An Ancient Black Art

Thomas J. Koehnle
Hiram College

The doctrine of pseudoreplication (DP) offers specific advice on how to ensure statistical independence and compute F-ratios properly when testing a null hypothesis. Our target article showed that this advice can lead to problems in experimental design and analysis. Though a few commenters attempt to defend DP, none offered substantive evidence that our modeling results were incorrect. In our response, we further highlight the complications surrounding definitions of experimental units. In particular, we show that the definition of independence assumed in DP is inconsistent with independence as defined in probability theory. We show that interconnectedness across levels of analysis is pervasive, and that no simple set of rules or procedures can help experimenters avoid this problem. We argue that the relevance of Hurlbert's experimental units can be tested in the context of nested analysis of variance (ANOVA) designs. Then, we show how the relevance of Hurlbert's experimental units can be tested and how the weakness of null testing and the inability of p values to predict whether a result will generalize or be replicated.

Keywords: pseudoreplication, statistical analysis, experimental design, statistical independence, Monte Carlo simulation

We will begin our reply to comments with a note of thanks to the editor, Gordon Burghardt. The doctrine of pseudoreplication (DP) is, and we suspect will remain, controversial. It is important to note the early advances in this debate, including the early pragmatic critique offered implicitly by Underwood (1997) and the more philosophical criticisms leveled by Oksanen (2001, 2004). We owe a great debt to them in helping to clear our own thinking on these issues. We take our results on the simulated pheromone experiments as dispositive: the advice offered on DP is not always and everywhere correct. The simulated experiments we offered provide simple and compelling counterexamples.

As we showed, Monte Carlo simulation can be a powerful tool for validating claims about experimental designs and statistical analyses. Though the advice offered to avoid pseudoreplication has been long accepted, there is no simple set of rules that will ensure statistical independence, provide adequate physical control, and allow identification of confounding factors. Indeed, the advice of Hurlbert (1984) seems logical at first blush, but following that advice can lead to numerous errors.

In our response, we will first consider our Monte Carlo results and the specific criticisms leveled against them. Then, we shall turn to other objections raised by Stuart Hurlbert. We consider statistical independence, physical controls, and confidence in experimental results alongside his comments on each subject. Next, we respond to the interesting historical and philosophical points raised by Coss, and show how the relevance of Hurlbert’s experimental units can be tested in the context of nested analysis of variance (ANOVA) designs. Then, we consider Wiley’s comments on sample size and Type I error rates at some length. Finally, we use a hypothetical playback experiment to reinforce the problem with the notion of experimental units, in our reply to Freeberg and Lucas.

Results of the Monte Carlo Simulations

Over and above what has already been said about the philosophical and practical limits of experimentation (Oksanen, 2001, 2004; Underwood, 1997), we turn to the particular findings from our target article. Crucially, we observed four things.

1. Hurlbert (1984) recommended averaging data from subjects in the same experimental unit. We found that averaging within experimental units decreases statistical power, while nested ANOVA and other multilevel modeling approaches can account for differences between and among experimental units. We are pleased to note that in his reply to our target article, Hurlbert announced his acceptance of these multilevel models. Presumably, averaging among subjects within a given unit is no longer an essential step in formal data analysis (also see Reply to Coss).

2. Modeling all levels of analysis increases the ability to detect confounding effects and contamination across blocks. No commenter offered substantive disagreement with this finding.

3. Prioritizing interspersion over randomization of blocks can reduce the ability of researchers to detect contamination and block effects. Rather than argue that there is some flaw in our results, in his reply Hurlbert claims that the simulated layout experiments are not a block design. In designing these simulated experiments, however, we used Hurlbert’s (1984) own terminology:

Thomas J. Koehnle, Department of Biology, Hiram College; Jeffrey C. Schank, Department of Psychology and Animal Behavior Graduate Group, University of California, Davis.

