randomization," "replication," "independence," or "error term" problems, but these concepts can apply in many ways in an experiment, and they apply in different ways to different kinds of experiments. For example, one often can replicate at several levels (e.g., blocks, experimental units, samples, subsamples, etc.) in the design of an experiment; at many levels the replication may be superfluous or optional, but there is usually at least one level (experimental unit) at which replication is obligatory, at least if significance tests are to be employed. Likewise, the term "error" is used as shorthand for many different quantities or concepts, including: type I and type II errors, random and systematic errors introduced by the experimenter, variation among replicates, variation among samples, the discrepancy between μ and $\bar{x}$, and so on. A slightly enlarged vocabulary, particularly one providing labels for various types of invalid procedures, may make things easier for us.

I begin this discussion at an elementary level, presuming that the reader has had the equivalent of a one-semester course in statistics but no training in experimental design. This approach, and indeed, the whole essay, will seem *too* elementary to some ecologists. But I wish my premises and arguments to be explicit, clear, and easily attacked if in error. Also it is the elementary principles of experimental design, not advanced or esoteric ones, which are most frequently and severely violated by ecologists.

The Experimental Approach

There are five components to an experiment: hypothesis, experimental design, experimental execution, statistical analysis, and interpretation. Clearly the hypothesis is of primary importance, for if it is not, by some criterion, "good," even a well-conducted experiment will be of little value.

By experimental design is meant only "the logical structure of the experiment" (Fisher 1971:2). A full description of the objectives of an experiment should specify the nature of the experimental units to be employed, the number and kinds of treatments (including "control" treatments) to be imposed, and the properties or responses (of the experimental units) that will be measured. Once these have been decided upon, the design of an experiment specifies the manner in which treatments are assigned to the available experimental units, the number of experimental units (replicates) receiving each treatment, the physical arrangement of the experimental units, and often, the temporal sequence in which treatments are applied to and measurements made on the different experimental units.

The execution of an experiment includes all those procedures and operations by which a decided-upon design is actually implemented. Successful execution depends on the experimenter's artistry, insight, and good judgment as much as it does his technical skill. While the immediate goal is simply the conduct of the technical operations of the experiment, successful execution requires that the experimenter avoid introducing systematic error (bias) and minimize random error. If the effects of DDT are being examined, the DDT must not be contaminated with parathion. If the effects of an intertidal predator are being assessed by the use of exclusion cages, the cages must have no *direct* effect on variables in the system other than the predator. If the effects of nutrients on pond plankton are being studied, the plankton must be sampled with a device the efficiency of which is independent of plankton abundance. Systematic error either in the imposition of treatments or in sampling or measurement procedures renders an experiment invalid or inconclusive.

Decisions as to what degree of initial heterogeneity among experimental units is permissible or desirable, and about the extent to which one should attempt to regulate environmental conditions during the experiment, are also a matter of subjective judgment. These decisions will affect the magnitude of random error and therefore the sensitivity of an experiment. They also will influence the specific interpretation of the re-

sults, but they cannot by themselves affect the formal validity of the experiment.

From the foregoing, it is clear that experimental design and experimental execution bear equal responsibility for the validity and sensitivity of an experiment. Yet in a practical sense, execution is a more critical aspect of experimentation than is design. Errors in experimental execution can and usually do intrude at more points in an experiment, come in a greater number of forms, and are often subtler than design errors. Consequently, execution errors generally are more difficult to detect than design errors, both for the experimenter himself and for readers of his reports. It is the insidious effects of such undetected or undetectable errors that make experimental execution so critical. Despite their pre-eminence as a source of problems, execution errors are not considered further here.

In experimental work, the primary function of statistics is to increase the clarity, conciseness, and objectivity with which results are presented and interpreted. Statistical analysis and interpretation are the least critical aspects of experimentation, in that if purely statistical or interpretative errors are made, the data can be reanalyzed. On the other hand, the only complete remedy for design or execution errors is repetition of the experiment.

Mensurative Experiments

Two classes of experiments may be distinguished: mensurative and manipulative. Mensurative experiments involve only the making of measurements at one or more points in space or time; space or time is the only "experimental" variable or "treatment." Tests of significance may or may not be called for. Mensurative experiments usually do not involve the imposition by the experimenter of some external factor(s) on experimental units. If they do involve such an imposition, (e.g., comparison of the responses of high-elevation vs. low-elevation oak trees to experimental defoliation), all experimental units are "treated" identically.

Example 1. We wish to determine how quickly maple (*Acer*) leaves decompose when on a lake bottom in 1 m of water. So we make eight small bags of nylon netting, fill each with maple leaves, and place them in a group at a spot on the 1-m isobath. After 1 mo we retrieve the bags, determine the amount of organic matter lost ("decomposed") from each, and calculate a mean decomposition rate. This procedure is satisfactory as far as it goes. However, it yields no information on how the rate might vary from one point to another along the 1-m isobath; the mean rate we have calculated from our eight leaf bags is a tenuous basis for making generalizations about "the decomposition rate on the 1-m isobath of the lake."

Such a procedure is usually termed an experiment *simply because the measurement procedure is somewhat elaborate,* often involving intervention in or prodding of the system. If we had taken eight temperature measurements or eight dredge samples for invertebrates, few persons would consider those procedures and their results to be "experimental" in any way.

Efforts at semantic reform would be in vain. Historically, "experimental" has always had "difficult," "elaborate," and "interventionist" as among its common meanings, and inevitably will continue to do so. The term *mensurative experiment* may help us keep in mind the distinction between this approach and that of the *manipulative experiment.* As the distinction is basically that between sampling and experimentation sensu stricto, advice on the "design" of mensurative experiments is to be found principally in books such as *Sampling techniques* (Cochran 1963) or *Sampling methods for censuses and surveys* (Yates 1960), and not in books with the word "design" in the title.

Comparative mensurative experiments

Example 2. We wish, using the basic procedure of Example 1, to test whether the decomposition rate of maple leaves differs between the 1-m and the 10-m isobaths. So we set eight leaf bags on the 1-m isobath and another eight bags on the 10-m isobath, wait a month, retrieve them, and obtain our data. Then we apply a statistical test (e.g., t test or U test) to see whether there is a significant difference between decomposition rates at the two locations.

We can call this a *comparative mensurative experiment.* Though we use two isobaths (or "treatments") and a significance test, we still have not performed a true or manipulative experiment. We are simply measuring a property of the system at two points within it and asking whether there is a real difference ("treatment effect") between them.

To achieve our vaguely worded purpose in Example 1, perhaps any sort of distribution of the eight bags on the 1-m isobath was sufficient. In Example 2, however, we have indicated our goal to be a comparison of the two isobaths with respect to decomposition rate of maple leaves. Thus we cannot place our bags at a single location on each isobath. That would not give us any information on variability in decomposition rate from one point to another along each isobath. We require such information before we can validly apply inferential statistics to test our null hypothesis that the rate will be the same on the two isobaths. So on each isobath we must disperse our leaf bags in some suitable fashion. There are many ways we could do this. Locations along each isobath ideally should be picked at random, but bags could be placed individually (eight locations), in groups of two each (four locations), or in groups of four each (two locations). Furthermore, we might decide that it was sufficient to work only with the isobaths along one side of the lake, etc.

Assuring that the replicate samples or measurements are dispersed in space (or time) in a manner appropriate

to the specific hypothesis being tested is the most critical aspect of the design of a mensurative experiment.

Pseudoreplication in mensurative experiments

Example 3. Out of laziness, we place all eight bags at a single spot on each isobath. It will still be legitimate to apply a significance test to the resultant data. However, and the point is the central one of this essay, if a significant difference is detected, this constitutes evidence only for a difference between two (point) *locations*; one "happens to be" a spot on the 1-m isobath, and the second "happens to be" a spot on the 10-m isobath. Such a significant difference cannot legitimately be interpreted as demonstrating a difference between the two *isobaths*, i.e., as evidence of a "treatment effect." For all we know, such an observed significant difference is no greater than we would have found if the two sets of eight bags had been placed at two locations on the *same* isobath.

If we insist on interpreting a significant difference in Example 3 as a "treatment effect" or real difference between isobaths, then we are committing what I term *pseudoreplication*. Pseudoreplication may be defined, in analysis of variance terminology, as the testing for treatment effects with an error term inappropriate to the hypothesis being considered. In Example 3 an error term based on eight bags at one location was inappropriate. In mensurative experiments generally, pseudoreplication is often a consequence of the actual physical space over which samples are taken or measurements made being smaller or more restricted than the inference space implicit in the hypothesis being tested. In manipulative experiments, pseudoreplication most commonly results from use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent. Pseudoreplication thus refers not to a problem in experimental design (or sampling) per se but rather to a particular combination of experimental design (or sampling) and statistical analysis which is inappropriate for testing the hypothesis of interest.

The phenomenon of pseudoreplication is widespread in the literature on both mensurative and manipulative experiments. It can appear in many guises. The remainder of this article deals with pseudoreplication in manipulative experiments and related matters.

Manipulative Experiments

More on terminology

Whereas a mensurative experiment may consist of a single treatment (Example 1), a manipulative experiment always involves two or more treatments, and has as its goal the making of one or more comparisons. The defining feature of a manipulative experiment is that the different experimental units receive different treatments and that the assignment of treatments to experimental units is or can be randomized. Note that in Example 2 the experimental units are not the bags of leaves, which are more accurately regarded only as measuring instruments, but rather the eight physical locations where the bags are placed.

Following Anscombe (1948), many statisticians use the term *comparative experiment* for what I am calling *manipulative experiment* and *absolute experiment* for what I am calling *mensurative experiment*. I feel Anscombe's terminology is misleading. It obscures the fact that comparisons also are the goal of many mensurative experiments (e.g., Example 2).

Cox (1958:92–93) draws a distinction between *treatment factors* and *classification factors* that at first glance seems to parallel the distinction between mensurative and manipulative experiments. However it does not. For Cox, "species" would always be a classification factor, because "species is an intrinsic property of the unit and not something assigned to it by the experimenter." Yet "species," like many other types of classification factors, clearly can be the treatment variable in either a mensurative or a manipulative experiment. Testing the effects of a fire retardant on two types of wood (Cox's example 6.3, simplified) or comparing decomposition rates of oak and maple leaves (my Example 5) represent manipulative experiments, with species being the treatment variable, and with randomized assignment of treatments to experimental units (=physical locations) being possible. However, to measure and compare the photosynthetic rates of naturally established oak and maple trees in a forest would be to conduct a mensurative experiment. Randomized assignment of the two tree species to locations would not be possible.

Cox's (1958) distinction of treatment factors vs. classification factors is a valid one. But because it does not coincide with any dichotomy in experimental design or statistical procedures, it is less critical than the mensurative-manipulative classification proposed here.

Critical features of a controlled experiment

Manipulative experimentation is subject to several classes of potential problems. In Table 1 I have listed these as "sources of confusion"; an experiment is successful to the extent that these factors are prevented from rendering its results inconclusive or ambiguous. It is the task of experimental design to reduce or eliminate the influence of those sources numbered 1 through 6. For each potential source there are listed the one or more features of experimental *design* that will accomplish this reduction. Most of these features are obligatory. Refinements in the *execution* of an experiment may further reduce these sources of confusion. However, such refinements cannot substitute for the critical

features of experimental design: controls, replication, randomization, and interspersion.

One can always *assume* that certain sources of confusion are not operative and simplify experimental design and procedures accordingly. This saves much work. However, the essence of a controlled experiment is that the validity of its conclusions is *not* contingent on the concordance of such assumptions with reality.

Against the last source of confusion listed (Table 1), experimental design can offer no defense. The meaning of demonic and nondemonic intrusion will be clarified shortly.

Controls.—"Control" is another of those unfortunate terms having several meanings even within the context of experimental design. In Table 1, I use control in the most conventional sense, i.e., any treatment against which one or more other treatments is to be compared. It may be an "untreated" treatment (no imposition of an experimental variable), a "procedural" treatment (as when mice injected with saline solution are used as controls for mice injected with saline solution plus a drug), or simply a different treatment.

At least in experimentation with biological systems, controls are required primarily because biological systems exhibit temporal change. If we could be absolutely certain that a given system would be constant in its properties, over time, in the absence of an experimentally imposed treatment, then a separate control treatment would be unnecessary. Measurements on an experimental unit prior to treatment could serve as controls for measurements on the experimental unit following treatment.

In many kinds of experiments, control treatments have a second function: to allow separation of the effects of different aspects of the experimental procedure. Thus, in the mouse example above, the "saline solution only" treatment would seem to be an obligatory control. Additional controls, such as "needle insertion only" and "no treatment" may be useful in some circumstances.

A broader and perhaps more useful (though less conventional) definition of "control" would include *all* the obligatory design features listed beside "Sources of confusion" numbers 1–6 (Table 1). "Controls" (sensu stricto) *control* for temporal change and procedure effects. Randomization *controls* for (i.e., reduces or eliminates) potential experimenter bias in the assignment of experimental units to treatments and in the carrying out of other procedures. Replication *controls* for the stochastic factor, i.e., among-replicates variability inherent in the experimental material or introduced by the experimenter or arising from nondemonic intrusion. Interspersion *controls* for regular spatial variation in properties of the experimental units, whether this represents an initial condition or a consequence of nondemonic intrusion.

In this context it seems perfectly accurate to state that, for example, an experiment lacking replication is also an uncontrolled experiment; it is not controlled for the stochastic factor. The custom of referring to replication and control as separate aspects of experimental design is so well established, however, that "control" will be used hereafter only in this narrower, conventional sense.

TABLE 1. Potential sources of confusion in an experiment and means for minimizing their effect.

