

PSEUDOREPLICATION REVISITED

Robert A. Heffner,¹ Mark J. Butler IV,^{1,2} and Colleen Keelan-Reilly¹

In 1984 Stuart Hurlbert published a review of the ecological literature wherein he scrutinized 176 experiments from 156 papers published during 1960–1980 for evidence of pseudoreplication (Hurlbert 1984). Pseudoreplication is defined by Hurlbert (1984: 18) 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 experimental units are not statistically independent.

His findings were disturbing. Of the 176 manipulative field experiments reviewed, 27% were guilty of pseudoreplication. Considering only the 101 studies applying inferential statistics, 48% were pseudoreplicated. More recently, Hurlbert and White (1993) reported that the frequency of pseudoreplication was 32% in papers describing invertebrate zooplankton research published in 1986–1990. Other reviews of statistical errors common in the ecological literature include those by Innis (1979) and Underwood (1981). Innis (1979) estimated that 20% of the scientific papers surveyed by students in a course on quantitative methods contained statistical or calculation errors. Underwood (1981) found that 78% of the papers on marine biology that he reviewed and that employed analysis of variance contained statistical errors of some sort. In addition to these reviews, there are numerous articles that warn of the lack of appreciation among ecologists of basic statistical issues, including Type I and II errors (Seaman and Jaeger 1990 and associated responses), power (Toft and Shea 1983, Peterman 1990), and adherence to parametric assumptions (Potvin and Roff 1993), to name a few (see Potvin and Travis [1993] for a recent bibliography). Hurlbert counseled us on replication.

A true “replicate” is the smallest experimental unit to which a treatment is independently applied. According to Hurlbert’s review (1984), pseudoreplication

most commonly results from wrongly treating multiple samples from one experimental unit as multiple experimental units, or from using experimental units that are not statistically independent. The implication of these errors is that chance events directly affecting one experimental unit are more likely to affect other experimental units within that treatment group than experimental units in other treatment groups. Although the definition is concise and seemingly straightforward, the concept of an experimental unit is perhaps best understood through example. Hurlbert (1984) provided several examples in his monograph, and we offer three more.

Mesocosms, within which some variable (e.g., nutrient level) is manipulated, are often the appropriate experimental unit in studies incorporating this useful methodology, but the individual samples or measurements (e.g., collections of phytoplankton) taken from within a mesocosm are not independent replicates. The experimental manipulation of nutrient concentration in the water, for example, is accomplished by altering conditions in an entire mesocosm, so the mesocosm itself is the smallest unit to which the treatment is independently applied.

Sometimes, experimental units are less easily distinguished, as may be the case when natural sampling units are used. One might remove sea urchins, for example, from boulders at different field sites and compare the abundance of benthic algae (an urchin food source) in several quadrats on each boulder with similar measurements taken from unmanipulated “control” boulders where urchins are still resident. The treatment effect of interest involves the manipulation of sea urchins. Thus, the experimental unit is the set of individual boulders at one site, not the quadrats from which measurements of algal growth were taken. The appropriate designation of a “replicate” may also depend on the hypothesis to be tested and the scale of inference desired. Are differences in the results among field sites of interest or only the general difference between boulder treatments (with field sites treated as a blocking variable)?

Pseudoreplication is an insidious beast, and although some occurrences are clear-cut, others are more subtle and require knowledge of the system to be studied if the problem is to be avoided. Let us say, for example, that one wanted to test whether a certain compound found within the young, growing shoots of annual grasses was responsible for the spring onset of the reproductive season in female rodents that eat these grasses. One might design a laboratory study in which female rodents are placed individually in 20 cages, half of which are chosen randomly and supplied with rodent

¹ Old Dominion University, Department of Biological Sciences, Norfolk, Virginia 23529 USA.

² Address reprint requests and correspondence to this author.

chow with the grass compound added and the other half with unaltered rodent chow (the "control"). Rodents are then examined daily for signs of estrus. Treatments are clearly replicated in this design. There are 20 cages with a single rodent in each; 10 cages are supplied with the compound and 10 cages are not. However, the replicates are probably not independent unless the cages are situated far apart in separate rooms. Why? Mammalogists know that estrus can be induced in female rodents via airborne chemical cues released by other females already in estrus. Thus, 20 caged rodents held in a single laboratory room will not respond independently to treatments affecting reproductive cycles. These examples only scratch the surface of what could be an exhaustive list of design or analysis infractions that are collectively referred to as "pseudoreplication."

