

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-m² 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.

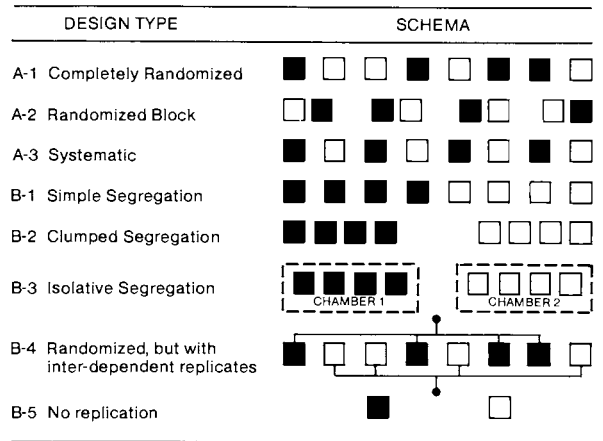


FIG. 1. Schematic representation of various acceptable modes (A) of interspersing the replicates (boxes) of two treatments (shaded, unshaded) and various ways (B) in which the principle of interspersion can be violated.

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

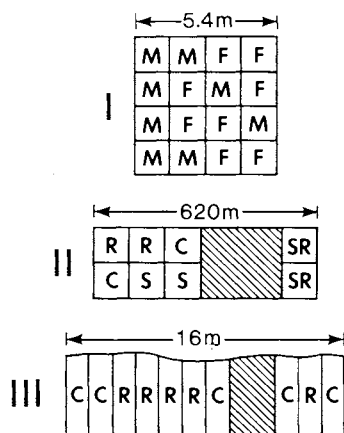


FIG. 2. Three experimental layouts exhibiting partial but inadequate interspersions of treatments. (I) test to compare predation rates on male (M) vs. female (F) floral parts placed on forest floor (Cox 1981, 1982); (II) test of effects on dispersal of removing from unfenced field plots one (S, R), both (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 interspersions 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 interspersions 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 interspersions *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-

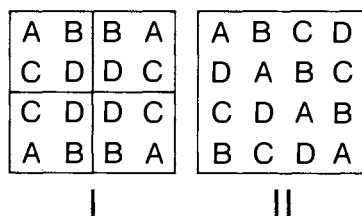


FIG. 3. Examples of segregated arrangements of four treatments, each replicated four times, that can result from use of restricted randomization procedures: (I) randomized block design, (II) Latin square design.

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 interspersions 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 interspersions 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. interspersions

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 interspersions 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; interspersions, 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,

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

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