Source of confusion	Features of an experimental design that reduce or eliminate confusion
1. Temporal change	Control treatments
2. Procedure effects	Control treatments
3. Experimenter bias	Randomized assignment of experimental units to treatments Randomization in conduct of other procedures "Blind" procedures*
4. Experimenter-generated variability (random error)	Replication of treatments
5. Initial or inherent variability among experimental units	Replication of treatments Interspersion of treatments Concomitant observations
6. Nondemonic intrusion†	Replication of treatments Interspersion of treatments
7. Demonic intrusion	Eternal vigilance, exorcism, human sacrifices, etc.

* Usually employed only where measurement involves a large subjective element.

† Nondemonic intrusion is defined as the impingement of chance events on an experiment in progress.

A third meaning of control in experimental contexts is regulation of the conditions under which the experiment is conducted. It may refer to the homogeneity of experimental units, to the precision of particular treatment procedures, or, most often, to the regulation of the physical environment in which the experiment is conducted. Thus some investigators would speak of an experiment conducted with inbred white mice in the laboratory at 25° ± 1°C as being "better controlled" or "more highly controlled" than an experiment conducted with wild mice in a field where temperature fluctuated between 15° and 30°. This is unfortunate usage, for the adequacy of the true controls (i.e., *control treatments*) in an experiment is independent of the degree to which the physical conditions are restricted or regulated. Nor is the validity of the experiment affected by such regulation. Nor are the results of statistical analysis modified by it; if there are no design or statistical errors, the confidence with which we can reject the null hypothesis is indicated by the value of P alone. These facts are little understood by many laboratory scientists.

This third meaning of control undoubtedly derives in part from misinterpretation of the ancient but ambiguous dictum, "Hold constant all variables except the one of interest." This refers not to temporal con-

stancy, which is of no general value, but only to the desired identity of experimental and control systems in all respects except the treatment variable and its effects.

Replication, randomization, and independence.—Replication and randomization both have two functions in an experiment: they improve estimation and they permit testing. Only their roles in estimation are implied in Table 1. Replication reduces the effects of "noise" or random variation or error, thereby increasing the *precision* of an estimate of, e.g., the mean of a treatment or the difference between two treatments. Randomization eliminates possible bias on the part of the experimenter, thereby increasing the *accuracy* of such estimates.

With respect to testing, the "main purpose [of replication], which there is no alternative method of achieving, is to supply an estimate of error [i.e., variability] by which the significance of these comparisons is to be judged . . . [and] the purpose of randomization . . . is to guarantee the validity of the test of significance, this test being based on an estimate of error made possible by replication" (Fisher 1971:63–64).

In exactly what way does randomized assignment of treatments to experimental units confer "validity" on an experiment? A clear, concise answer is not frequently found. It guarantees "much more than merely that the experiment is unbiased" (Fisher 1971:43), though that is important. It guarantees that, on the average, "errors" are independently distributed, that "pairs of plots treated alike are* not nearer together or further apart than, or in any other relevant way distinguishable from pairs of plots treated differently "except insofar as there is a treatment effect (Fisher 1926:506). (*In her paraphrase of this statement, Box [1978:146] inserts at this point the very important qualifier, "on the average.")

In operational terms, a lack of independence of errors prohibits us from knowing α, the probability of a type I error. In going through the mechanics of a significance test, we may specify, for example, that $\alpha = 0.05$ and look up the corresponding critical value of the appropriate test criterion (e.g., t or F). However, if errors are not independent, then true α is probably higher or lower than 0.05, but in any case unknown. Thus interpretation of the statistical analysis becomes rather subjective.

Demonic and nondemonic intrustion.—If you worked in areas inhabited by demons you would be in trouble regardless of the perfection of your experimental designs. If a demon chose to "do something" to each experimental unit in treatment A but to no experimental unit in treatment B, and if his/her/its visit went undetected, the results would be misleading. One might also classify the consequences of certain design or execution errors as demonic intrusion. For example, if effects of fox predation are studied using fenced and unfenced fields, hawks may be attracted to the fence posts and use them as perches from which to search for prey. Later, foxes may get credit for treatment effects generated in the fenced fields by the hawks. Whether such non-malevolent entities are regarded as demons or whether one simply attributes the problem to the experimenter's lack of foresight and the inadequacy of procedural controls is a subjective matter. It will depend on whether we believe that a reasonably thoughtful experimenter should have been able to foresee the intrusion and taken steps to forestall it.

By nondemonic intrusion is meant the impingment of chance events on an experiment in progress. This sort of intrusion occurs in *all* experimental work, adding to the "noise" in the data. Most of the time the effect of any single chance event is immeasurably slight. However, by definition, the nature, magnitude, and frequency of such chance events are not predictable, nor are their effects. If an event impinges on all experimental units of all treatments there is no problem. Every change in weather during a field experiment would represent such a "chance" event. Potentially more troublesome are chance events that affect only one or a few experimental units. An experimental animal may die, a contamination event may occur or a heating system may malfunction. Some chance events may be detected, but most will not be. Experimenters usually strive to minimize the occurrence of chance events because they reduce the power of an experiment to detect real treatment effects. However, it is also important to minimize the probability of concluding there is a treatment effect when there is not one. Replication and interspersion of treatments provide the best insurance against chance events producing such spurious treatment effects (Table 1).

Interspersion of Treatments

By their very nature, the "treatments" in a mensurative experiment (Example 2) usually are *isolated* from each other in space and/or time. In contrast, treatments in a manipulative experiment always must be *interspersed* with each other in space and time. This interspersion/isolation criterion is the principal operational distinction between the two types of experiments.

In many, perhaps most kinds of manipulative experiments, adequate interspersion of treatments results more or less automatically when experimental units are assigned to treatments by randomization procedures. However, in some ways, interspersion is the more critical concept or feature; randomization is simply a way of achieving interspersion in a way that eliminates the possibility of bias and allows accurate specification of the probability of a type I error. Also, for preliminary assessment of the adequacy of experimental designs, interspersion is a more practical criterion than is randomization. The latter refers only to the process, but the former suggests what the physical layout of the experiment should look like, roughly how the experimental units should be distributed in space.

Example 4. We return to our 1-m isobath to test whether oak (*Quercus*) leaves will decompose more rapidly than will maple (*Acer*) leaves at that depth. This will be a manipulative experiment, though our operations in the field will be very similar to those of our earlier mensurative experiments (Examples 2, 3). Now we are actually altering a single variable (species) and not just comparing a system property at two points in space or time.

We place eight bags of maple leaves at random within a 0.5-m^2 plot (A) on the 1-m isobath and eight bags of oak leaves at random within a second "identical" plot (B) contiguous to the first one. Because the treatments are segregated and not interspersed, this is an uninteresting experiment. The only hypothesis tested by it is that maple leaves at location A decay at a different rate than do oak leaves at location B. The supposed "identicalness" of the two plots almost certainly does not exist, and the experiment is not controlled for the possibility that the seemingly small initial dissimilarities between the two plots will have an influence on decomposition rate. Nor is it controlled for the possibility of nondemonic intrusion, i.e., the possibility that an uncontrolled extraneous influence or chance event during the experiment could increase the dissimilarity of the two plots.

Example 5. We use eight leaf bags for each species and distribute them all at random within the *same* plot on the 1-m isobath. This experiment will allow us validly to test whether the two species decompose at the same rate at this location. If our interest is primarily in a comparison of the two species, we may feel this experiment is sufficient, and it is. However, if it is important to us to state how the two species' rates compare *on the 1-m isobath,* then we should carry out an experiment in which both sets of leaves are dispersed over two or more randomly selected points on the 1-m isobath. Also, if we wish to generalize to the 1-m isobaths of a certain class of lakes, obviously two sets of leaf bags must be distributed in some randomized fashion over all or a random sample of these lakes. The appropriate dispersion of replicates is as important in manipulative as in mensurative experiments.

Modes of spatial interspersion and segregation

Fig. 1 illustrates schematically three acceptable ways and four (*not* five; B-4 is equivalent to A-1, with respect to the interspersion criterion) unacceptable ways of interspersing treatments in a two-treatment experiment. The boxes or experimental units could be aquaria on a laboratory bench, a string of ponds, or a row of plots, with either real (structural) or imaginary boundaries, in a field or in the intertidal zone. Each unit is assumed to have been treated (fish introduced, insecticide applied, starfish removed) independent of the other units in the same treatment.

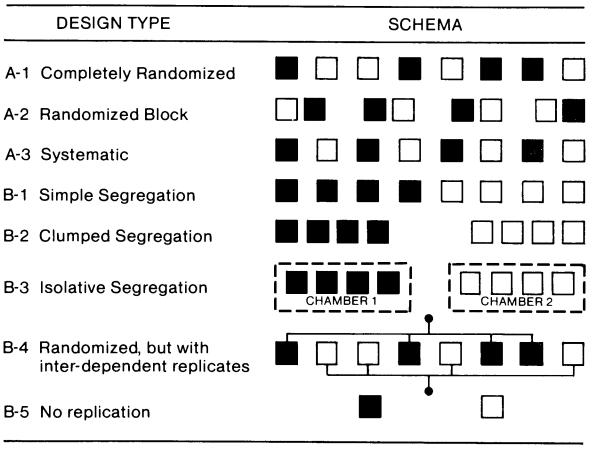

FIG. 1. Schematic representation of various acceptable modes (A) of interspersing the replicates (boxes) of two treatments (shaded, unshaded) and various ways (B) in which the principle of interspersion can be violated.

A few comments are now offered concerning each design illustrated in Fig. 1.

Completely randomized design (A-1).—Simple randomization is the most basic and straightforward way of assigning treatments to experimental units. However, it is not frequently employed in ecological field experiments, at least not when the experimental units are large (ponds, 1-ha plots, etc.). In these cases there usually are available only a few experimental units per treatment, replication as great as four-fold being uncommon. In that circumstance, a completely random assignment process has a good chance of producing treatments which are segregated rather than spatially interspersed. For example, the chances of the random numbers table giving us simple segregation (B-1 in Fig. 1) are $\approx$ 3% when there is four-fold replication and 10% when there is three-fold replication. I strongly disagree with the suggestion (Cox 1958:71; Cochran and Cox 1957:96) that the completely randomized design may be most appropriate in "small experiments." Clearly we cannot count on randomization always giving us layouts as "good" as A-1 (Fig. 1).

Few examples of strict randomization leading to inadequate interspersion of treatments are found in the ecological literature. Perhaps experimental ecologists fall primarily into two groups: those who do not see the need for any interspersion, and those who do recognize its importance and take whatever measures are necessary to achieve a good dose of it. In Fig. 2 are shown three actual experimental layouts in which the degree of interspersion seems unsatisfactory. Fig. 2-I is the only example I have found of poor interspersion having resulted from clearly specified and formally correct randomization procedures. And even in this case, the experimental layout is only that of one block in a four-block randomized complete block design. For the other two experiments (Fig. 2-II, III) the authors did

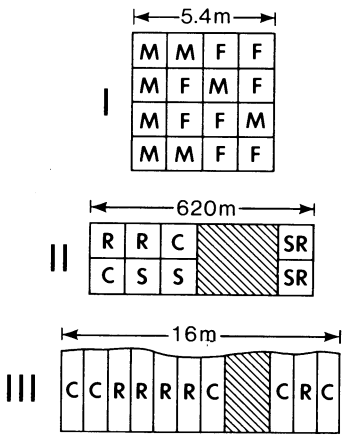

FIG. 2. Three experimental layouts exhibiting partial but inadequate interspersion of treatments. (I) test to compare predation rates on male (M) vs. female (F) floral parts placed on forest floor (Cox 1981, 1982); (II) test of effects on dispersal of removing from unfenced field plots one (S, R), both (SR), or neither (C) of two rodent species (Joule and Cameron 1975); (III) test to compare effects on algae, of removing grazers (R) vs. not doing so (Slocum 1980); shading represents unused portion of study areas.

not indicate what procedures or criteria were used in assigning experimental plots to treatments. In any event, it would not be unusual for such segregated layouts to result from random assignment. The potential for pre-existing gradients or nondemonic intrusion to produce spurious treatment effects was high in all three cases.

Randomized block design (A-2).—This is a commonly used design in ecological field experiments, and it is a very good one. In the example, four blocks were defined, consisting of two plots each, and each treatment was randomly assigned to one plot in each block. Like other modes of "restricted randomization," a randomized block design reduces the above-mentioned probability of chance segregation of treatments. And it helps prevent pre-existing gradients and nondemonic intrusion from obscuring real effects of treatments or from generating spurious ones. As insurance against non-demonic intrusion, blocking or some other procedure which guarantees interspersion is *always* highly desirable. It should *not* be regarded as a technique appropriate only to situations where a premanipulation gradient in properties of experimental units is known or suspected to exist.

This design has one disadvantage if the results are to be analyzed with nonparametric statistics. A minimum of six-fold replication is necessary before significant ($P \leq .05$) differences can be demonstrated by Wilcoxon's signed-ranks test (the appropriate one for design A-2), whereas only four-fold replication is necessary before significant differences can be demonstrated by the Mann-Whitney U test (the appropriate one for design A-1). However, there is probably nothing wrong, at least in a practical sense, in applying a U test to data from an experiment of design A-2; doing so should not increase our chances of generating a spurious treatment effect (i.e., of raising the probability of a type I error)—and that is probably the best single criterion for assessing the validity of such a hybrid approach.

Systematic Design (A-3).—This achieves a very regular interspersion of treatments but runs the risk that the spacing interval coincides with the period of some periodically varying property of the experimental area. That risk is very small in most field situations.