The relevance of this error of pseudoreplication, considering that the primary function of statistics in experimental work is to "... increase clarity, conciseness, and objectivity with which results are presented and interpreted" (Hurlbert 1984:189), is now clear for many ecologists. Without proper replication an investigator's scope of inference and resulting conclusions are limited or invalid. Improper replication usually results in the underestimation of true variation or the confounding of its sources, thereby increasing the risk of a Type I error (i.e., the chance of rejecting a null hypothesis that is true).

Hurlbert (1984) defined four types of pseudoreplication: simple, temporal, sacrificial, and implicit pseudoreplication. *Simple pseudoreplication* was the most common form of pseudoreplication in the field experiments Hurlbert examined. It occurs when samples from a single experimental unit are treated as replicates representing multiple experimental units. Typically, inferential statistics are then erroneously applied to these samples and used to support conclusions drawn from what are essentially unreplicated treatment groups. *Temporal pseudoreplication* occurs when an experimental unit is sampled repeatedly through time and the samples treated as if they represented independent experimental units. *Sacrificial pseudoreplication* results when an investigator pools multiple samples from multiple experimental units under the same treatment prior to statistical analysis, which confounds two sources of variation within the data set (i.e., variance among samples within an experimental unit and variance among experimental units). Lastly, *implicit pseudoreplication* refers to manipulative studies with unreplicated but subsampled treatments where tests of significance are not directly applied but the "significance" of treatment effects is nonetheless discussed, often with reference to graphs showing treatment means with non-overlapping standard errors or confidence limits. A thorough

explanation of these four types of pseudoreplication can be found in Hurlbert's paper (1984).

Hurlbert's incisive description of the central tenets of proper experimental design for field studies and his convincing documentation of the ubiquity of the problem in the ecological literature struck a chord among ecologists. His 1984 paper in *Ecological Monographs* is recognized as a science citation classic (Hurlbert 1993) and has been cited in >600 published articles. The American Statistical Association also honored Hurlbert's contribution with the Snedecor Award for the best paper in the field of biometry in 1984. The term "pseudoreplication" is now in the lexicon of biologists and statisticians.

Yet it remains to be seen whether the experimental design and statistical analysis employed in ecological field studies have improved in the decade since Hurlbert's review. At that time he suggested that ecologists could be made more aware of misapplied statistics, and pseudoreplication in particular, if (1) statistical texts provided clearer, non-technical descriptions and examples of proper experimental design and (2) editors of scientific journals became more knowledgeable of statistics and more hard-nosed about accepting flawed manuscripts. A few years ago, Hurlbert stated that "... this [critiquing of statistical practice] remains a fertile field of endeavor" (Hurlbert 1993). The purpose of this paper is to assess the current state of pseudoreplication in ecological field experiments.

Methods

We examined the experimental design of papers recently published in the same well-known, ecological journals originally reviewed by Hurlbert. All of the manipulative ecological field experiments found in articles from the 1992 volumes of *Ecology*, *American Midland Naturalist*, *Limnology and Oceanography*, *Journal of Experimental Marine Biology and Ecology*, *Journal of Animal Ecology*, *Canadian Journal of Fisheries and Aquatic Sciences*, and the *Journal of Mammalogy* were examined for evidence of pseudoreplication. Several 1991 issues of the *American Midland Naturalist* and the *Journal of Mammalogy* were also included to increase the sample size for these journals in the analysis. Following criteria set by Hurlbert, we initially scanned each article to determine: (1) if it was a manipulative study, (2) if it was a field experiment, and (3) if inferential statistics were used for data analysis. *Manipulative experiments* involve direct manipulation of the independent variable in such a way that the experimental units can be randomly assigned to treatment groups. *Mensurative studies*, in which treatment groups are not randomly assigned and tests are of differences among physical locations or points in time and not designated treatments, were not evaluated

TABLE 1. Assignment of recently published ecological field experiments to pseudoreplication type, by subject area and by journal for the years 1991 and 1992. A total of 892 articles were reviewed; 119 of them met our criteria for further review for instances of pseudoreplication. The number of studies classified as questionable pseudoreplication and the frequency of pseudoreplication when this group is considered pseudoreplicated appear in parentheses.