Correspondence concerning this article should be addressed to Jeffrey C. Schank, Department of Psychology, University of California, One Shields Avenue, Davis, CA 95616. E-mail: jcschank@ucdavis.edu
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 [emphasis added] Latin square) employing restricted randomization reduce the possibility of getting extremely segregated layouts, but still allow degrees of segregation unacceptable to thoughtful experimenters (see Figure 3). (Hurlbert, 1984, p. 196)

We fail to see how our simulated pheromone experiments are not block designs. Further, we reiterate another of our basic points, that no one has offered a rational set of criteria by which “thoughtful experimenters” might determine whether a given block layout is adequately interspersed. Interspersion may or may not be necessary (see Torgerson, 1985, for a discussion the pragmatics of this issue in the context of preference studies), but one cannot start with the assumption that the spatial distribution of a contaminant will not coincide with the layout of experimental blocks, particularly (as we showed) if the contaminant comes from subjects within each block.

4. Monte Carlo approaches can be used to find heuristics about when and how to pool error variance (e.g., Bancroft, 1964). (For a complete overview of Monte Carlo methods, see Manly (2007) and Morgan (2003).) Hurlbert (1984) initially opposed pooling or using a mixed pooling rule. Later, he seemed to weaken his stance on this issue somewhat, provisionally accepting a mixed rule (Hurlbert, 2004). In his reply, Hurlbert seems to again reject pooling in most cases, even though criterion $p$ value levels for pooling have long been established by Monte Carlo simulation methods. Hurlbert’s recommendation that every scientist read Kreft and de Leeuw (1999) is a good one, but as those authors show, pooling across levels is at the heart of multilevel modeling. One cannot reject pooling and accept multilevel modeling.

Reply to Hurlbert

Hurlbert begins his rebuttal with a list of errors we allegedly made in the target article. It is essential to show that the problems we identified in DP are correct and how they are related to a misunderstanding of independence in statistical inference. To be clear, we will quote Hurlbert’s list of nine “errors” and reply to each below.

[1] pseudoreplication is not a “doctrine”, nor a synonym or neologism for “experimental confounds” or “failure of statistical independence”; [2] I have never recommended that any “experimental designs . . . [be] labeled pseudoreplicated” or argued “that certain experimental designs are inherently invalid” (though they can be invalid for a particular objective); [3] I have never argued “that locations close in space and time are by their very nature statistically dependent”; [4] I have never stated or implied that “study [of] a particular lake or watershed . . . [can provide no] knowledge about lakes or watersheds in general . . .”; [5] I have never stated or implied that “temporal pseudoreplication . . . occurs [automatically, is implied] when repeated measures are taken on an experimental unit over time;” [6] in every discipline there are “satisfactory criteria for drawing boundaries around experimental units;” [7] in no type of experimentation is it appropriate to “control physical conditions . . . as much as possible;” [8] I do not “reject the importance of physical control or regulation of the environment” where regulation is useful to the objectives of an experiment, but such regulation “is often [nonessential for well-controlled experiments],” as in the case of ecological field experiments; [9] if there is high potential for “interactions or contamination among units,” then the experimental unit has been inappropriately defined or constructed, results will be compromised and resources wasted. It should be easy to locate in SK where the corresponding problematic statements are found. (p. 435)

Hurlbert opens up his reply by claiming that pseudoreplication is not a doctrine [1]. A doctrine is merely a set of beliefs that one adheres to and pseudoreplication is surely that. In retrospect, we agree with Hurlbert that pseudoreplication is not merely a neologism for “experimental confounds” or “failure of statistical independence,” although others do interpret pseudoreplication this way (see Freeberg and Lucas, this issue). We take no issue with general concerns about the assumption of independence, or detecting and correcting confounding influences (see, e.g., Rosenthal, 1976). If this were all that pseudoreplication meant, we would have no objection other than the introduction of an unnecessary term.

Unfortunately, DP also comprises a set of specific practices and advice about how to avoid these issues through experimental design and analysis. Certain types of experimental design are to be avoided, data from subjects within experimental units is to be averaged, never pool across levels, interspersion is to be prioritized over randomization, and physical control is not essential. As we have shown, this advice can lead to diverse errors in the design and analysis of experiments.