An example where a systematic design seemed definitely preferable to a randomized one concerns an experiment on the effects of flamingo grazing on lacustrine microbenthos (Hurlbert and Chang 1983). Four exclosures were established in a linear arrangement with equal spacing between them and with 10 control areas interspersed systematically among and around them. Our rationale was that the flamingos might be shy of the exclosure fences, in which case the variability in the distance between exclosures would have led to increased variability among control areas in their use by flamingos. In our statistical analysis, we employed a procedure (Mann-Whitney U test) strictly appropriate only for a completely randomized design.

In both systematic and randomized block designs, we can base the assignment process not on the locations of the experimental units but rather on their internal properties prior to imposition of treatments. If our study concerns soil mites, for example, we could rank experimental plots on the basis of premanipulation total soil mite densities, assigning odd-ranked plots to one treatment and even-ranked plots to the other. In this process, ideally we would use premanipulation mite densities that were averages based on two or more premanipulation sampling dates.

The danger of basing the assignment process on internal properties rather than on location is that we run a risk of ending up with spatially segregated treatments (e.g., B-1), just as we run this risk with a completely randomized design. Again, the magnitude of this risk decreases as the number of replicates per treatment increases.

A combined or hybrid approach is to consider *both* location and premanipulation internal properties of units, and to assign treatments to units in an essentially subjective manner. The goal would be to achieve spatial interspersion *and* minimization of premanipulation differences between treatment means and equalization of premanipulation variability among replicate units (within treatments). We have employed this approach in studies of the effects of an insecticide (Hurlbert et al. 1972) and of fish on plankton populations (Hurlbert and Mulla 1981). In the latter experiment there were (initially) three treatments (0, 50, and 450 fish per pond), limited and unequal replication (5, 4, and 3 ponds per treatment), and marked premanipulation variability among ponds. The unequal replica-

tion reflected our judgment that postmanipulation among-pond variability in plankton populations would be inversely related to fish density. Given these circumstances, it is hard to imagine that some other way of assigning treatments would have been preferable to the hybrid approach taken.

Simple and clumped segregation (B-1, 2).—These types of design are rarely employed in ecological field experiments. Vossbrinck et al. (1979), Rausher and Feeny (1980), and Warwick et al. (1982) provide three examples. Presumably persons perceptive enough to see the need for physically independent replicates also will recognize the need for treatment interspersion. Treatment segregation is much more commonly found in laboratory experiments.

The danger of treatment segregation of any sort is that it very easily leads to spurious treatment effects, i.e., to type I error. Such effects can result from either or both of two causes. First, differences between "locations" of the two treatments may exist prior to the carrying out of the experiment; in theory these could be measured, but that requires both effort and knowledge of what to measure. Second, as a result of nondemonic intrusion, differences between "locations" can arise or become greater *during* the experiment independently of any true treatment effect.

Example 6. To test the effects of DDT on phytoplankton populations, we set up eight plankton-containing aquaria on a laboratory bench and apply DDT to the four tanks on the left, keeping the other four as controls. It is relatively easy to establish initial conditions that are extremely similar from one aquarium to another and we do so. This includes assuring the equivalence of inocula, light conditions, etc., for all aquaria.

In such an experiment, the most likely source of spurious treatment effects would be events that occur after the experimental systems are established. For example, a light bulb at one end of the bench may dim, producing a light gradient along the bench unperceived by us. A spurious effect could easily result. Or the bulb might fail altogether but not be detected until 48 h later. If our wits are improving we will replace the bulb, throw the whole experiment out, and start over again with a better design. Otherwise a spurious treatment effect is highly probable.

Example 7. Another possibility: someone leaves an uncapped bottle of formaldehyde on one end of the bench for an entire afternoon, creating a gradient of formaldehyde fumes along the bench. We do not find out. What we *do* "find out" is that DDT stimulates phytoplankton photosynthesis, because the formaldehyde bottle had been left near the "control" end of the bench!

In this example, and in many laboratory experiments, treatment interspersion is not very necessary or critical as a means of assuring that initial conditions for the two treatments are, on average, quite similar. It *is* critical, however, as a control for nondemonic intrusion, for differential impingement of chance events during the experiment. If DDT and control aquaria had been reasonably interspersed, then the light bulb failure or a formaldehyde gradient would have had little or no effect on the difference between treatment means, but probably they would have increased markedly the variance among aquaria in each treatment. This by itself would have precluded spurious treatment effects and also made the detection of any true treatment effect more difficult.

Example 8. We repeat our DDT-plankton experiment, this time conducting it in experimental ponds with treatments again arranged in simple segregated fashion (B-1). Here, as in many field experiments, segregation poses a double danger. The experiment is controlled neither for possible preexisting locational differences (e.g., a gradient in soil type) nor for the possibility of locational differences arising during the experiment (e.g., if one end of the row of ponds is closer to a woods, ponds at that end may be more heavily utilized for breeding by amphibians; ponds upwind might receive more debris during a windstorm than would ponds downwind).

Isolative segregation (B-3).—Isolative segregation is a common design in laboratory experiments, but one rarely used by field ecologists. It poses all the dangers of simple segregation but in more extreme form, and spurious treatment effects are much more likely to occur. Studies of temperature effects commonly use constant-temperature rooms, growth chambers, or incubators. These are expensive, usually limited in number, and often shared by many workers. Though two such chambers might be *considered* to be identical except for one being at 10°C and the other at 25°, they in fact usually must differ in many other characteristics (lighting, volatile organics, etc.) despite efforts to prevent this.

Studies of fish physiology and growth often use a single tank, containing a fixed number of fish, for each experimental treatment (temperature, food level, etc.). In the sense that the individual fish are the units of direct interest, such experiments may be viewed as exemplifying isolative segregation of treatments (design B-3). In the sense that the tanks are the units directly manipulated or treated, such experiments may be viewed as simply *lacking* replicated treatments (design B-5).

The increased likelihood of spurious treatment effects with isolative segregation of treatments is illustrated by again considering the effect of a chance formaldehyde spill. In Example 7, a spurious treatment effect requires the somewhat improbable circumstance that a marked concentration gradient of formaldehyde persists in the air along the row of aquaria for an effectively long period of time despite normal air turbulence in the room. In our new examples, however, a small spill of formaldehyde on the floor of one constant-temper-

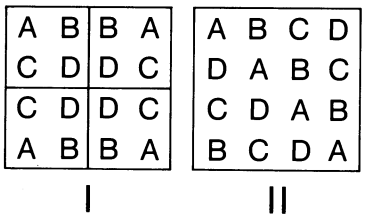

FIG. 3. Examples of segregated arrangements of four treatments, each replicated four times, that can result from use of restricted randomization procedures: (I) randomized block design, (II) Latin square design.

ature room or in one fish tank *guarantees* differential exposure of treatments to this extraneous variable. Moreover, the replicates of the contaminated treatment may be more equally exposed than are the replicates in Example 7. This will further increase the likelihood of a spurious treatment effect, as within-treatment variances are less likely to be increased.

Physically interdependent replicates (B-4).—So far we have focused on spatial interspersion as a way of achieving and assuring statistical independence. This will not always be sufficient. Design B-4 (Fig. 1) shows an arrangement which could represent two sets of aquaria, where the four aquaria in each set share a common heating, aeration, filtration, circulation, or nutrient supply system. Though meeting the interspersion requirement, such a design is no better than the isolative segregation. It is subject to the same easy generation of spurious treatment effects. For experiments involving such systems, each replicate should have its own independent maintenance systems. In that way a single chance motor failure, contamination event, or other kind of nondemonic intrusion will only affect a single experimental unit and be unlikely to produce a "treatment effect." Equally satisfactory would be to have, when possible, all experimental units of all treatments hooked up to the same maintenance system.

Randomization vs. interspersion

From the foregoing it is apparent that there is often a conflict between the desirability of using randomization procedures and the desirability of having treatments interspersed. Randomization procedures sometimes produce layouts with treatments markedly segregated from each other in space, especially when replication is low and a completely random design is employed. Designs (randomized block, Latin square) employing restricted randomization reduce the possibility of getting extremely segregated layouts, but still allow degrees of segregation unacceptable to thoughtful experimenters (Fig. 3).

Cox (1958:85–90) discusses three possible solutions to this problem. Of these, the simplest and most widely useful is the second: simply reject highly segregated layouts when they arise, and "rerandomize" until a layout with an acceptable degree of interspersion is obtained. Ideally, the criterion or criteria of acceptability are specified beforehand. This procedure leads to designs which, on the average, are more interspersed (or systematic or balanced) than those obtained by strict randomization procedures. But the procedure also precludes our knowing the exact value of α, the probability of a type I error. For that reason, this solution would have been anathema to Fisher. For him, the exact specification of α was the sine qua non of proper experimental design. His hard-nosed rejection of any departure from strict randomization procedures, and of systematic designs in particular (Barbacki and Fisher 1936, Fisher 1971:64–65, 76–80), was an attitude that was passed on to his followers and that has set the tone of the literature on the topic. It was not an entirely rational attitude, however; interspersion, systematic or otherwise, merits more weight, vis-a-vis randomization, than he gave it.

A historical perspective.—To understand Fisher's attitude and its consequences, history is as important as mathematics. The notion of randomization was Fisher's "great contribution to the scientific method" (Kempthorne 1979:121) and he knew it. Yet W. S. Gossett ("Student"), his mentor and friend, and one of the other giants in the history of statistics, never fully accepted Fisher's arguments in favor of strict randomization. Worse yet, Gossett argued that systematic designs were superior. They corresponded on the matter, off and on, for 13 yr, and publicly argued the subject at the Royal Statistical Society (e.g., Gossett 1936). But to the end, Gossett "stood his ground against Fisher and left him seething with rage" (Box 1978:269). Traces of that rage passed, I think, into Fisher's writings. Though certain as to the correctness of his own ideas, he undoubtedly felt defensive with respect not only to Gossett but also to the large number of older agricultural experimenters who were inclined to use systematic designs.

Gossett's (1937) clearest defense of systematic designs was written during his last year of life and published after his death. His basic arguments (pp. 363–367) seem irrefutable. Yates (1939) responded at length and in moderate tones, admitting several of Gossett's points but in general adhering to the Fisherian view. Fisher (1939:7) never really responded except to comment that Gossett's failure to "appreciate the necessity of randomization . . . was perhaps only a sign of loyalty to colleagues whose work was in this respect open to criticism."

It was unfortunate that Gossett could not have lived to resolve this controversy, because there was no one to fill his shoes in the debate. If he and Fisher had been able to focus on fundamentals (many of their arguments concerned a specific agricultural technique called the "half-drill strip method"), more common ground might have been found. But it also may have been inevitable that the Fisherian view on systematic or

TABLE 2. Comparison of some properties of pre-layout alpha (α_{PL}) and layout-specific alpha (α_{LS}).

α	Applies to	Exactly knowable or specifiable?	Affected by assignment procedure?	Affected by the nature of variation among experimental units?
α_{PL}	The general procedure; the average for all possible layouts	Yes*	Yes†	No
α_{LS}	The one specific layout being used	No	No	Yes

* Only on the assumption that randomization procedures are employed wherever appropriate.
† In that it can be specified only if randomization procedures are employed wherever appropriate.

balanced designs prevailed. Fisher not only outlived Gossett by a quarter of a century, but out-published him (more than 300 articles, plus seven books, to Gossett's 22 articles) and had a tremendous direct influence as a teacher, consultant and adviser of agricultural and other scientists throughout the world. Gossett's position as statistician and brewer for the Guinness breweries was a much more modest podium.

There is no question that Fisher recognized the importance of interspersion for minimizing bias and the possibility of spurious treatment effects (see: Fisher 1926:506, 1971:43). Almost all his work in experimental design was focused on those techniques employing restricted randomization, which not only guarantee *some* degree of interspersion but also often increased the power of experiments to detect treatment effects. Fisher differed from Gossett primarily in stipulating that interspersion was a secondary concern and should never be pursued at the expense of an exact knowledge of α.

To judge this controversy further, we must ask how important it is to know the value of α precisely. If we do know it, what do we know? If we sacrifice knowledge of it, what have we given up?

Prelayout and layout-specific alpha.—Clarity is served by distinguishing two alphas, which I will call *prelayout alpha* (α_{PL}) and *layout-specific alpha* (α_{LS}). They are contrasted in Table 2. The distinction was clearly made by Gossett (1937:367) and presumably is widely understood by statisticians.

α_{PL} is the conventional alpha, the one Fisher and other statisticians have been most concerned about, the one that the experimenter usually specifies. It is the probability, *averaged over all possible layouts of a given experiment,* of making a type I error, i.e., of concluding there is a treatment effect when in fact there is not one. In more symbolic form,

$$\alpha_{PL} = \frac{\sum \alpha_{LS}}{\text{All possible layouts}} \Big/ \frac{}{\text{Number of possible layouts}}$$

Once a specific experimental layout has been selected and treatments assigned to experimental units, one can define α_{LS}, the probability of making a type I error *if that layout is used.* Since a given experiment is usually performed only once, using a single layout, α_{LS} is of much greater interest to experimenters than is α_{PL}. Usually α_{LS} will be less than or greater than α_{PL}. For example, if spatial gradients in influential variables exist across the row or grid of experimental units, α_{LS} will usually be lower than α_{PL} when treatments are well interspersed and higher than α_{PL} when treatments are segregated to some degree.

The problem is that α_{LS} cannot be known or specified exactly. This is true whether the particular layout has been obtained through randomization methods or not. Thus, experimenters must fall back on α_{PL} as the only objective way of specifying acceptable risk, even though α_{PL} may be of marginal relevance to the one experiment actually conducted. This does not mean, however, that if we set $\alpha_{PL} = 0.05$ we must adhere to all the procedures (strict randomization, in particular) necessary for guaranteeing the accuracy of that specification. More exactly, if one opts for a systematic or balanced design as recommended by Gossett (1937), or adopts Cox's (1958) second solution, or achieves interspersion by some more ad hoc approach, the particular experiment is likely to be a better one, with an $\alpha_{LS} < 0.05$. That is, with respect to type I error, the experiment will be conservative.

