Opinion is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise as possible and references should be kept to a minimum. A summary is not required.

On misinterpretations of pseudoreplication and related matters: a reply to Oksanen

Stuart H. Hurlbert, Dept of Biology and Center for Inland Waters, San Diego State Univ., San Diego, California 92182, USA. (shurlbert@sunstroke.sdsu.edu)

Pseudoreplication has become a widely accepted label for a certain class of statistical error common in the literature of ecology as well as of other fields. A wide-ranging critique by L. Oksanen recently published in this journal criticizes the term and concept and concludes it to be a "pseudoissue," one reflecting an intellectual disease, "a totally outdated epistemology" known as "inductionism." The present article addresses some of Oksanen's complaints. His critique is based on a misconception of pseudoreplication, reflects unawareness of most of the literature on the topic, and mistakenly argues that the seriousness of the error is a function of whether an experiment is conducted in an inductive or deductive spirit. Oksanen's advocacy of using resources available for large scale ecology more for large numbers of experiments with unreplicated treatments than for fewer experiments with modest replication of treatments is unrealistic. It is based on an overly optimistic view of the ability of a meta-analysis to compensate for deficiencies, such as very noisy estimates of treatment effects, of the individual studies that are fed into it. A definition is offered of the term manipulative experiment, since adequate ones are lacking in the literature. Attention is called to the fact that for certain types of manipulative experiments lacking treatment replication, there are valid ways to test for treatment effects.

> Authors who cite Hurlbert would do better if they had read his paper!
>
> – A.J. Underwood (1998:344)

Twenty years ago I wrote a review of a particular category of statistical error that I termed pseudoreplication, assessed the frequency with which it occurred in articles reporting ecological field experiments, and commented on related issues of experimental design and statistical analysis (Hurlbert 1984). Since that time the term pseudoreplication has become widely used, and many ecologists have become more aware of the need for close concordance of design, analysis, and interpretation of experiments. A wide-ranging recent paper titled "Logic of experiments in ecology: is pseudoreplication a pseudoissue?" (Oksanen 2001) finds many faults of logic and epistemology in my 1984 paper, and answers the question in its title in the affirmative.

If indeed pseudoreplication is a "pseudoissue", that will be a shock to the American Statistical Association, which awarded the original pseudoreplication paper the G.W. Snedecor Award for the best paper in biometry in 1984.

The present report responds to key points in Oksanen's (2001) critique but does not attempt to cover many collateral issues he discusses. I focus in particular on his misunderstanding of the nature of pseudoreplication and experiments, his crediting of me with the revival of "long dead" epistemologies, and his over-valuation of the statistical treatment of experiments lacking treatment replication.

While the present report was under review, Cottenie and De Meester (2003) also published a critique of Oksanen's (2001) key claims, reinforcing many points that will be made here.

There has been much published on pseudoreplication since 1984 (Machlis et al. 1985, Hairston 1989, Krebs 1989, Kroodsma 1989a, b, 1990, Hurlbert and White 1993, Heffner et al. 1996, Lombardi and Hurlbert 1996, García-Berthou and Hurlbert 1999, Jenkins 2002, Hurlbert and Meikle 2003). One of the best recent texts on experimental design devotes several pages to discussing various types of pseudoreplication, though without using the label (Mead 1988:107-122; reviewed in Hurlbert 1990). In its 1995 edition, one of the most widely used statistics texts quietly removed an example where pseudoreplication had long been advocated in earlier editions as the correct way of doing things (Sokal and Rohlf 1969:438, 1981:488, 1995). The problematic aspects of Oksanen (2001) derive in part from its attempt to critique the concept of pseudoreplication while ignoring most of the literature on it.

Nature of pseudoreplication

Oksanen (2001) claims that "The concept of 'pseudoreplication' amounts to entirely unwarranted stigmatization of a reasonable way to test predictions referring to large-scale systems. . .[it is] introduced as a stigmatizing label for experimental studies, where inferential statistics have been used in the context of unreplicated or compound treatments ... Referees should preferentially refrain from using [the term]."

Pseudoreplication in any of its various guises is simply an error of statistical analysis and interpretation. It is not committed only in experiments where treatments are unreplicated. One of the most common types is sacrificial pseudoreplication (Hurlbert 1984, Hurlbert and White 1993, Lombardi and Hurlbert 1996, García-Berthou and Hurlbert 1999, Hurlbert and Meikle 2003), and that is possible only when there is at least two-fold replication of treatments. What might be termed test-qualified sacrificial pseudoreplication may actually be on the rise: there are now at least three books (Hairston 1989:33ff, Underwood 1997:268ff, Quinn and Keough 2002:260ff) that, in their discussions of "pooling", essentially recommend sacrificial pseudoreplication when tests for differences among experimental units (within treatments) yield high P-values (Hurlbert 1997, Hurlbert and Lombardi 2003). Jenkins (2002) shows the error of that approach, though without reference to those books or the term pseudoreplication.