Statistical Independence

Informally, the assumption of statistical independence means that taking one measurement has no effect on, and provides no information about, the next measurement taken (relative to any effects not accounted for by factors shared in common). Consider, for example, fish sharing the same pond. At any given time, sampling one of the fish and measuring its body size will not provide any information about the body size of the other fish. These measures are independent. Suppose further that body size is determined by an intervening variable, such as pesticides in agricultural runoff. This factor may be completely unknown to the experimenter until more studies have been done. However, even in this case body size is still an independent measure: the runoff constitutes a shared effect for all fish in the pond. Deviations of each individual fish from the population mean are still random (assuming no other systematic variables are affecting fish size), and valid inferences can be made using standard statistical techniques.

The assumption of statistical independence is violated when these other factors are nonrandom. Suppose that a researcher is testing aggression or dominance between pairs of fish in small riffs in a stream. With a robust personality test for fish, the fact that the first fish scored extremely high on some scale of dominance may provide information about the neighboring fish. These measures may not be independent relative to social interactions (e.g., see Krause, Lusseau, & James, 2009 for the analysis of social behavior). What we see from these simple thought experiments is that the mere fact that two fish share the same pond (experimental unit) does not tell us whether statistical independence is violated.
To illustrate this point, we will consider this problem more formally using the general linear model for a one-way ANOVA in (1)

\[ y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \]  

(1)

where measurements of units \( e_{ij} \) are modeled by the variable \( y_{ij} \), which consists of three additive components: the effects, \( \mu \), of factors affecting measurements that are shared in common by all units (e.g., fish); the effects, \( \alpha_j \), of factors affecting measurements that are shared in common by units within the treatment conditions, \( j \); and the effects, \( \varepsilon_{ij} \), of all other factors affecting measurements that are neither shared in common, \( \mu \), nor are treatment factors, \( \alpha_j \).

In statistical inference, the assumption of independence means that the magnitude of the errors, \( \varepsilon_{ij} \) (i.e., all the factors not accounted for by \( \mu \) and \( \alpha_j \)) are statistically independent of each other. This assumption can be violated by any factors that are shared in common by two or more measurements, which are not part of \( \mu \) and \( \alpha_j \). Thus, one of the most important reasons for imposing experimental control is to maximize factors held in common across the subjects and thereby minimize \( \varepsilon_{ij} \).

Hurlbert, however, presents a radically different definition of independence in the context of defining an experimental unit:

\begin{align*}
\text{Experimental unit.} & \quad \text{"The smallest system or unit of experimental material to which a single treatment (or treatment combination) is assigned by the experimenter and which is dealt with independently of other such systems under that treatment at all stages in the experiment at which important variation may enter. By "independently" is meant that, aside from both receiving the same treatment, two systems or experimental units assigned to the same treatment will not be subject to conditions or procedures that are, on average, more similar than are the conditions or procedures to which two systems each assigned to a different treatment are subject." (Kozlov & Hurlbert, 2006)}
\end{align*}

Hurlbert’s definition of independence appears to neither imply nor be implied by statistical independence. In the general linear model above, error terms for each measurement are independent if and only if

\[ P(\varepsilon_{ij}) = P(\varepsilon_{ij}|\varepsilon_{nk}) \]  

(2)

where the probability of \( \varepsilon_{ij} \) is equal to the conditional probability of \( \varepsilon_{ij} \) given any other error term \( \varepsilon_{nk} \). In a one-way ANOVA, this is assumed to hold for measurements of all units across treatment conditions. Hurlbert’s concerns about average similarity of units within treatments fails to guarantee that independence will hold for all error terms.

Formally, suppose \( S \) is a similarity measure such that for any two experimental units \( e_u \) and \( e_{uk}, i \neq k \). If \( e_u \) and \( e_{uk} \) are exactly similar to each other in the conditions and procedures that they are subjected to, then \( S(e_u, e_{uk}) = 1 \). If they are completely dissimilar, then \( S(e_u, e_{uk}) = 0 \). Hurlbert’s definition of independence requires the comparison of average similarity of pairs of experimental units within a treatment is equal to the averages in all other treatments. The average similarity of all pairs of experimental units is \( E[S(e_{ij}, e_{jk})] \), where \( j \) is the treatment condition and this expectation is defined in (3).

\[ E[S(e_{ij}, e_{jk})] = \frac{\sum_{i=1}^{n-1} \sum_{j=1}^{n} S(e_{ij}, e_{jk})}{m_i} \]  