Cox (1958:88) summarizes the philosophy of this approach succinctly:

> *. . . to adopt arrangements that we suspect are bad, simply because things will be all right in the long run, is to force our behavior into the Procrustean bed of a mathematical theory. Our object is the design of individual experiments that will work well: good long-run properties are concepts that help us in doing this, but the exact fulfillment of long-run mathematical conditions is not the ultimate aim.*

Is it more useful (1) to know that the chosen value of α represents a probable upper bound to α_{LS}, or (2) to know that it equals α_{PL} exactly and have little idea as to what the upper bound of α_{LS} may be? Every experimenter must decide for himself.

Biased estimation of treatment effects?—A second classical objection to systematic designs is that "Biases may be introduced into treatment means, owing to the pattern of the systematic arrangement coinciding with some fertility pattern in the field, and this bias may persist over whole groups of experiments owing to the arrangement being the same in all" (Yates 1939:442). This objection would also apply to all designs where

TABLE 3. Categorization of recent (post-1960) ecological field experiments according to type of experimental designs and statistical analyses employed.

Subject matter	Total number of studies (papers)*	Design and analysis category I	II "pseudo-replication"	III	IV
Treatments replicated?		No	No†	Yes	Yes
Inferential statistics applied?		No	Yes	No	Yes
Freshwater plankton	48 (42)	14	5‡ (10%)	15	14
Marine benthos	57 (49)	13	18§ (32%)	15	11
Small mammals	24 (21)	1	12‖ (50%)	2	9
Other topics	47 (46)	6	13¶ (28%)	9	19
Totals	176 (156)	34	48 (27%)	41	53

* If a paper presented two or more experiments and these were assignable to different categories, the paper has *sometimes* been listed under more than one category. Hence the number of studies listed is somewhat greater than the number of papers examined (in parentheses).

† In some studies in this category, treatments were replicated but the manner in which significance tests were employed assumed that replication was of a different sort than it actually was (see section on "sacrificial pseudoreplication"). It also is recognized that there are special cases where treatment effects can be assessed statistically even in the absence of treatment replication, but such cases were not encountered in this survey.

‡ Jones and Moyle (1963), Cowell (1965), Giguere (1979: clutch size), Fry and Osborne (1980), Marshall and Mellinger (1980: ELA experiment).

§ Harger (1971: two cages), Menge (1972), Haven (1973), Paine (1974, 1980: *Katharina, Acmea* experiments), Young et al. (1976), Peterson (1977), Virnstein (1977), Bell and Coull (1978), Reise (1978: part), Rogers (1979: part), Vance (1979), Bell (1980), Hixon (1980), Holland et al. (1980), Lubchenco (1980), Markowitz (1980), Sherman and Coull (1980).

‖ Spitz (1968), Cameron (1977: part), Grant et al. (1977), Price (1978: competitive density), Abramsky et al. (1979), Crowner and Barrett (1979), Dobson (1979), Gaines et al. (1979), Holbrook (1979), Reichman (1979), Spencer and Barrett (1980), Munger and Brown (1981: matched pairs test).

¶ Gilderhus (1966), Clarke and Grant (1968), Cope et al. (1969), Malone (1969), Hurlbert et al. (1971: ducks), Werner and Hall (1976), Bakelaar and Odum (1978), Durbin et al. (1979: litter respiration), McCauley and Briand (1979: "1976 expt."), Vossbrink et al. (1979), Hall et al. (1980), Rausher and Feeny (1980).

ad hoc efforts to achieve good interspersion had produced a marked degree of regularity in the experimental layout. However, though widely repeated in experimental design and statistics textbooks, the objection is without foundation. In small experiments, randomization will often produce systematic or nearly systematic layouts. Do even hardcore Fisherians reject such nicely interspersed "legitimate" layouts because of this slight chance of coinciding periodicities? One expects not. They probably beam with delight, knowing that they're getting the best of both worlds: they can specify α_{PL} and they have good reason to expect that $\alpha_{LS} < \alpha_{PL}$. Furthermore, when some factor does fluctuate in magnitude across an experimental area, it most commonly does so irregularly and not periodically. In that case, the greatest bias in estimating a treatment effect will result from some particular *non*systematic design (or class of such) and not from a systematic one.

Nevertheless, Fisher himself was so zealous that he actually may have preferred the worst of both worlds, rather than concede any of the points above. When asked in 1952 what he would do if randomization produced, by chance, a particular systematic Latin Square design, "Sir Ronald said he thought he would draw again and that, ideally, a theory explicitly excluding regular squares should be developed" (Savage et al. 1962:88). In a talk in 1956, Youden (1972) described a "constrained randomization" procedure in which exact knowledge of α_{PL} is retained by rejecting both highly segregated *and* highly interspersed layouts. In his four-treatment, two replicates per treatment example, Youden thus rejects the following layouts: AABBCCDD, AABBCDCD, ABCDABCD, and ABCDBADC, among others. Possibly this procedure would have been acceptable to Fisher. In any case, the latter two well-interspersed layouts are much less likely to lead to spurious treatment effects than are many of the layouts acceptable to Youden (e.g., ABACCDDB). While one could attempt to minimize such absurdities by modifying Youden's criteria of acceptability, I believe that any approach is undesirable which rejects certain designs a priori because of a perceived "excessive" degree of interspersion or regularity.

As to experiments which are repeated many times or to "whole groups of experiments," it is obviously undesirable to use a particular systematic design over and over, just as it would be undesirable to obtain a single design by randomization and use it over and over. Yet it must be admitted that in practice, particular systematic designs have been used over and over in certain types of work. Usually this has been done not on statistical grounds but rather because they offered some operational convenience. The classic example is the design yielded by the "half drill strip" method of planting grain in two-variety trials (Gossett 1923, Neyman and Pearson 1937). This yielded strips of grain alternating in the manner ABBAABBAABBAAB. The merits and faults of such a layout, used repeatedly, were the focus of much of the debate between Fisher and Gossett.

PSEUDOREPLICATION IN MANIPULATIVE EXPERIMENTS

If treatments are spatially or temporally segregated (B-1, 2, 3), if all replicates of a treatment are somehow interconnected (B-4), or if "replicates" are only samples from a single experimental unit (B-5), then replicates are not independent (Fig. 1). If one uses the data from such experiments to test for treatment effects, then one is committing pseudoreplication. Formally, all the B designs (Fig. 1) are equally invalid and are equivalent to that of Example 4 (above); at best they

can only demonstrate a difference between "locations." Naturally, if we know the precise details of an experiment with a B design, we most likely could find grounds for subjectively appraising whether there was a treatment effect and, if so, how great a one. Common sense, biological knowledge, and intuition should be applied to that task; inferential statistics should not be.

Two literature surveys

To assess the frequency of pseudoreplication in the literature, I examined the experimental designs and statistical analyses of 156 papers reporting the results of manipulative ecological field experiments. These papers represent all the field experiments reported in recent issues of selected journals (*Ecology* 1979, 1980; *American Midland Naturalist* 1977, 1978, 1979, 1980; *Limnology and Oceanography* 1979, 1980; *Journal of Experimental Marine Biology and Ecology* 1980; *Journal of Animal Ecology* 1979, 1980; *Canadian Journal of Fisheries and Aquatic Sciences* 1980 (Number 3 only): *Journal of Mammalogy* 1977, 1978, 1979, 1980), the experiments reported in the volume edited by Kerfoot (1980), and those listed in the bibliographies of several recent papers and reviews (Connell 1974, Hurlbert 1975, Chew 1978, Hayne 1978, Hayward and Phillipson 1979, Paine 1980, Peterson 1980, Virnstein 1980, Hurlbert and Mulla 1981, Munger and Brown 1981). Each paper was placed into one of four categories according to whether or not treatments were replicated and whether or not significance tests were carried out. The results are given in Table 3.

Some papers that were part of the sample (as defined above) were not included in the tabulation because I was unable to obtain them soon enough or because their descriptions of experimental design and statistical procedures were too vague. A couple of papers were included in the tabulation simply because they crossed my desk at the time I was carrying out the survey.

These papers are reasonably regarded as a representative, though not random, sample of the recent literature. Most of the tabulated papers were published in the late 1970s. All papers published before 1960 were excluded from the tabulation.

Three assemblages that have been the subject of much recent field experimentation are the freshwater plankton, the marine intertidal and shallow subtidal benthos, and terrestrial small-mammal (rodent) populations. The experiments on each of these subjects have been tabulated separately and all other studies lumped under "other topics" (Table 3).

The survey suggests that overall $\approx$27% of recent manipulative field experiments have involved pseudoreplication. This represents 48% [=48/(48 + 53)] of all studies applying inferential statistics. These figures are disturbingly high, especially given that the analysis considers only this one class of statistical error.

The distribution of studies among design and analysis categories varies significantly among the three specific subject matter areas ($\chi^2 = 20.5$, df = 6, $P < .005$). Where field experiments confront great logistical difficulties (small mammals), pseudoreplication is not only common but dominant. Where field experiments are easiest (freshwater plankton), pseudoreplication is infrequent. Studies of marine benthos are intermediate in both regards. However, if only those studies employing inferential statistics are considered (categories II and IV), then marine benthologists seem the worst pseudoreplicators (62% of studies), followed by the mammalogists (57%), then the relatively virginal planktologists (26%).

A second survey of the literature was carried out by 11 graduate students in a course on experimental design. Each was instructed to select a topic of interest to them, to find $\approx$50 reports of manipulative experiments on that topic, and to examine them for adequacy of design and statistical analysis. Pseudoreplication was only one of several problems for which they were told to keep their eyes open.

Table 4 shows the frequency with which pseudoreplication was found by the students. Of the 537 reports examined, 12% (62) were blemished by the problem. A large number of these 537 reports used no inferential statistics, and for them, pseudoreplication as I have defined it was not a possibility, of course. Of the 191 reports which described their designs clearly and which used inferential statistics, 26% (50) involved pseudoreplication (data from Gasior, Rehse, and Blua [Table 4] not used in this calculation). The difference between this figure and the 48% obtained in my own survey probably resulted from several factors. Among these would be the fact that the student survey was not restricted to ecological field experiments but included laboratory studies of various sorts as well. The frequent lack of clarity in descriptions of designs and analyses was perhaps more of a hindrance to students than to myself in our detective work. The figure of 26% pseudoreplication may be compared with G. S. Innis's (1979) estimate that $\approx$20% of the papers surveyed by students in his course on quantitative methods contained statistical or calculation errors (based only on those papers giving sufficient details for evaluation). And in a very thorough survey of how analysis of variance has been misused by marine biologists, Underwood (1981) found 78% of 143 papers examined to contain statistical errors of one sort or another.

Simple pseudoreplication

The most common type of "controlled" experiment in field ecology involves a single "replicate" per treatment. This is neither surprising nor bad. Replication is often impossible or undesirable when very large-scale systems (whole lakes, watersheds, rivers, etc.) are studied. When gross effects of a treatment are anticipated, or when only a rough estimate of effect is required, or when the cost of replication is very great,

TABLE 4. Occurrence of pseudoreplication in various segments of the biological journal literature, as determined by several student reviewers.

Subject matter	Journal	Number of reports: Examined	... which adequately described design and used inferential statistics	... and which committed pseudoreplication	Reviewer
Marine field experiments	*Journal of Experimental Marine Biology and Ecology*	50	18	7	J. Johnson
Marine organisms	*Marine Behaviour and Physiology; Biological Bulletin*	44	25	15	M. Chiarappa
Heavy metal effects on marine plankton	Articles in bibliography of Davies (1978)	50	5	1	A. Jones
Temperature effects on fish	Various	50	29	7	T. Foreman
Salt-marsh plants	Various	50	31	4	P. Beare
Temperature-plant relationships	Various	50	11	7	J. Gilardi
Life-history traits of animals	Various	44	38	8	M. Russell
Animal physiology	*Physiological Zoology*	50	?*	7	C. Gasior
Effects of ionizing radiation	*Radiation Research; Health Physics*	50	34	1	J. DeWald
Animal ecology	*Journal of Animal Ecology*	50	?*	2	M. Rehse
Plant-herbivore interactions	Various	49	?*	3	M. Blua
	Totals	537	191+	62	

* The number of studies falling under this heading was not reported.

experiments involving unreplicated treatments may also be the only or best option.

What is objectionable is when the tentative conclusions derived from unreplicated treatments are given an unmerited veneer of rigor by the erroneous application of inferential statistics (e.g., Barrett 1968, Spitz 1968, Malone 1969, Young et al. 1976, Waloff and Richards 1977, Buzas 1978, Bell and Coull 1978, Rogers 1979, Vance 1979, Holland et al. 1980, Sherman and Coull 1980, Spencer and Barrett 1980). In these investigations the "strong similarity," "replicability," or "identicalness" of experimental units prior to manipulation sometimes is assessed by "eyeballing" or by subsampling and measurement. When quantitative data are obtained, tests of significance are usually applied to them, and it usually is found that "no significant difference" exists between the one experimental and one control unit prior to manipulation. This result is used, implicitly, to validate the claim that significant differences found between the two units after manipulation represent a treatment effect. Crowner and Barrett (1979) exemplify this approach.

The validity of using unreplicated treatments depends on the experimental units being *identical* at the time of manipulation and on their remaining *identical* to each other after manipulation, except insofar as there is a treatment effect. The lack of significant differences prior to manipulation cannot be interpreted as evidence of such identicalness. This lack of significance is, in fact, only a consequence of the small number of samples taken from each unit. In any field situation (and probably any laboratory situation as well) we *know*, on first principles, that two experimental units *are* different in probably every measurable property. That is, if we increase the number of samples taken from each unit, and use test criterion (e.g., t) values corresponding to an α of 0.05, our chances of finding a significant premanipulation difference will increase with increasing number of samples per unit. These chances will approach 1.0 as the samples from an experimental unit come to represent the totality of that unit (at least if the finite correction factor is employed in the calculation of standard errors).