	No. of articles reviewed	Pseudoreplication type				Frequency of pseudoreplication (%)
		Simple	Temporal	Sacrificial	Implicit	
A. Subject matter						
Terrestrial plants	19	3 (2)	0	0	0	16 (26)
Terrestrial invertebrates	19	2	0	0	0	11
Terrestrial vertebrates	24	2 (1)	1	0	0	13 (17)
Freshwater nekton	11	1	0	0	0	9
Other freshwater organisms	15	2	1	0	0	20
Marine benthic organisms	14	0	0	0	1	7
Other marine organisms	17	1	0	0	0	6
B. Journal						
<i>Ecology</i> (1992)	38	2 ^a (1)†	0	0	0	5 (8)
<i>American Midland Naturalist</i> (1991 and 1992)	15	3 ^b (1)‡	0	0	0	20 (27)
<i>Limnology and Oceanography</i> (1992)	4	1 ^c	0	0	0	25
<i>Journal of Experimental Marine Biology and Ecology</i> (1992)	25	0	1 ^d	0	1 ^e	8
<i>Journal of Animal Ecology</i> (1992)	16	3 ^f	0	0	0	19
<i>Canadian Journal of Fisheries and Aquatic Sciences</i> (1992)	15	2 ^g	0	0	0	13
<i>Journal of Mammalogy</i> (1991 and 1992)	6	0 (1)§	1 ^h	0	0	17 (33)

Sources: ^a Ehrlén, J., pp. 1820–1831; Harvell, C. D., pp. 1567–1576; ^b Bollinger, E. K., et al. (1991), pp. 114–125; Hazlett, D. L., pp. 276–289; Yahner, R. H., pp. 381–391; ^c Levine, S. N. and D. W. Schindler, pp. 917–935; ^d Kaartvedt, S. and E. Nordby, pp. 279–293; ^e Borsa, P., et al., pp. 169–181; ^f Volke, W., pp. 273–281; Bolton, M., et al., pp. 521–532; Gibbons, D. W. and D. Pain, pp. 283–289; ^g Deegan, L. A. and B. J. Peterson, pp. 1890–1901; Rand, P. S., et al., pp. 2377–2385; ^h Dietz, B. A. and G. W. Barnett, pp. 577–581; † Megonigal, J. P. and F. P. Day, pp. 1182–1193; ‡ Davis, M. A., et al. (1991), pp. 150–161; § Simons, L. H. (1991), pp. 518–524.

in our review. See Hurlbert (1984) for more information on the distinctions between manipulative and mensurative studies.

We considered an experiment to be a *field study* if the manipulation was physically conducted outside in a natural setting where many environmental variables were not controlled; we included mesocosm experiments in our review. Only manipulative field studies employing inferential statistics were further reviewed. If a paper describing more than one study included at least one manipulative field study, we then considered the paper in our review. If a paper reported multiple manipulative experiments, we counted it as a case of pseudoreplication if at least one statistical analysis included that error.

We evaluated papers for pseudoreplication by asking whether treatments were applied randomly to experimental units and whether the replicate experimental units of each treatment group were likely to be independent. We considered the study to be pseudoreplicated if the data collected from experiments that violated one of these conditions were used to test for, or imply that, treatment effects differed significantly among treatments. Each of these papers was then placed into one of the four categories defined by Hurl-

bert (i.e., simple, temporal, sacrificial, or implicit). Articles with vague descriptions of experimental design or statistical procedures were not included in the tabulation; this occurred in <1% of the articles reviewed.

Following our initial screening, we contacted the authors whose papers we had initially identified as being pseudoreplicated to give them the opportunity to comment on or rebut our conclusion. We sent each of them a copy of our results, a copy of the abstract of this manuscript, and a description of the specific error we found in their papers; authors had four weeks to respond to our request for further information. Following our receipt of comments from the authors, we reviewed each article yet again to reconsider our conclusions in light of the clarifications provided by the authors.