(3)

For any two treatment conditions \( j \) and \( h \) (\( j \neq h \)), Hurlbert’s definition of independence states that neither \( E[S(e_{ij}, e_{jh})] \) nor \( E[S(e_{ij}, e_{hj})] \) nor \( E[S(e_{ij}, e_{ik}, e_{jk})] \) violates the assumption of statistical independence. Put simply, the average similarity for all pairs of experimental units in treatment \( j \) should equal the average similarity of all pairs of experimental units in every other treatment condition \( h \). This condition is not found in formal accounts of statistical inference, so what is its relationship to statistical independence?

Hurlbert’s definition of independence would allow \( P(\varepsilon_{ij}) \neq P(\varepsilon_{ij}|\varepsilon_{nk}) \) so long as the average similarity of conditions and procedures is the same across treatment conditions for units, violating the probability theory requirement for independence in equation (2). In other words, some units could be more similar in conditions and procedures than others so long as the average similarity is the same across treatments. Thus, Hurlbert’s definition of independence (Kozlov & Hurlbert, 2006) is not consistent with the definition of independence in probability theory.

Interconnectedness and Experimental Units

As the above arguments show, spatiotemporal proximity does not automatically lead to statistical dependence and certainly not in a way that prohibits appropriate statistical inferences. Whether a given experimental design violates this assumption can be settled after the experiment is over, for example with a multilevel model that rules out the effect of some possible level of analysis (see Reply to Coss). The basic truth of our argument is best illustrated in the context of our hypothetical urn experiment. In his reply, Hurlbert attempts to define away this example by offering a new definition of experimental unit (Kozlov & Hurlbert, 2006), which he claims does not apply to urns because urns by definition cannot be experimental units. The measurement of the color of marbles in this case is an example of completely independent measurements. Spatial relationships, with appropriate controls, introduce no dependencies. We do not see how the newer definition would exclude an urn, or a ladle of balls taken from the urn, but Hurlbert’s argument begs the question. Further discussion of the issue of interconnectedness across the levels of analysis is clearly necessary.

In his reply, Hurlbert asserts that he does not label certain types of experimental designs as invalid [2], and that he has never argued that spatiotemporal proximity violates the assumption of statistical independence [3]. In this case, it is worth reconsidering the original article. Hurlbert (1984), lists different types of experimental designs in his Figure 1, and after several pages of describing why designs of Type B are pseudoreplicated because their spatiotemporal structure violates statistical independence, he nicely summarizes his view:

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 [emphasis added] (see Figure 1).
If one uses the data from such experiments to test for treatment effects, then one is committing pseudoreplication [emphasis added]. Formally, all the B designs (see Figure 1) are equally invalid [emphasis added] and . . . at best they can only demonstrate a difference between “locations” (Hurlbert, 1984, pp. 198–199).

There are no caveats about particular research objectives. Hurlbert goes on to say that there might be subjective support for treatment effects, but that valid inferences cannot be drawn from designs of Type B.

Hurlbert also believes that his concerns about spatiotemporal proximity and thus his definition of experimental units eliminates subjects, or more generally, individual organisms, as experimental units:

“Subject” is a widely used term whose relegation to the dustbin of history would be propitious for statistics and science. Textbooks and literature on experimental design and statistics typically incorporate full understanding of the historically core term and concept of the experimental unit and present other terms and the subject matter in language fairly similar from one book to another [citations omitted]. However, there are many widely used textbooks that avoid experimental unit and much of the classical terminology of experimental design and attempt to make do with terms such as subject, participant, cluster, group, subgroup, and so on. These often are books aimed primarily at researchers in medicine, psychology, and education [citations omitted].

What if we take this view seriously? Consider, for example, a multilevel analysis of student performance on a standardized test used to compare different curricula. Students would be the units of interest, but there might be many other levels relevant to their performance, including classrooms, school districts, counties, and states. If students are not the relevant units, what are the appropriate experimental units? Classrooms? School districts? Counties? States? Which ones should be included the final analysis, and which can be excluded? There is no way to settle this question before the experiment is done.