The above may be contrasted with the result of increasing the number of independent experimental units per treatment. If treatments are assigned to units in randomized fashion and if we again use test criterion values corresponding to an α of 0.05, then our chances of finding a significant premanipulation difference between treatments remain unchanged at 0.05 *regardless* of the number of experimental units per treatment and the number of subsamples taken per experimental unit. This provides an excellent criterion for distinguishing true replication from pseudoreplication.

Example 9. We have a beetle population distributed over a large field with a true mean density (μ) of 51 beetles/m^2 and a true variance (σ^2) (for a 1-m^2 sampling unit) of 100. We wish to test whether a herbicide has any short-term effect on beetle density; but let us assume that we are omniscient and *know* that, under our

experimental conditions, the herbicide will have no effect on beetles whatsoever. Let us conduct the experiment with two different experimental designs:

1. *Design A.*—The field is divided into two subfields (1 and 2) which are "essentially identical" but which in fact differ slightly in beetle density, with $\mu_1 = 52$, $\mu_2 = 50$, and $\sigma_1^2 = \sigma_2^2 = 64$. A preapplication sampling of both subfields finds no significant difference between them. The herbicide is then applied to one sub-field and the other kept as a control. After 48 h, control and treated field are both sampled again.

2. *Design B.*—The entire field is partitioned into an imaginary grid of 4 × 4 m plots. A certain number (n) of these are selected at random to serve as control plots and an equal number to serve as herbicide plots. A preapplication sampling (let us assume a nondestructive census of a 1-m^2 subplot in each plot) of both sets of experimental plots finds no significant difference between the sets. The herbicide is then applied to one set, and 48 h later both sets are sampled again. (I omit here any consideration of execution problems, e.g., whether plots or subfields should be fenced, etc.).

The essential difference between these two designs can be illustrated by repeating each experiment (design) many times, increasing the number of replicates (i.e., samples in Design A, plots in Design B) each time. Since we know the true parameters (μ, σ^2) of the field situation, we can calculate for each experiment the probability of finding a statistically significant difference, given the number of replicates and assuming application of the t test (Steel and Torrie 1980:113–121). In this example, that probability is the probability of a type I error. The results of such calculations are shown in Fig. 4. In the properly designed experiment (B), α remains at the specified value of 0.05 and is unaffected by n. In the design (A) relying on the "identicalness' or "replicability" of the subfields, α is >0.05 for all n and approaches 1.0 as n becomes very large. This illustrates how pseudoreplication, which is what Design A exemplifies, increases the probability of spurious treatment effects. In other words, with Design A the null hypothesis we are testing is not that of "no herbicide effect" but rather that of "no difference between subfields." The difference between subfields may and, in the example, does exist independently of the herbicide treatment. Thus when we conclude that there is a significant effect of the herbicide, we are making a type I error with respect to the hypothesis of interest ("no herbicide effect"). But with respect to the only hypothesis actually testable with Design A ("no difference between subfields"), statistical significance leads us to *avoid* making a type II error. With Design A the probability of a type I error with respect to the hypothesis of "no herbicide effect" is therefore equal to the probability of *avoiding* a type II error with respect to the hypothesis of "no difference between subfields." It is this latter probability, usually called the "power of the test" and denoted symbolically as 1-β (where β is the probability of a type II error), which has been calculated and plotted for Design A in Fig. 4. (Note: this example should not be construed as recommending repeated t tests as the best approach to analysis of a "Design B" experiment. That approach is used here only for simplicity of illustration and ease of calculation of α.)

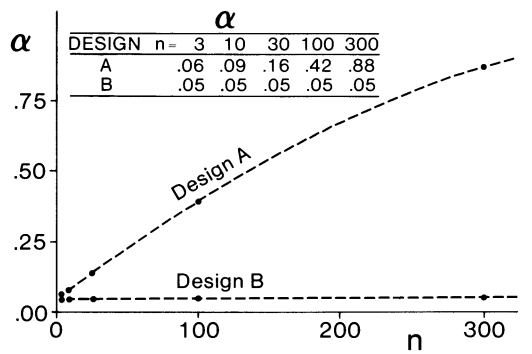

FIG. 4. The relationship between the probability of a type I error (α) and number of replicates (n) for two experimental designs (see text). The α values apply to both the preapplication and postapplication comparisons, since we have specified the herbicide to have no effect.

Multiple samples per experimental unit.—None of the above should be interpreted as arguing against the taking of multiple samples or measurements from each experimental unit. This clearly is often desirable. It increases the sensitivity of the experiment by increasing the precision with which properties of each experimental unit, and hence each treatment, are estimated. However, multiple samples per experimental unit *do not* increase the number of degrees of freedom available for testing for a treatment effect. In such tests, the simplest and least error-prone approach usually is to use only a single datum (mean of the samples) for each experimental unit and to omit completely any formal analysis of the data for individual samples and subsamples. Fancier approaches, e.g., nested analyses of variance, will not be any more powerful in detecting treatment effects, but will be more susceptible to calculation and interpretation error.

Replicability: a red herring.—The confusing notion of *replicability* is a major contributor to the popularity of simple pseudoreplication. The idea is that replicate experimental units must be extremely similar if not identical at the beginning (premanipulation period) of an experiment. Such a view usually reflects a presumption or prior decision that treatments are not going to be replicated, i.e., it indicates a lack of understanding of the basics of experimental design. Replicability has also been called "reproducibility" (Abbott 1966), and the desired state of similarity has been called "close duplication" (Abbott 1966) and even "replication" (Takahashi et al. 1975, Grice et al. 1977, Menzel 1977, Menzel and Case 1977), in disregard for the conventional statistical meaning of the latter term.

TABLE 5. Variability among replicate microcosms, as observed in various investigations.

Study	Variable	Number of microcosms	Range	Coefficient of variation [100($s/\bar{x}$)]	Standard deviation (s)
Abbott 1966	Community respiration	18	2.02–5.21	32	4.78
	Gross production	18	2.11–3.43	14	2.88
	Re-aeration constant	18	1.13–0.12	172	0.54
	Nitrate	18	8.0–16.8	22	2.49
	Nitrite	18	0.19–0.26	11	0.024
	Orthophosphate	18	0.16–1.30	74	0.36
	Gross production	12	1.97–3.13	14	0.367
Beyers 1963	Community respiration	12	1.86–3.02	14	0.358
	Efficiency of gross photosynthesis	12	2.0–4.0	22	0.706
McIntire 1964 (15 August data)	Community respiration	6	1.6–3.2	33	0.782
	Gross production	6	2.9–4.1	14	0.455
	Biomass	6	98.0–148.0	17	21.4
Takahashi et al. 1975	Phytoplankton standing crop (day 15)	4	457–2290	76	827
	Thalassiosira (% of total phytoplankton)	4	0.18–0.63	46	0.19
	Photosynthetic productivity	4	45–146	45	47

In fact, replicability refers to nothing more than the degree of similarity that exists or can be obtained among experimental units. It is thus a superfluous term: the advantages of homogeneity of experimental units are well understood. It is also a misleading term in that it suggests the idea that if the degree of similarity is great enough, true replication of treatments is unnecessary to the conduct of a rigorous experiment; that will never be the case in ecological work.

Cooke (1977:64), in a review of the use of laboratory aquatic microcosms in ecological studies, provides an example of the misplaced concern that the notion generates:

> *The extent of replicability with regard to basic characteristics such as population density and rates of succession has not been adequately established in many studies. Some divergence, even in carefully cross-seeded replicate systems, has been noted, and the variation to be tolerated in experimental ecosystems remains to be established. A larger variance than customarily found in experimental work may have to be accepted, since small differences at the outset of the experiment may be magnified as succession proceeds . . . . Further work with regard to replicability is needed.*

Clearly what is needed is not "further work with regard to replicability" but rather replication of treatments!

In summarizing evidence that replicability is achievable, Cooke (1977:64, 86) states:

> *There is good evidence to show that replicate microcosms do not differ significantly with respect to levels of community metabolism (Abbott 1966) . . . . The replicability of the six streams* [experimental ones at Oregon State University], *at least with respect to rates of community metabolism has been demonstrated (McIntire et al. 1964) . . . .*

What is the meaning of these conclusions? Both of the studies cited by Cook, as well as that of Beyers (1963), found that replicate microcosms varied in all properties investigated (Table 5), with standard deviations ranging between 7 and 170% of the means. Abbott's (1966) failure to detect significance is irrelevant, since it is largely a matter of sample size (see earlier discussion of Example 8). He referred (p. 267) to coefficients of variation in the range of 13–15% as indicating "reasonable reproduceability." He makes no direct comment on whether replication of treatments is made unnecessary by such values, but in his later experimental work (Abbott 1967) he did not replicate his treatments. McIntire et al. (1964) likewise made no mention of the need for replication and failed to replicate treatments in a later experiment (McIntire 1968).

A larger example of how the notion of replicability can misdirect research efforts is provided by the Controlled Ecosystem Pollution Experiment (CEPEX) program. This was an expensive, "cooperative, multi-disciplinary research program designed to test effects of chronic exposure to low levels of pollutants on pelagic marine organisms" using large columns of ocean water enclosed in situ in polyethylene bags, some with a capacity of 1700 m^3 (Menzel and Case 1977). Early results of the program are reported in Takahashi et al. (1975), Grice et al. (1977) and in 17 papers in the January 1977 issue (**27**[1]) of the *Bulletin of Marine Science*. These reports consistently use the term "replication" to mean "similarity among experimental units treated alike." Though one of the reported experiments used two control bags ("Copper I" experiment), in all

other instances treatments were unreplicated. Nowhere in any of these papers is there any evidence of recognition that the rather "soft" biological results of the CEPEX experiments would have been much more conclusive if treatments had been replicated. Twofold replication would have been sufficient *if* the CEPEX systems were as similar as the investigators implied they were.

In their introductory paper, Menzel and Case (1977: 2) state that "it is necessary . . . to have replication of controls and experimentally manipulated enclosures." This sounds fine but they apparently mean only that the various enclosures must be initially similar, not that *treatments* must be replicated. Later Menzel (1977: 142) states:

> *A second consideration is not how closely enclosures duplicate the outside environment but whether they duplicate each other if treated identically. In the case of CEPEX, replication experiments have been conducted by Takahashi et al. (1975) which demonstrated reasonable similarities in four containers over 30 days. This study described a sequence of events of sufficient similarity in unpolluted identically treated containers to allow us to expect that when pollutants were added a realistic assessment could be made of their effect on the enclosed populations."*

To assess these "reasonable similarities" objectively, I calculated measures of variability for three variables from the graphs of Takahashi et al. (1975). The results are given in Table 5. Again, there is nothing in them to suggest that true replication can be dispensed with. To be sure, "the four containers behaved biologically in a very similar manner" (Takahashi et al. 1975), as similar experimental units almost always do to some extent. But such general similarities notwithstanding, variances are high; we must presume that the effects of manipulated variables in the early CEPEX experiments have been assessed rather imprecisely.

The notion of replicability often includes the idea that if two identically treated microcosms are initially similar they will remain so. A CEPEX report gives us a clear statement of this "principle":

> *It has been demonstrated that there was good initial species and numerical similarity among the CEEs* [Controlled Experimental Ecosystems]. *It is evident, therefore, that subsequent variations in population levels or species composition cannot be attributed to differences in the captured water columns" (Gibson and Grice 1977:90).*

This idea is counter to logic. And the experience of every ecosystem experimentalist who has bothered to use replication probably is like that of Whittaker (1961: 162), who found that

> *Experiments with indoor aquaria were affected by the phenomenon of aquarium individuality . . . the magnitude of contrasts between aquaria which supposedly represented the same conditions much exceeded expectation . . . . Differences in aquaria which were already significant in the earliest phase of an experiment were usually increased, rather than evened out, by their further development.*

Unlike a large number of their nonreplicating colleagues who work in the intertidal zone, the CEPEX investigators for the most part refrained from the application of inferential statistics. They did not, as Green (1979:71) would put it, "attempt to cover up . . . by executing statistical dances of amazing complexity around their untestable results." In the 19 CEPEX reports considered here, only one occurrence of pseudoreplication was found (Thomas and Seibert 1977).

More recently, the notion of replicability is discussed by many contributors to the symposium volume *Microcosms in Ecological Research* (Giesy 1980). Here again one finds much undisciplined terminology, much hand-wringing over coefficients of variation and similarity of experimental units and much neglect of the need for replication of treatments. This produces statements such as ". . . replication [of microcosms] may not be achievable, even under careful laboratory conditions" (Harte et al. 1980:106), and "The replicability of two microcosms that are subsets of the same naturally occurring environment is variable and it is difficult to stipulate the degree of conformity required to deem two microcosms subsets of the same ecosystem" (Giesy 1980:xlv). The implied problems are imaginary. Many of the experiments reported in this symposium volume did not employ replicated treatments and, in at least three instances (Maki 1980, Manuel and Minshall 1980, Rodgers et al. 1980), pseudoreplication was committed. At the other end of the spectrum, there are also reported in this volume numerous well-designed experiments that used replicated treatments. Yet not even one of their authors saw fit to make any clear, general statement about the necessity of treatment replication in microcosm research; perhaps to these latter authors it was too obvious.

The conclusion is that replicability is a red herring, a false issue. The question to be asked is not: "Are experimental units sufficiently similar for one to be used per treatment?" Rather it is: "Given the observed or expected variability among experimental units, how many should be assigned to each treatment?"