Results

We reviewed 892 articles, 119 (13%) of which met our criteria as manipulative field studies employing inferential statistics, which is a sample size similar to that examined by Hurlbert (1984; $n = 101$). These articles were evaluated for pseudoreplication; 14 (12%) of the 119 papers reviewed by us included pseudoreplicated studies. Three additional studies were placed in a category we call “questionable pseudoreplication”

(Table 1). If those studies are considered to be pseudoreplicated, then the frequency of pseudoreplication is slightly higher (14%). Both of these values are markedly lower than the rate of pseudoreplication (48%) reported by Hurlbert a decade earlier, but the incidence of pseudoreplication today remains disturbingly high. We suspect that Hurlbert's message, warning of the consequences of pseudoreplication, has been heeded by many ecologists and we found evidence of this in the frequent citing of Hurlbert's monograph in the papers we reviewed. Although the experimental design of ecological field studies may have improved over the last decade or so, the fact remains that one published paper in eight involves pseudoreplication. Our survey found simple pseudoreplication (9% without questionable cases, 12% with questionable cases) to be the most common type, with temporal (2%) and implicit (1%) pseudoreplication occurring less frequently. We found no instances of sacrificial pseudoreplication.

The frequency of pseudoreplication in our sample dropped from 29% to 12% between our initial and final readings (with the author's comments in hand) of the studies. The reasons for this change came from expected and unexpected sources. Some papers were incorrectly interpreted by the authors of this review due to missing ANOVA tables and/or explicit statements of the number of degrees of freedom used in the analysis. We also found that some of the descriptions of the experimental design led to incorrect determinations of the appropriateness of the statistical tests applied to the results. If the reader is unable to properly evaluate a study's experimental design, even if replicates have been correctly identified, the conclusions drawn from that study may be viewed with an undeserved skepticism.

Other papers, initially included in our tally, were later dropped from our sample because they did not meet the criteria of being a manipulative study. The independent variable under investigation was one that could be manipulated, but in these cases, the authors took advantage of an existing condition that allowed the variable of interest to be examined. The source of the manipulation was not explicitly stated and led to misclassification on our part. For example, a study may look at the difference in the foraging distances of ants in a mowed and unmowed field. If the mowing treatment was applied by the researcher the study is clearly manipulative; however, if the researchers are working in mowed and unmowed fields that exist coincident to their study, the "experiment" is a mensurative one. Without explicit statements concerning the source of the manipulations, it is a matter of chance whether or not the study is properly classified in this regard.

Discussion

Why does pseudoreplication still occur? There must be several reasons. Several authors who responded to us had not read Hurlbert's monograph and seemed generally unfamiliar with issues of experimental design and analysis. Another likely answer, it seems, is that it is widely held that statistical analyses add some measure of quantitative rigor to a study—even if such statistics are inappropriate under the circumstances. At best, such analyses yield the vague statistical result that there is a "treatment effect" that cannot be statistically separated from a "location effect." As Hurlbert stated in his 1984 paper (p. 190): "It will 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." It was clear from the titles and discussion of several papers that the authors were interested in broader ecological questions regarding potential treatment effects, as opposed to location effects, which are of little interest to readers and editors. There is little appeal, for example, to a study in which plants in two fields are cataloged for several years, and such results are not likely to be published in reputable journals. However, the effect of fire on species composition of plant communities is of broader interest. The problem arises when the comparison is of one field that experienced a fire and a second that did not.

As mentioned above, a few articles in our sample fell within what we believe to be a "gray area" with respect to what is commonly considered pseudoreplication. These articles contain experiments that are unreplicated. The authors were aware of this, at least at the time the manuscript was submitted, if not when the study was conducted. Recognizing the problem in their studies, they offered various caveats in the text to avoid the pseudoreplication label. Somewhere in each paper the authors state either that the study was unreplicated, that the statistical results should not be used to infer specific treatment effects or at least should be viewed skeptically, or that their conclusions about specific treatment effects is based on logic and biological intuition, rather than statistical inference. But the titles, subtitles, and focus of the discussion in these papers centers on statistically significant "treatment effects." Are these studies pseudoreplicated or not?