In his reply, Hurlbert claims that there are “satisfactory criteria for drawing boundaries around experimental units” in every discipline (point [6], his reply), but never gives any criteria for any discipline (see Reply to Coss and Reply to Freeberg and Lucas). He offers examples, such as rooms, pens, aquaria, and bottles, but there is no logical argument for why they are so special. They are among the many possible levels of analysis that may be considered in a statistical analysis, but whether or not they are important depends on the outcome of the statistical analysis and the context in which the research is executed. Above all, we are not typically interested in studying rooms, pens, aquaria, bottles, and so forth. We are interested in studying the things we put into them such as rats, goats, fish, and fruit flies.

This problem is magnified in experiments where multiple subjects are nested in multiple higher levels. Consider rats housed as pairs in cages. Each cage might be nested on a shelf, and each rack might be nested in a room. For example, Hurlbert asserts that Heffner, Butler, and Reilly (1996) “got it exactly right” in their hypothetical design of an experiment that controls for the possible release of pheromones by a species of rodent by placing them in separate rooms. The problem is that animals housed in a single room are within a single experimental unit—the room.

Rats housed in the same room give rise to simple pseudoreplication (see quote above). Placing control animals in one room and treatment animals in another is also “clumped segregation,” which leads to pseudoreplication (see Hurlbert, 1984, Figure 1, B-2). So, how is it that Heffner et al. (1996) “got it exactly right”? Why are rodents housed in the same cage interconnected, but cages housed in the same room not? Is it inconceivable there was a burned out light bulb (to borrow Hurlbert’s own example) in one room but not the other? Is it impossible that the janitorial staff allowed pheromones to leak through open doors, vents, or on their clothing? There is no reason to exclude “room” as an experimental unit, and with multilevel models there is no need to do so (see point [9], Hurlbert’s reply).

Again, one might claim, as Hurlbert does, that there are “satisfactory criteria for drawing boundaries around experimental units.” If, however, Hurlbert truly accepts the value of multilevel modeling, he would have to provide very strong criteria, indeed, to avoid inclusion of higher levels of organization in any statistical analysis. Alternatively, one might reasonably argue that to have sufficient statistical power to model effects at all levels, one would need huge numbers of rats, cages, rooms, and so forth. This is part of the criticism offered by Oksanen (2001, 2004) and Underwood (1997). To truly isolate all the relevant factors would require a heroic effort and a monumental budget (see Reply to Wiley). Hurlbert (1984) went to great lengths to show how effects at higher levels of analysis, effects as simple as a burnt out light bulb, could permit nondemonic intrusion into an experiment. Under this logic, one cannot fix the highest relevant level by fiat. If certain disciplines do have criteria (we know of none), or have established tacit limits to this infinite regress by long tradition (see Reply to Freeberg and Lucas), it was done precisely to escape an infinite regress of this sort, not because effects at other levels do not exist.

It is also important to note that this problem does not merely apply to studies of schools or rats; it has broad implications for the whole of ecology. For example, from a strict reading of the original 1984 article, a pseudoreplicationist must conclude that no inferential knowledge can be properly drawn from within a single lake or watershed. We do not attribute this view to Hurlbert personally (see point [4], his reply). However, within a single stream, statistical independence is violated by interconnectedness (design of Type B-4), and all replicates (individual plots within a stream) come from the same experimental unit (i.e., the same stream, design of Type B-5). Similarly, groups of fish sampled in different reaches off the main stream suffer from clumped segregation. We see no way to argue that two fish sharing the same aquarium violate the assumption of statistical independence, while two fish sharing the same stream do not. This problem highlights the inconsistency in the application of the logic of DP.

This inconsistency extends to samples connected in time. Hurlbert (1984) assumed that samples taken at different times are automatically correlated (see point [5], his reply). To requote him at greater length:

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 . . . Dates are then taken to represent replicated treatments and significance test 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 could 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. (Hurlbert, 1984, pp. 204–205).