Optimal impact study design.—The principles of sampling as they apply to ecological field studies are perhaps nowhere more clearly discussed, or in a more lively way, than in a recent book by Green (1979). The book contains a pleasantly large ratio of common sense to equations, yet without sacrificing specificity.

On one topic I must take issue with it, however. Green suggests (pp. 29–30, 68–71) that it is valid to use inferential statistics to test for environmental impacts of an externally imposed factor even in situations

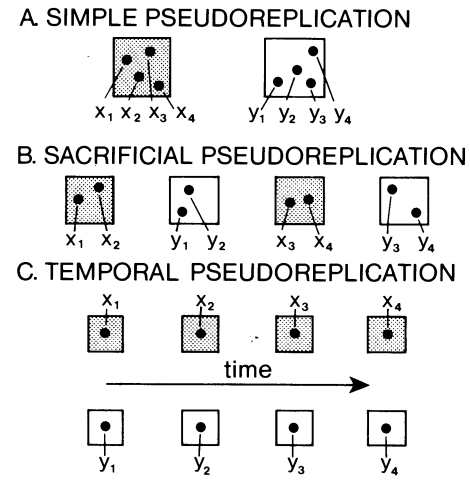

FIG. 5. Schematic representation of the three most common types of pseudoreplication. Shaded and unshaded boxes represent experimental units receiving different treatments. Each dot represents a sample or measurement. Pseudoreplication is a consequence, in each example, of statistically testing for a treatment effect by means of procedures (e.g., *t* test, *U* test) which assume, implicitly, that the four data for each treatment have come from four independent experimental units (=treatment replicates).

where only a single control area and single impact area are available.

One example Green uses is that of wastes being discharged into a river. If it is possible to take replicate samples both upstream (control area) and downstream from the discharge point and to do this both before and after the discharging of wastes begins, Green suggests carrying out what he terms an "optimal impact study." Once the data are gathered, he recommends that some procedure such as analysis of variance be applied and that "the evidence for impact effects is a significant areas-by-times interaction" (p. 70). I would argue that this is improper, and that the best one can do in such a situation is to develop graphs and tables that clearly show both the approximate mean values *and* the variability of the data on which they are based.

Though the statistical procedure (ANOVA) recommended by Green is more sophisticated than the *t* tests, *U* tests, and χ^2 tests used in most of the earlier studies cited for pseudoreplication (Table 3), pseudoreplication is no less the result. The ANOVA can only demonstrate significant differences between locations, not significant effects of the discharge. Since the treatments cannot be interspersed or assigned randomly to experimental plots (the several sampling sites, both upstream and downstream), the experiment is not controlled except in a subjective and approximate way.

More specifically, the "areas-by-times interaction" can be interpreted as an impact effect *only* if we assume that the differences between upstream and downstream locations will remain constant over time if no wastes are discharged or if they are without effect. This is unreasonable. The magnitude of the true differences ($\Delta\mu$) between two "similar" segments of a river, or two "similar" ponds, or two "similar" field plots changes constantly over time.

If ANOVA were appropriate, we would have to make arbitrary decisions about how to measure difference. For example, upstream mayfly density is X_u and downstream mayfly density is X_d. Should our null hypothesis by that (X_u/X_d) will not change with time, or should it be that ($X_u - X_d$) will not change? (Eberhardt [1976: 33] suggests the former.) Or is some other measure of difference more appropriate? Different procedures probably would be appropriate for different kinds of variables.

Eberhardt (1976, 1978) addresses this same problem of how to assess impact when there is a single site exposed. His conclusions are similar to those of Green (1979), in that he acknowledges the before–after, upstream–downstream sampling study to be the best available option. However, Eberhardt offers many caveats, clearly states the statistical difficulty, and invents the properly pejorative terms "pseudoexperiment" and "pseudodesign" for the procedure. In his own words:

> *What cannot presently be done is to insure that classical inferential methods can actually be applied to pre- and post-operational data on one impacted site* [1976:321] . . . . *The whole formal technology of experimental design is not properly admissible* [1978: 210]. . . . [Such work] *is really more in the area of sample survey design than a part of the design of experiments* [1976:32] . . . . *We have as yet not progressed very far in trying to carry the pseudodesign idea to an operationally effective stage. I am not even sure that goal is either feasible or desirable* [1976: 35].

As examples of first-rate "optimal impact studies" may be cited the Hubbard Brook deforestation experiment (e.g., Likens et al. 1970, 1977) and the Canadian whole-lake fertilization experiments (e.g., Schindler et al. 1971, Schindler 1974). Replicate experimental units were not used in these investigations, yet the effects of the experimental variables were convincingly demonstrated. Inferential statistics were not used (with minor exceptions). They were not applicable, and they would not have made the results any clearer or the conclusions any firmer. All experimenters who do not or cannot employ true replication would do well to emulate the straightforwardness of these two outstanding research groups.

Temporal pseudoreplication

This differs from simple pseudoreplication only in that the multiple samples from each experimental unit are not taken simultaneously but rather sequentially over each of several dates (Fig. 5C). Dates are then taken to represent replicated treatments and signifi-

cance tests are applied. Because successive samples from a single unit are so obviously going to be correlated with each other, the potential for spurious treatment effects is very high with such designs.

It should be remarked that repeated sampling of experimental units and the use of such data in statistical analyses can be quite proper in some circumstances. It is only the treating of successive dates as if they were independent replicates of a treatment that is invalid.

Examples of temporal pseudoreplication may be found in Cowell (1965), Clarke and Grant (1968), Thomas and Seibert (1977), Abramsky et al. (1979), McCauley and Briand (1979), and Hixon (1980).

Sacrificial pseudoreplication

This results when an experimental design involves true replication of treatments but where the data for replicates are pooled prior to statistical analysis (see next section) or where the two or more samples or measurements taken from each experimental unit are treated as independent replicates (Fig. 5B). Information on the variance among treatment replicates exists in the original data, but is confounded with the variance among samples (within replicates) or else is effectively thrown away when the samples from the two or more replicates are pooled (hence "sacrificial").

Surprisingly this convoluted approach is only slightly less common than simple pseudoreplication. Recent examples are found in Hurlbert et al. (1971), Cameron (1977), Grant et al. (1977), Virnstein (1977), Bakelaar and Odum (1978), and Bell (1980). It may be significant that all these studies involved only twofold replication of treatments; if they had restricted themselves to valid statistical procedures, they would have found fewer or no significant differences.

In some of these studies (e.g., Grant et al. 1977, Virnstein 1977), the samples from the two replicates were not pooled automatically. Rather, a significance test (e.g., t test) first was applied to test whether two replicates of a treatment were significantly different. They usually were not significantly different, and pooling was carried out. But "in the few cases where replicates were quite different, each replicate was treated separately" (Virnstein 1977).

Though, as I have indicated, the pooling of samples from separate experimental units was not justified in any circumstance, the above testing procedure is inappropriate in its own right. Certainly in any field situation, we *know* that two replicate plots or ponds in the same treatment are not identical. It may be of interest to us to know roughly *how* different they are, but a significance test of the difference is irrelevant.

Chi-Square and pseudoreplication

Chi-square is one of the most misapplied of all statistical procedures. In the manipulative ecological field experiments I reviewed it was not used frequently except in small-mammal studies. In such studies, animals are commonly caught one at a time in small traps and each capture can be regarded as an independent observation. Thus chi-square seems appropriate for testing hypotheses concerning sex ratios, distribution among microhabitats, etc. However, when it is used specifically to assess treatment effects in manipulative experiments, it seems invariably to be misapplied.

When treatments are unreplicated and chi-square is used to compare the sex ratios of one experimental and one control plot (e.g., Dobson 1979, Gaines et al. 1979) one is again only testing for a location difference, not for a treatment effect. And, as usual, if one fails to realize that, one is pseudoreplicating. This would be "simple pseudoreplication."

When two replicate plots have been available per treatment (Cameron 1977, Grant et al. 1977, Hansen and Batzli 1979), the capture data for the two replicates are invariably combined and chi-square applied to the totals. This represents "sacrificial pseudoreplication."

Then what is the correct approach? A hypothetical example (Table 6) has been contrived to demonstrate that, contrary to established tradition, chi-square is inappropriate and that the methods called for are the same ones (t test, U test, or ANOVA) that are used to analyze for treatment effects on variables such as body mass, vegetation biomass, etc.

The procedures followed in Table 6 are those used by Grant et al. (1977) and others. This example shows how they lead to a conclusion that fox predation does affect sex ratio when in fact the putative significance of the effect is attributable to a single sex ratio (B_2) being out of line with the others. Any time that happens one should suspect that something is wrong.

Pooling is wrong on four related counts. First, the 35 mice caught in A_1 can be regarded as 35 independent observations and so can the 16 mice in A_2. Thus a chi-square test to compare the sex ratios of these two plots is valid (though irrelevant). However, when the data for these two plots are pooled the resultant 51 observations are *not* independent; they represent two sets of interdependent or correlated observations. The pooled data set thus violates the fundamental assumption underlying the chi-square test.

Second, pooling treatment replicates throws out the information on the variability among replicate plots. Without such information there is no proper way to assess the significance of the difference between treatments.

Third, if one carries out a test on the pooled data, one is implicitly redefining the experimental units to be the individual mice and not the field plots. That is not allowable. Other more standard sorts of pooling (e.g., Winer 1971:378–384, Sokal and Rohlf 1981:285) usually do not imply any redefinition of the nature of the experimental unit. When they do, they should be regarded with suspicion, as redefinition of the experimental unit alters the specific hypothesis being tested.

Fourth, pooling weights the replicate plots differ-

TABLE 6. A hypothetical example of sacrificial pseudoreplication resulting from misuse of chi-square.
Question: Does fox predation affect the sex ratio of *Microtus* populations?
Experimental design: Establish four 1-ha experimental plots in a large field where foxes hunt; put fox-proof fences around two plots selected at random (A_1, A_2), keep the other two plots as controls (B_1, B_2); 1 mo later sample *Microtus* population in each plot.

	Results of sampling				
	Plot	% males	No. males	No. females	Statistical analysis
Foxes	A_1	63	22	13	Test for homogeneity with χ^2 Result: $\chi^2 = .019$, $P > .50$ So: pool the data (see below)
	A_2	56	9	7	
No foxes	B_1	60	15	10	Test for homogeneity with χ^2 Result: $\chi^2 = 2.06$, $P > .15$ So: pool the data (see below)
	B_2	43	97	130	
	Pooled data				
Foxes	$A_1 + A_2$	61	31	20	Test for homogeneity with χ^2 Result: $\chi^2 = 3.91$, $P < .05$ *Conclusion:* foxes affect sex ratio
No foxes	$B_1 + B_2$	44	112	140	

entially. The example (Table 6) is contrived to convince you on intuitive grounds that such weighting is improper; it produces a nonsense conclusion. (Note that we have said nothing about whether the number of *Microtus* captured per plot represents the total number present, is proportional to the total number present, or is proportional to a possibly variable capture effort; the matter is not relevant here.) The mean sex ratios (% males) for the two treatments should be 59.5 and 51.5% (unweighted), not 61 and 44% (weighted). Because we caught more *Microtus* in plot B_2 it is reasonable to assume that we have a more precise estimate of the true sex ratio *in that plot.* But there is no basis for the assumption, implicit in the pooling procedure, that the "true" B_2 sex ratio is a better estimate of the "true" sex ratio for the treatment than is the B_1 ratio.

Let us say that instead of studying sex ratio, we measured the body mass of every one of the 143 (=22 + 9 + 15 + 97) males caught and that the data listed under "% males" in Table 6 now represent mean masses (e.g., in grams). The effect of fox predation could be properly assessed by applying a conventional analysis of variance to the original data. That approach entails calculating treatment means as the *unweighted* averages of plot means, even though sample size varies from plot to plot. Differential weighting would be unwarranted for the body mass data, and it is equally unwarranted for the sex ratio data.

I believe the only appropriate test for the example in Table 6 would be either a *t* test or a *U* test. With twofold replication, these do not have much power, but neither will they mislead.

The commonness of this type of chi-square misuse probably is traceable to the kinds of examples found in statistics texts, which too often are only from genetics, or from mensurative rather than manipulative experiments, or from manipulative experiments (e.g., medical ones) in which individual organisms *are* the experimental units and not simply components of them, as in the mammal field studies cited. It does seem incongruous that chi-square *can* be used to test for a sex ratio difference between two populations (mensurative experiment), but *cannot* be used to test for such a difference between these two populations and two other populations subjected to a different treatment (manipulative experiment). Yet it seems to be a fact. I know of no statistics textbook that provides clear and reliable guidance on this matter.

Implicit pseudoreplication

In the examples discussed so far, pseudoreplication is a consequence of the faulty but explicit use of significance tests to test for treatment effects. However, in some manipulative studies involving unreplicated but subsampled treatments (e.g., Menge 1972, Lubchenco 1980), the authors present standard errors or 95% confidence intervals along with their means and discuss the putative effects of the imposed variable, but they do not apply any direct tests of significance. In such cases, the appropriateness of the label "pseudoreplication" depends on how aware the authors seem to be of the limitations of their experimental design and data. If they seem to regard their paired and nonoverlapping 95% confidence intervals as equivalent to significance tests, and if they offer no specific disclaimer acknowledging that their data are, in fact, inadequate for assessing treatment effects, then their procedures seem reasonably labelled "implicit pseudoreplication."

The presentation of information on variability within experimental units sometimes may be of interest even if treatments are not replicated. I believe, however, that the least misleading way to present this might be in the form of standard deviations rather than standard errors or 95% confidence intervals. This will help emphasize what the authors should acknowledge explicitly: that the variability within experimental units is useless for assessing possible treatment effects. Sample sizes can be indicated independently; that will allow

rough determination of standard errors for those wishing to know them.