Hurlbert did not face this dilemma in his original literature review. Of course, the problem was not as generally recognized as it is today, and certainly none of the papers he read referred specifically to pseudoreplication, and so could not offer the compulsory language to absolve them of the error. We chose to classify these articles in a separate category (questionable pseu-

doreplication) and leave it to the reader to decide whether the peculiar circumstances presented in these few papers lands them squarely within the realm of pseudoreplication or not. Our position is not an abrogation of responsibility; rather, it reflects our judgment that this is truly a gray area within the current definition of pseudoreplication.

We also recognize that some have taken an unflinching view of unreplicated experiments, dubbing each with the ignominious term of pseudoreplication, without recognizing the inherent scientific value of many such studies. We do not condone such an approach and neither did Hurlbert (1984:188) in his original review: "... the quality of an investigation depends on more than good experimental design. . . . Most of them, despite errors of design or statistics, nevertheless contain useful information." Others have also argued for the logistical necessity and merit of unreplicated ecological studies (Hawkins 1986, Carpenter 1990). ^{But} After all, it is reasonably certain that the earth revolves around the sun and science came to know this through means other than replicated experiments! Yet, depending on the circumstance and questions of interest, some unreplicated studies can be analyzed using statistical techniques such as time-series analysis (Jassby and Powell 1990), resampling-based analyses (Crowley 1992), ANOVA (Underwood 1994), and analyses based on Bayesian inference (Reckhow 1990).

Periodic scrutiny, whereby we drag "... statistical malpractice into the sunshine" (Hurlbert 1993), permits us to assess the state of proper statistical analysis and experimental design in our science and ensures progress towards increasing statistical savvy among ecologists. Clearly, there is progress still to be made in the areas of identifying independent experimental units and designing field experiments. Hurlbert had hoped that his review would stimulate a reduction in the frequency of pseudoreplication in the ecological literature. Our review suggests that his hopes have been at least partially realized.

Acknowledgments: We thank Raymond Alden, Gerald Levy, Robert Rose, an anonymous reviewer, and especially Stuart Hurlbert for their comments on earlier

versions of this manuscript. We are also grateful for the prompt and generally courteous responses we received from authors whose papers we reviewed and cited.

Literature Cited

- Carpenter, S. R. 1990. Large-scale perturbations: opportunities for innovation. *Ecology* 71:2038–2043.
- Crowley, P. H. 1992. Resampling methods for computation-intensive data analyses in ecology and evolution. *Annual Review Ecology and Systematics* 23:405–447.
- Hawkins, C. P. 1986. Pseudo-understanding of pseudoreplication: a cautionary note. *Bulletin of the Ecological Society of America* 67:184–185.
- Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187–211.
- . 1993. Dragging statistical malpractice into the sunshine. *Current Contents* 24(12):8.
- Hurlbert, S. H., and M. D. White. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analysis. *Bulletin of Marine Science* 53:128–153.
- Innis, G. S. 1979. Letter to the editor. *Bulletin of the Ecological Society of America* 60:142.
- Jassby, A. D., and T. M. Powell. 1990. Detecting changes in ecological time series. *Ecology* 71:2044–2052.
- Peterman, R. 1990. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences* 47:2–15.
- Potvin, C., and D. A. Roff. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? *Ecology* 74:1617–1628.
- Potvin, C., and J. Travis. 1993. Concluding remarks: a drop in the ocean. *Ecology* 74:1674–1676.
- Reckhow, K. H. 1990. Bayesian inference in non-replicated ecological studies. *Ecology* 71:2053–2059.
- Seaman, J. W., and R. G. Jaeger. 1990. Statistical dogmatism: a critical essay on statistical practice in ecology. *Herpetologia* 46:337–346.
- Toft, C. A., and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. *American Naturalist* 122:618–625.
- Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Oceanography and Marine Biology Annual Reviews* 19:513–605.
- . 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbance. *Ecological Applications* 4:3–15.

*Manuscript received 28 November 1994;
revised 21 December 1994;
accepted 8 January 1996;
final version received 20 February 1996.*