Use of repeated measures-ANOVA is one of the circumstances in which successive dates can be properly analyzed. In the calculation of a repeated-measures ANOVA, successive measures on the same organism are treated as independent replicates. Indeed, it is precisely the within-subjects correlations that are valuable in repeated-measures ANOVA. This method can reduce the number of subjects required and reduce the amount of variation in the model coming from within subjects effects. Violations of the independence of error terms (i.e., assumption of sphericity) can be detected by examining the variance-covariance matrices and corrected by adjusting the degrees of freedom and thereby $p$ values (Huyhn & Feldt, 1970). Often there is no violation of independence and the replicate measures are represented in the degrees of freedom.

Experimental Controls

We take it as a given that physical control is of utmost importance in deriving causal inference from experiments. We control physical conditions in the laboratory and in the field as much as possible precisely because we want to identify causal mechanisms, exclude “nondemonic intrusions,” and reduce error variance. Hurlbert states that he views physical control as “often nonessential” in ecological experiments (see points [7] and [8], his reply). It is not clear how he came to this view, but his argument is implicitly undermined by the importance of interspersion in DP. Interspersion is a form of control, exercised to homogenize experimental groups by their physical position in space or time.

We agree with Oksanen (2001, 2004) and many others that obtaining complete physical control in an ecological experiment may not be possible. Nonetheless, we also hold to the view that maximizing control is desirable. When this is not possible, especially when intervening factors are not readily obvious to human observers, the utility of multilevel modeling becomes even greater. In these models, block effects can be dealt with separately from the main effects of interest. In other situations, multilevel models might not be applicable. In these situations, one can turn to, for example, quasi-experimental designs (Campbell & Stanley, 1963).

True Replication

DP also prizes the calculation of a $p$ value for a null test: “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” (Hurlbert, 1984, p. 191). Thinking of this sort has elevated the Type I error rate into a sort of bronze bull (Cohen, 1994; Gigerenzer, Swijtink, Porter, Daston, & Beatty, 1990; McCloskey & Ziliak, 2008; Oakes, 1986;). It is a simple mistake, however, to assume that the results of a single experiment can be generalized based merely on rejection of the null hypothesis or the $p$ value (Cohen, 1994; Oakes, 1986).

This point is important and bears repeating: the probability of replication is not contingent on the $p$ value of from any given test. One cannot use $p$ values to make any inferences about whether a Type I error has occurred, or whether a given result will generalize to other situations. Ultimately, true replication occurs only across multiple experiments.

Reply to Coss

Richard Coss helps to put our observations into a broader historical and philosophical context (also see, Oksanen, 2001, 2004). His assertion that particular bits of methodological and statistical advice are testable is well taken. Our results in the target article show that the advice offered to solve the problem of pseudoreplication can lead to errors. The result, for the scientist, is double jeopardy: a Type II error and no way to detect it.

As Coss mentions, one of the uses of replication within experiments is to estimate error variance. In multilevel models, one can use this estimate to assess the effects of different levels and to detect block effects. Hurlbert (1984), however, urged us to average measurements within experimental units, to get a better estimate of error variance. Averaging, and emphasizing the degrees of freedom in a test, suggests that multilevel analyses are error prone and likely to increase the Type I error rates. This is not correct, as we showed in the target article.

If the measurements are not averaged and measurements within experimental units are used in the analysis, then we can explicitly represent and test whether there are experimental units by introducing them as nested blocks of measurements, illustrated in (4).$$y_{ikj} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk} \quad (4)$$

The variable $y_{ikj}$ is a measurement made within each block, and it is modeled linearly as consisting of shared effects, $\mu$, nested blocks effects, $\beta_{ij}$, within fixed effects $\alpha_i$, and individual measurement errors, $\epsilon_{ijk}$. The block effects, $\beta_{ij}$, are the experimental units that concerned Hurlbert (1984), and may or may not have a significant effect. If it were true that we can accept or reject effects based on $p$ values alone (Hurlbert, 1984), nonsignificant block effects would constitute sufficient reason to reject the existence of certain experimental units. As we showed in the target article and above, if $p$ values are sufficiently high, pooling across Hurlbert’s experimental units is justified.