Original sin at Rothamstead

It may be of comfort to know that pseudoreplication is not the invention of modern ecologists but in fact was first committed by Fisher himself. We thus have a theological "out": the father of modern experimental design committed original sin, so what can be expected of mere mortals like ourselves?

The story is well told by his daughter (Box 1978: 110–112) and Cochran (1980). The slip came in a factorial experiment involving 12 potato varieties, 3 types of potassium fertilization, and 2 levels (0, +) of farmyard manure (Fisher and Mackenzie 1923). In the layout, "the total area was divided into two equal parts, one of which was used for the farmyard manure series, and the other for the series without farmyard manure," and the other factors were distributed in plots and subplots over both halves of the area. This layout clearly does not permit a valid test for a manure effect, but Fisher nevertheless used analysis of variance to test for one (and found none). He soon recognized his error, prodded perhaps by comments sent him by Gossett (J. F. Box, *personal communication*). In 1925 in the first edition of *Statistical Methods for Research Workers* he presented, as an example, an analysis of variance for the data from the manured half of the study area *only,* remaining silent about the other half of the area and his original analysis (Fisher 1958:236–241). Since the experiment had been designed by other persons and without Fisher's collaboration, this incident might be considered only an "original misdemeanor"—*if* Fisher had instructed us with an open confession, biting the bullet as well as the apple.

For Statisticians

Where did you fail us? We took your courses; we read your books. Here are some suggestions.

1) Include in your statistics books concise, nonmathematical expositions of the basic principles of experimental design. Steel and Torrie (1980) do an excellent job of this, but most other texts do not even try. Do not presume that more than a minority of your students who go on to become experimental scientists will take a formal course in experimental design.

2) In your statistics books, when using examples, give more details on the physical layout and conduct of the experiments from which the data sets are obtained. Discuss alternative layouts and their validity or lack thereof. Cite and discuss actual examples of the more common sorts of design errors, such as pseudoreplication.

3) Emphasize that although most statistical methods can be applied to either experimental or observational data, their proper use in the former case requires that several conditions be met concerning the physical conduct of the experiment.

4) Be more hard-nosed and suspicious when you are being consulted by experimenters. Do not let them sweet-talk you into condoning a statistical analysis where accuracy would be better served by not applying inferential statistics at all. Some statisticians may be too willing, for example, to accept as substitutes for proper design the self-interested claims of experimenters about the homogeneity of their experimental material or the "certain" absence of nondemonic intrusion.

5) When you do assist with analysis of data from experiments, encourage the experimenter to include in his report explicit description of the physical layout of the experiment. When the design contains weaknesses, encourage the experimenter to discuss these in his report.

For Editors

Poorly designed or incorrectly analyzed experimental work literally is flooding the ecological literature. In my survey, I found that 48% of recent, statistically analyzed, ecological field experiments have involved pseudoreplication. My students, Innis's (1979) students, and Underwood (1981) confirm the magnitude of the statistical malpractice problem. How can the flood be stemmed?

Many remedies might be proposed. Better training in statistics and experimental design for all ecologists is the most obvious one. But how can this be accomplished effectively and quickly? Rather easily. Though the typical manuscript is reviewed and critiqued by its authors, some of their colleagues, a few anonymous reviewers, and an editor, only the editor determines whether it will be published or not. If editors collectively were to become only *slightly* more knowledgeable in statistics, and if they, as a matter of routine, were to scrutinize manuscripts for a certain few common errors, a major improvement in the ecological literature could be effected in 1 or 2 yr. When the coin of the realm is the published paper, nothing educates so well as an editorial rejection or request for major revision. A barrage of clearly explained rejection notices would educate more ecologists more rapidly than any general attempt to upgrade statistics books and statistics courses, matters which are, in any case, beyond our control.

Statistical sophistication, or lack of it, is not the main problem. At least in field ecology, the designs of most experiments are simple and when errors are made they are of a gross sort. There will be instances where a valid but perhaps complicated experimental design is employed or where error intrudes only in some difficult-to-discern misstep in statistical analysis. Such errors can be hard to catch, even for professional statisticians. *Their* elimination can be brought about only gradually, as investigators and editors both advance in understanding of statistics.

For the larger class of errors, including pseudoreplication in its various forms, detection requires only modest familiarity with the elementary principles of statistics and experimental design. Lack of this on the part of ecologists and their editors is the major proximate cause of our present plight. But perhaps it is not so much that the familiarity is lacking, as that the principles are so easily lost sight of, in most books and courses, among the multitudinous mathematical aspects of statistical analysis.

Some specific actions that editors might take to combat pseudoreplication and related errors are as follows:

1) Insist that the physical layout of an experiment either be presented in a diagram or be described in sufficient detail that the reader can sketch a diagram for himself. Especially in many marine experiments, this information on the physical layout is either not provided or is given only vaguely. In such cases, the validity of the experimental design cannot be assessed.

2) Determine from the above whether the design involves true replication and interspersion of treatments.

3) Determine from the description of procedures the manner in which treatments were assigned to experimental units. If this was accomplished by means other than randomization (simple or restricted), then examine the experimenter's justification for not employing randomization. Pass judgment, and this will have to be subjective, as to the likelihood that his procedure for assigning treatments to experimental units may have introduced bias or generated spurious treatment effects. As long as the procedure produced good interspersion of treatments, the lack of true randomization may not be a deficiency. On the other hand, if randomization procedures were used but produced a high degree of segregation of treatments, the consequent potential for error should be explicitly acknowledged by the authors.

4) Insist that the statistical analysis applied be specified in detail. Sometimes this can be done by referring to specific pages in a statistics book. More often additional information must be supplied.

5) Disallow the use of inferential statistics where they are being misapplied. Where they are marginally allowable, insist on disclaimers and explicit mention of the weaknesses of the experimental design. Disallow "implicit" pseudoreplication which, as it often appears in the guise of very "convincing" graphs, is especially misleading.

6) Be liberal in accepting good papers that refrain from using inferential statistics when these cannot validly be applied. Many papers, both descriptive and experimental, fall in this category. Because an obsessive preoccupation with quantification sometimes coincides, in a reviewer or editor, with a blindness to pseudoreplication, it is often easier to get a paper published if one uses erroneous statistical analysis than if one uses no statistical analysis at all.

Conclusion

During a discussion at a meeting of the Royal Statistical Society in 1934, a Mr. Page suggested that "we had now moved a long way from the position of a certain distinguished Professor of Agriculture who said, 'Damn the duplicate plot; give me one plot and I know where I am'" (Wishart 1934:56). Doubtless that was and is true for many areas of agricultural science. Ecologists, however, have marched to a different drummer. A large percentage of modern experimental field ecologists would seem quite willing to clap this "distinguished professor" on the back, slide him his ale, and toast his health. To demonstrate their modernity, perhaps they would add: "As long as the bloody thing's big enough to subsample, we'll give Mr. Fisher his error term!"

Pseudoreplication is probably the single most common fault in the design and analysis of ecological field experiments. It is at least equally common in many other areas of research. It is hoped that this review will contribute to a reduction in its frequency. Such reduction should be a manageable, short-term task.

Acknowledgments

This paper is based on a talk given at the Florida State University Wakulla Springs symposium in March 1981. The manuscript has been improved substantially by the suggestions of C. Chang, B. D. Collier, C. F. Cooper, P. G. Fairweather, D. A. Farris, W. J. Platt, A. J. Underwood, D. Wise, P. H. Zedler, and two anonymous reviewers. Any errors that remain are their responsibility and theirs alone. J. F. Box kindly provided information on W. S. Gossett's correspondence with her father, R. A. Fisher.

I dedicate this paper to Lincoln P. Brower, who introduced me to experimental ecology.

Literature Cited

Abbott, W. 1966. Microcosm studies on estuarine waters. I. The replicability of microcosms. Journal of the Water Polution Control Federation **38**:258–270.

———. 1967. Microcosm studies on estuarine waters. II. The effects of single doses of nitrate and phosphate. Journal of the Water Polution Control Federation **39**:113–122.

Abramsky, Z., M. I. Dyer, and P. D. Harrison. 1979. Competition among small mammals in experimentally perturbed areas of the shortgrass prarie. Ecology **60**:530–536.

Anscombe, F. J. 1948. The validity of comparative experiments. Journal of the Royal Statistical Society (London) **A 111**:181–211.

Bakelaar, R. G., and E. P. Odum. 1978. Community and population level responses to fertilization in an old-field ecosystem. Ecology **59**:660–665.

Barbacki, S., and R. A. Fisher. 1936. A test of the supposed precision of systematic arrangements. Annals of Eugenics **7**:183–193.

Barrett, G. W. 1968. The effects of an acute insecticide stress on a semi-enclosed grassland ecosystem. Ecology **49**:1019–1035.

Bell, S. S. 1980. Meiofauna-macrofauna interactions in a high salt marsh habitat. Ecological Monographs **50**:487–505.

Bell, S. S., and B. C. Coull. 1978. Field evidence that shrimp predation regulates meiofauna. Oecologia **35**:245–248.

Beyers, R. J. 1963. The metabolism of twelve aquatic lab-

oratory microecosystems. Ecological Monographs **33**:281–306.

Boaden, P. J. S. 1962. Colonization of graded sand by an interstitial fauna. Cahiers de Biologie Marine **3**:245–248.

Box, J. F. 1978. R. A. Fisher: the life of a scientist. Wiley, New York, New York, USA.

Brown, J. H., J. J. Graver, and D. W. Davidson. 1975. A preliminary study of seed predation in desert and montane habitats. Ecology **56**:987–992.

Buzas, M. A. 1978. Foraminifera as prey for benthic deposit feeders: results of predator exclusion experiments. Journal of Marine Research **36**:617–625.

Cameron, G. N. 1977. Experimental species removal: demographic responses by *Sigmodon hispidus* and *Reithrodontomys fulvescens*. Journal of Mammalogy **58**:488–506.

Chew, R. M. 1978. The impact of small mammals on ecosystem structure and function. Pages 167–180 *in* D. P. Snyder, editor. Populations of small mammals under natural conditions. Pymatuning Symposia in Ecology Number 5, Pymatuning Laboratory of Ecology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

Clarke, R. D., and P. R. Grant. 1968. An experimental study of the role of spiders as predators in a forest litter community. Part I. Ecology **49**:1152–1154.

Clark, R. W. 1976. The life of Bertrand Russell. Knopf, New York, New York, USA.

Cochran, W. G. 1963. Sampling techniques. Third edition. Wiley, New York, New York, USA.

———. 1980. Fisher and the analysis of variance. Pages 17–34 *in* E. Fienberg and D. V. Hinckley, editors. R. A. Fischer: an appreciation (Lecture Notes in Statistics, Volume 1). Springer-Verlag, New York, New York, USA.

Cochran, W. G., and G. M. Cox. 1957. Experimental designs. Second edition. Wiley, New York, New York, USA.

Connell, J. H. 1974. Field experiments in marine ecology. Pages 21–54 *in* R. Mariscal, editor. Experimental marine biology. Academic Press, New York, New York, USA.

Cooke, G. D. 1977. Experimental aquatic laboratory ecosystems and communities. Pages 59–103 *in* J. Cairns, editor. Aquatic microbial communities. Garland, New York, New York, USA.

Cope, O. B., J. P. McCraren, and L. Eller. 1969. Effects of dichlobenil on two fishpond environments. Weed Science **17**:158–165.

Cowell, B. C. 1965. The effects of sodium arsenite and Silvex on the plankton populations in farm ponds. Transactions of the American Fisheries Society **94**:371.

Cox, D. R. 1958. Planning of experiments. Wiley, New York, New York, USA.

Cox, P. A. 1981. Vertebrate pollination and the maintenance of unisexuality in *Freycinetia*. Dissertation. Harvard University, Cambridge, Massachusetts, USA.

———. 1982. Vertebrate pollination and the maintenance of dioecism in *Freycinetia*. American Naturalist **120**:65–80.

Crowner, A. W., and G. W. Barrett. 1979. Effects of fire on the small mammal component of an experimental grassland community. Journal of Mammalogy **60**:803–813.

Davies, A. G. 1978. Pollution studies with marine plankton. Part II. Heavy metals. Advances in Marine Biology **15**:381–508.

Dobson, F. S. 1979. An experimental study of dispersal in the California ground squirrel. Ecology **60**:1103–1109.

Durbin, A. G., S. W. Nixon, and C. A. Oviatt. 1979. Effects of the spawning migration of the alewife, *Alosa pseudoharengus*, on freshwater ecosystems. Ecology **60**:8–17.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

———. 1978. Appraising variability in population studies. Journal of Wildlife Management **42**:207–238.

Fisher, R. A. 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture (London) **33**:503–513.

———. 1939. "Student." Annals of Eugenics **9**:1–9.

———. 1958. Statistical methods for research workers. 13th edition. Oliver and Boyd, London, England.

———. 1971. The design of experiments. Ninth edition. Hafner, New York, New York, USA.

Fisher, R. A., and W. A. Mackenzie. 1923. Studies in crop variation. II. The manurial response of different potato varieties. Journal of Agricultural Science **13**:311–320.

Fisher, R. A., and J. Wishart. 1930. The arrangement of field experiments and the statistical reduction of the results. Imperial Bureau of Soil Science (London), Technical Communication Number **10**:1–23.

Fry, D. L., and J. A. Osborne. 1980. Zooplankton abundance and diversity in central Florida grass carp ponds. Hydrobiologia **68**:145–155.

Gaines, M. S., A. M. Vivas, and C. L. Baker. 1979. An experimental analysis of dispersal in fluctuating vole populations: demographic parameters. Ecology **60**:814–828.