Reply to Wiley

R. Haven Wiley in turn makes several valuable points about the inevitability of trade offs in experimental design. This very problem has been stressed in the context of pseudoreplication by Oksanen (2001, 2004), but a few points are worth expanding upon. Experimental designs of the sort demanded by Hurlbert (1984) are, in many cases, simply impossible. Whether the study is of an endangered species, the effect in question is limited in time or space (say, an oil spill in a particular bay or stream), or the pragmatic issues involved are restrictive, not every experiment can achieve total control (also see Campbell & Stanley, 1963).

Wiley also makes an important point about the stress placed on setting criteria for rejecting the null hypothesis under different modes of inquiry. It is worth mentioning that setting the critical $p$ value at 0.05 is meant to control the overall rate of Type I errors, and does not necessarily restrict or eliminate the probability of a Type I error in any given experiment. Because we prize true
replication across multiple experiments, we believe that the rate of Type II errors is at least as damaging to science as a whole.

There are many ways out of this morass. One is to simply reject null hypothesis testing altogether (see, e.g., Cohen, 1994; Furr & Rosenthal, 2003; Gigerenzer et al., 1990; McCloskey & Ziliak, 2008; Oakes, 1986) and test actual theories and real hypotheses. Goodness of fit testing within optimality studies provides a simple example of this approach, and contrast analysis provides another method in this regard (Furr & Rosenthal, 2003). Another approach is to coordinate the use of several experiments to test potential confounds (e.g., Freed-Brown, King, Miller, & West, 2006). Under the latter approach, each test must be, alone or in combination, severe (Mayo, 1996). Indeed, Deborah Mayo has written extensively on this latter stratagem, and we recommend her work highly.

Finally, Wiley takes us to task for misunderstanding his point about the risks of large sample sizes in accidentally creating, rather than automatically avoiding, small but truly random biases in experiments. Perhaps our reading was too facile, and indeed the multidimensionality inherent in optimizing sample sizes is problematic, as he points out. We were not, however, advocating large sample sizes. The point, though admittedly not clear, was that small and moderate block effects might be missed with small sample sizes. We agree with Wiley that large sample sizes can detect effects with small size that have little or no impact on the results of interest. Whether or not an effect—small or large—matters depends on the research context and the theoretical background. For example, small effects may be important if they are in a nonlinear system highly sensitive to perturbations.

Our worry in this case, though, is not small, truly random biases emerging from the study design. Rather, if an experiment “fails,” it might be due to either an incorrect hypothesis, or a systematic bias. We stressed that maximum interspersion, long advocated on a priori grounds by pseudoreplicationists, seems ideally suited to avoiding the detection of exactly the sorts of systematic biases that it is intended to prevent. In designing an experiment, it is clear that both sample size and the arrangement of blocks will contribute to the isolation of effects not yet understood and not yet under adequate experimental control.

Reply to Freeberg and Lucas

Todd Freeberg and Jeff Lucas offer a spirited defense of DP. We largely agree with their opening paragraph. If DP were just the claim that “data should be independent when the statistics call for independence,” then this would be fine, but why give the problem of statistical independence a new name? Hurlbert, however, does not agree with this definition as he explicitly states in his commentary. We strongly disagree with Freeberg and Lucas’ claim that “one consequence of requiring that statistics match experimental design is that it is more difficult to design acceptable experiments.” The real problem is to find or develop acceptable statistical analyses for the experiments we design (viz., Campbell & Stanley, 1963; Furr & Rosenthal, 2003).

Turning to the interesting point they raise about playback designs, as we stated near the beginning of our paper, “The question of pseudoreplication is whether data gathered with any degree of spatiotemporal proximity can be said to be independent. The standard answer to this question is, and always has been, ‘yes,’ if appropriate experimental controls were implemented, and if you can detect contaminating influences, either within experiments or across multiple experiments.” A simple example should suffice to illustrate the endless downward spiral of experimental units.

In this case, we shall consider playback studies of birdsong. Coss (this issue) highlights some of the many difficulties involved in setting up such a study. Much advice has been offered on this sort of design (e.g., McGregor et al., 1992), most of it well within the genre on experimenter effects (Rosenthal, 1976). Unfortunately, even if you follow this advice you cannot escape from pseudoreplication in any playback study.

The reason for this problem will be clear once a few more of the possible experimental units are illustrated. Songs are nested within individual birds. Birds are often nested within social neighborhoods or local populations, which are in turn nested within regions or across large areas (as with migratory birds). Birds are also nested in time, and there are large seasonal and diurnal effects on many characteristics of song (see, e.g., Liu & Kroodsma, 2007). Finally, direct comparisons between all possible playback stimuli are time and cost prohibitive. Several strategies have been developed to deal with this issue (e.g., the Latin Square), but they only work under very strict assumptions. Hurlbert (1984), for example, assumes that any intrusion or contamination is unlikely to correspond with the spatial scales used in his regular interspersion designs, an assumption that will automatically fail when the contamination comes from the individuals within blocks. See Torger son, (1985) for a more complete discussion of such assumptions and how to manage sampling layouts in the context of comparative judgment, as with playback studies.

At the limit, any experiment about birdsong that relies on recordings or playback will sample only a tiny fraction of the available songs, birds, locales, times, seasons, and so on to gather and playback their exemplars. It is hard to invent a richer opportunity for “nondemonic intrusion” (Oksanen, 2004). There is little reassurance available, other than long tradition, or an author’s word, that the exemplars chosen indeed sample the repertoires, individuals, populations, times, and seasons efficiently and exhaustively. But let us grant that it is possible to do so, or at least that one can avoid the need for a completely exhaustive sampling (Torgerson, 1985). We have still not found all the levels of analysis (and hence, potentially relevant experimental units).

Each experimenter will be using a recording device and a playback device. All of these devices necessarily truncate or change the song stimuli, and introduce artifacts, both during recording and during playback. This is true of both analog (Espmark & Fonstad, 1983) and digital (Joachim & Goodale, 2005; Logue, Gammon, & Baker, 2005) recording methods. That is, even with the best technology, animals exposed to playbacks will always have some level of potential stimulus artifact imposed upon them.

Thus, the recording and playback method serves as the next logical experimental unit in any playback experiment. Without parametric variation of recording and playback device, there is ample opportunity for the artifacts introduced by equipment to intrude into an experiment without being detected (see Hurlbert’s, 1984, example of the burnt out light bulb). In some cases, intrusion of a stimulus artifact can be quite helpful (Hubel, 1982), but in most cases, artifacts are not desirable. Should we change the degrees of freedom, to reflect the number of tape recorders used in an experiment?

457COMMENTARIES
We are ultimately left with only the expertise, judgment, and experience of field workers to assure us that any given playback experiment adequately controls all of these possible intrusions, and that the equipment is not itself an experimental unit. Unfortunately, we cannot rely solely on the experiences and traditions of researchers. Even if experiences and traditions are largely valid, mistakes can be made. We have to shortcut this infinite regress of levels of analysis, and rely on actual replication, between distinct experiments.

Conclusion

The degree of independence at a given level cannot be deduced on the basis of spatiotemporal proximity, but can be tested after the experiment is done. There is no logical limit on what levels might be important in an experiment, and thus, there are no a priori criteria for demarcating a given level as the level of experimental units. To escape this infinite regress of levels, true replication across multiple experiments is necessary with increasing levels of precision and control over time. Indeed, true replication is the only way to identify and correct Type I errors.

Our overriding concern is that statistical analyses should be designed to fit experiments, and not vice versa. Because of this, it is impossible to evaluate experiments and their analyses before the results are in. If a given data set does not conform to the assumptions of a particular test, labeling the study as “pseudoreplicated” is of no help. Our approach places considerable burden on researchers, reviewers, and editors. They must attend to the assumptions of various tests, estimate the effects of assumption violations on statistical calculations, and be open to new and innovative methods. This is not just a burden, though, but also a substantial opportunity for innovation in experimental design and analysis.

There is considerable interest in improving methods in behavioral and ecological research, as evidenced by the wide impact Hurlbert’s (1984) article has had. This discussion is just one step on the way toward more empirical and analytical clarity. We hope that we have illustrated the basic problems with the advice offered to avoid pseudoreplication. If nothing else, it should increase interest in investigating how methods work, validating much beloved but seldom questioned traditions, and looking for circumstances under which particular methodological assumptions lead to errors.

References


Received July 21, 2009
Revision received August 13, 2009
Accepted August 14, 2009