Gibson, V. R., and G. D. Grice. 1977. Response of macrozooplankton populations to copper: controlled ecosystem pollution experiment. Bulletin of Marine Science **27**:85–91.

Giesy, J. P., Jr., editor. 1980. Microcosms in ecological research. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Giguere, L. 1979. An experimental test of Dodson's hypothesis that *Ambystoma* (a salamander) and *Chaoborus* (a phantom midge) have complementary feeding niches. Canadian Journal of Zoology **57**:1091–1097.

Gilderhus, P. A. 1966. Some effects of sublethal concentrations of sodium arsenite on bluegills and the aquatic environment. Transactions of the American Fisheries Society **95**:289–296.

Gossett, W. S. 1923. On testing varieties of cereals. Biometrika **15**:271–293.

———. 1936. Cooperation in large scale experiments (a discussion). Journal of the Royal Statistical Society, Supplement **3**:115–136.

———. ("Student"). 1937. Comparison between balanced and random arrangements of field plots. Biometrika **29**: 363–379.

Grant, W. E., N. R. French, and D. M. Swift. 1977. Response of a small mammal community to water and nitrogen treatments in a shortgrass prarie ecosystem. Journal of Mammalogy **58**:637–652.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York, New York, USA.

Grice, G. D., M. R. Reeve, P. Koeller, and D. W. Menzel. 1977. The use of large volume, transparent enclosed seasurface water columns in the study of stress on plankton ecosystems. Helgoländer Wissenschaftliche Meeresuntersuchungen **30**:118–133.

Hall, R. J., G. E. Likens, S. B. Fiance, and G. R. Hendrey. 1980. Experimental acidification of a stream in the Hubbard Brook Experimental Forest, New Hampshire. Ecology **61**:976–989.

Hansen, L. P., and G. O. Batzli. 1979. Influence of supplemental food on local populations of *Peromyscus leucopus*. Journal of Mammalogy **60**:335–342.

Harger, J. R. E. 1971. Variation and relative "niche" size in the sea mussel *Mytilus edulis* in association with *Mytilus californianus*. Veliger **14**:275–281.

Harte, J., D. Levy, J. Rees, and E. Saegebarth. 1980. Making microcosms an effective assessment tool. Pages 105–137 *in*

J. P. Giesy, Jr., editor. Microcosms in ecological research. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Haven, S. B. 1973. Competition for food between the intertidal gastropods *Acmaea scabra* and *Acmaea digitalis*. Ecology **54**:143–151.

Hayne, D. W. 1978. Experimental designs and statistical analyses. Pages 3–10 *in* D. P. Snyder, editor. Populations of small mammals under natural conditions. Pymatuning Symposia in Ecology Number 5. Pymatuning Laboratory of Ecology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

Hayward, G. F., and J. Phillipson. 1979. Community structure and functional role of small mammals in ecosystems. Pages 135–211 *in* D. M. Stoddart, editor. Ecology of small mammals. Chapman and Hall, London, England.

Hixon, M. A. 1980. Competitive interactions between California reef fishes of the genus *Embiotoca*. Ecology **61**:918–931.

Holbrook, S. J. 1979. Habitat utilization, competitive interactions, and coexistence of three species of cricetine rodents in east-central Arizona. Ecology **60**:758–759.

Holland, A. F., N. K. Mountfort, M. H. Hiegel, K. R. Kaumeyer, and J. A. Mihursky. 1980. The influence of predation on infaunal abundance in upper Chesapeake Bay. Marine Biology **57**:221–235.

Hurlbert, S. H. 1975. Secondary effects of pesticides on aquatic ecosystems. Residue Reviews **58**:81–148.

Hurlbert, S. H., and C. C. Y. Chang. 1983. Ornitholimnology: effects of grazing by the Andean flamingo (*Phoenicoparrus andinus*). Proceedings of the National Academy of Sciences (USA) **80**:4766–4769.

Hurlbert, S. H., and M. S. Mulla. 1981. Impacts of mosquitofish (*Gambusia affinis*) predation on plankton communities. Hydrobiologia **83**:125–151.

Hurlbert, S. H., M. S. Mulla, J. O. Keith, W. E. Westlake, and M. E. Dusch. 1971. Biological effects and persistence of Dursban in freshwater ponds. Journal of Economic Entomology **63**:43–52.

Hurlbert, S. H., M. S. Mulla, and H. R. Willson. 1972. Effects of an organophosphorus insecticide on the phytoplankton, zooplankton, and insect populations of freshwater ponds. Ecological Monographs **42**:269–299.

Innis, G. S. 1979. Letter to the Editor. Bulletin of the Ecological Society of America **60**:142.

Jones, B. R., and J. B. Moyle. 1963. Populations of plankton animals and residual chlorinated hydrocarbons in soils of six Minnesota ponds treated for control of mosquito larvae. Transactions of the American Fisheries Society **92**:211–219.

Joule, J., and G. N. Cameron. 1975. Species removal studies. I. Dispersal strategies of sympatric *Sigmodon hispidus* and *Reithrodontomys fulvescens* populations. Journal of Mammalogy **56**:378–396.

Kempthorne, D. 1979. The design and analysis of experiments. Krieger, Huntington, New York, USA.

Kerfoot, W. C., editor. 1980. Evolution and ecology of zooplankton communities. (Special Symposium Volume 3, American Society of Limnology and Oceanography) University Press of New England, Hanover, New Hampshire, USA.

Likens, G. E., F. H. Bormann, N. M. Johnson, D. W. Fisher, and R. S. Pierce. 1970. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed ecosystem. Ecological Monographs **40**:23–47.

Likens, G. E., F. H. Bormann, R. S. Pierce, J. S. Eaton, and N. M. Johnson. 1977. Biogeochemistry of a forested ecosystem. Springer-Verlag, New York, New York, USA.

Lubchenco, J. 1980. Algal zonation in the New England rocky intertidal community: an experimental analysis. Ecology **61**:333–344.

Maki, A. W. 1980. Evaluation of toxicant effects on structure and function of model stream communities: correlation with natural stream effects. Pages 583–609 *in* J. P. Giesy, Jr., editor. Microcosms in ecological research. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Malone, C. R. 1969. Effects of Diazinon contamination on an old-field ecosystem. American Midland Naturalist **82**: 1–27.

Manuel, C. Y., and G. W. Minshall. 1980. Limitations on the use of microcosms for predicting algal response to nutrient enrichment in lotic systems. Pages 645–667 *in* J. P. Giesy, Jr., editor. Microcosms in ecological research. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Markowitz, D. V. 1980. Predator influence on shore-level size gradients in *Tegula funebralis*. Journal of Experimental Marine Biology and Ecology **45**:1–13.

Marshall, J. S., and D. L. Mellinger. 1980. Dynamics of cadmium-stressed plankton communities. Canadian Journal of Fisheries and Aquatic Sciences **37**:403–414.

McCauley, E., and F. Briand. 1979. Zooplankton grazing and phytoplankton species richness: field tests of the predation hypothesis. Limnology and Oceanography **24**:243–252.

McIntire, C. D. 1968. Structural characteristics of benthic algal communities in laboratory streams. Ecology **49**:520–537.

McIntire, C. D., R. L. Garrison, H. K. Phinney, and C. E. Warren. 1964. Primary production in laboratory streams. Limnology and Oceanography **9**:92–102.

Menge, B. A. 1972. Competition for food between intertidal starfish species and its effect on body size and feeding. Ecology **53**:635–644.

Menzel, D. W. 1977. Summary of experimental results: controlled ecosystem pollution experiment. Bulletin of Marine Science **27**:142–145.

Menzel, D. W., and J. Case. 1977. Concept and design: controlled ecosystem pollution experiment. Bulletin of Marine Science **27**:3–7.

Munger, J. C., and J. H. Brown. 1981. Competition in desert rodents: an experiment with semipermeable exclosures. Science **211**:510–512.

Neyman, J., and E. S. Pearson. 1937. Notes on some points in "Student's" paper on "Comparison between balanced and random arrangements of field plots." Biometrika **29**: 380–388.

Paine, R. T. 1974. Intertidal community structure: experimental studies on the relationship between a dominant competitor and its principal predator. Oecologia **15**:93–120.

———. 1980. Food webs, linkage, interaction strength and community infrastructure. Journal of Animal Ecology **49**: 667–685.

Pearson, E. S. 1939. William Sealey Gossett, 1876-1937. (2) "Student" as statistician. Biometrika **30**:210–250.

Peterson, C. H. 1977. Competitive organization of the soft-bottom macrobethnic communities of southern California lagoons. Marine Biology **43**:343–359.

Price, M. 1978. The role of microhabitat in structuring desert rodent communities. Ecology **59**:910–921.

Rausher, M. D., and P. Feeny. 1980. Herbivory, plant density, and plant reproductive success: the effect of *Battus philenor* on *Aristolochia reticulata*. Ecology **61**:905–917.

Reichman, O. J. 1979. Desert granivore foraging and its impact on seed densities and distributions. Ecology **60**:1085–1092.

Reise, K. 1978. Experiments on epibenthic predation in the Wadden Sea. Helgoländer Wissenschaftliche Meeresuntersuchungen **31**:55–101.

Rodgers, J. H., Jr., J. R. Clark, K. L. Dickson, and J. Cairns, Jr. 1980. Nontaxonomic analyses of structure and function of aufwuchs communities in lotic microcosms. Pages 625–644 *in* J. P. Giesy, Jr., editor. Microcosms in ecological research. Technical Information Center, United States Department of Energy, Washington, D.C., USA.

Rogers, C. S. 1979. The effect of shading on coral reef structure and function. Journal of Experimental Marine Biology and Ecology **41**:269–288.

Savage, L. J., moderator. 1962. The foundations of statistical inference: a discussion. Wiley, New York, New York, USA.

Schindler, D. W. 1974. Eutrophication and recovery in experimental lakes: implications for lake management. Science **184**:897–898.

Schindler, D. W., F. A. J. Armstrong, S. K. Holmgren, and G. J. Brunskill. 1971. Eutrophication of lake 227, Experimental Lakes Area, northwestern Ontario, by addition of phosphate and nitrate. Journal of the Fisheries Research Board of Canada **28**:1763–1782.

Sherman, K. M., and B. C. Coull. 1980. The response of meiofauna to sediment disturbance. Journal of Experimental Marine Biology and Ecology **46**:59–71.

Slocum, C. J. 1980. Differential susceptibility to grazers in two phases of an intertidal alga: advantages of heteromorphic generations. Journal of Experimental Marine Biology and Ecology **46**:99–110.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. Second edition. Freeman, San Francisco, California, USA.

Spencer, S. R., and G. W. Barrett. 1980. Meadow vole population response to vegetational changes resulting from 2, 4-D application, American Midland Naturalist **103**:32–46.

Spitz, F. 1968. Interaction entre vegetation epigée d'une luzerniere et des populations enclose ou non enclose de *Microtus arvalis* Pallas. Terre et Vie 1968:274–306.

Steel, R. G. D., and J. H. Torrie. 1980. Principles and procedures of statistics. Second edition. McGraw-Hill, New York, New York, USA.

Suttman, C. E., and G. W. Barrett. 1979. Effects of Sevin on arthropods in an agricultural and an old-field plant community. Ecology **60**:628–641.

Takahashi, M., W. H. Thomas, D. L. R. Seibert, J. Beers, P. Koeller, and T. R. Parsons. 1975. The replication of biological events in enclosed water columns. Archiv für Hydrobiologie **76**:5–23.

Thomas, W. H., and D. L. R. Seibert. 1977. Effects of copper on the dominance and diversity of algae: controlled ecosystem pollution experiment. Bulletin of Marine Science **27**:17–22.

Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and Marine Biology Annual Reviews **19**:513–605.

Vance, R. R. 1979. Effects of grazing by the sea urchin, *Centrostephanus coronatus*, on prey community composition. Ecology **60**:537–546.

Vance, R. R., and R. J. Schmitt. 1979. The effect of the predator-avoidance behavior of the sea urchin, *Centrostephanus coronatus*, on the breadth of its diet. Oecologia (Berlin) **44**:21–45.

Virnstein, R. W. 1977. The importance of predation by crabs and fishes on benthic infauna in Chesapeake Bay. Ecology **58**:1199–1217.

Vossbrinck, C. R., D. C. Coleman, and T. A. Wooley. 1979. Abiotic and biotic factors in litter decomposition in a semiarid grassland. Ecology **60**:265–271.

Waloff, N., and O. W. Richards. 1977. The effect of insect fauna on growth mortality and natality of broom, *Sarothamnus scoparius*. Journal of Applied Ecology **14**:787–798.

Warwick, R. M., J. T. Davey, J. M. Gee, and C. L. George. 1982. Faunistic control of *Enteromorpha* blooms: a field experiment. Journal of Experimental Marine Biology and Ecology **56**:23–31.

Werner, E. E., and D. J. Hall. 1976. Niche shifts in sunfishes: experimental evidence and significance. Science **191**:404–406.

Whittaker, R. H. 1961. Experiments with radiophosphorus tracer in aquarium microcosms. Ecological Monographs **31**: 157–188.

Winer, B. J. 1971. Statistical principles in experimental design. McGraw-Hill, New York, New York, USA.

Wishart, J. 1934. Statistics in agricultural research. Journal of the Royal Statistical Society **1**:26–61.

Yates, F. 1939. The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. Biometrika **30**:440–466.

———. 1960. Sampling methods for censuses and surveys. Third edition. Hafner, New York, New York, USA.

Youden, W. J. 1972. Randomization and experimentation. Technometrics **14**:13–22.

Young, D. K., M. A. Buzas, and M. W. Young. 1976. Species densities of macrobenthos associated with seagrass: a field experimental study of predation. Journal of Marine Research **34**:577–592.