In any case, pseudoreplication indeed is a "stigmatization", but a warranted one, of statistical or interpretational errors. It seems a useful label, even if some will misuse it, as happens with everything else useful. Cottenie and De Meester (2003) concur: "Pseudoreplication is thus not a pseudoissue, but a valid and important statistical problem that should be taken into account by referees when applicable."

Compound treatments

Confusion between effects of procedures used to impose treatments and effects of chance events impinging on an experiment (= non-demonic intrusion, Hurlbert 1984) is introduced in Oksanen's (2001) discussion of what he calls "compound treatments". He states, "If the concept of pseudoreplication is used in the broader sense, including compound treatments, then all experiments are pseudoreplicated ..."

Oksanen correctly notes that in an experiment, treatment effects may result from either the nominal treatment factor (e.g. vole density) or as unintended side-effects of procedures (e.g. exclosure cages) used to impose treatments on experimental units. The sum of the nominal treatment factor(s) and the procedures is what Oksanen means by "compound treatments." He is also correct in noting that failure to distinguish procedure effects from effects of the nominal treatment factor is a potential problem in all experiments. Traditional and effective ways of dealing with the problem include: 1) the use of placebo treatments instead of "do nothing" controls (e.g. control mice get injection with saline solution, experimental mice an injection with saline solution plus a drug), and 2) the use of multiple control treatments, each controlling for one or more types of possible procedure effect. Both marine ecologists and small mammal ecologists in particular have been ingenious in devising multiple control treatments to sort out various unintended effects of cages and exclosures when these are used to manipulate densities of mobile organisms. The challenge in field experiments can be sufficiently great that some ecologists have taken the extreme position that "experiments with only two treatments are not usually much good" (Underwood 1997:139). This general problem of how one controls for procedure effects has, however, nothing to do with pseudoreplication.

Oksanen goes astray when he mistakenly states that "Hurlberts philosophy is that even compound treatments are regarded as pseudoreplication," and then refers to an example of pseudoreplication presented in Hurlbert (1984, Fig. 1, case B-4). In that example, multiple aquaria are set up under each of two treatments, but all the aquaria in a given treatment are hooked into the same water circulation system – and statistical analysis ignores this fact. The problem being illustrated has nothing to do with procedure effects or "compound treatments," and exists prior to the imposition of treatments.

The problem is that the interconnection of all tanks in a given treatment will destroy their statistical independence. In the absence of a real treatment effect, this will greatly increase the likelihood of "detecting" a spurious one, i.e. of biasing P-values downward. Intuitively this may be seen in how easily spurious treatment effects, or biased estimates of real ones, could be generated by a single incident of non-demonic intrusion. A chemical contaminant or pathogen that is accidentally introduced into a single tank could quickly spread to all tanks of one treatment but to no tank in the other treatment. The "replicate" tanks in a given treatment in such an experiment lack statistical independence to the same degree that replicate plots in an agricultural experiment would lack it if all plots for one treatment are put at the north end of a field and all those for the other treatment at the south end.

In sum, procedure effects are an issue that must be dealt with appropriately in all experiments. Pseudoreplication is a separate issue and is avoidable in all experiments.

Induction versus deduction

Oksanen (2001) embeds his critique of pseudoreplication in an extended discussion of epistemology and in particular, of the relative values of inductive and deductive modes of reasoning or scientific research. This does not seem particularly germane to a critique of Hurlbert (1984), but is consistent with Oksanen's misunderstanding of the simple technical nature of pseudoreplication.

At one point Oksanen acknowledges that "Both approaches [deduction and induction] have their roles in science. Inductive experiments can provide new, unexpected insights." But elsewhere he takes a hard stand, saying "As a method for basic sciences, inductionism [= excessive reliance on inductive reasoning?] has been dead for decades, and its resurrection in ecology in 1984 [by Hurlbert] is truly amazing ... [It is] a totally outdated epistemology."

Now I am pleased to be credited with resurrecting something as so grand-sounding as "inductionism" even if it is not found in dictionaries. Hurlbert (1984) concerns itself in no way, however, with the relative roles or importance of induction versus deduction, but only with whether analyses and interpretations of experiments are concordant with the way experiments were designed and conducted. Cottenie and De Meester (2003) also have pointed out "Oksanen's misinterpretation of Hurlbert's supposed inductionism."

Some of the 176 experiments reviewed in Hurlbert (1984) were perhaps carried out in a purely "deductive spirit" and some in a purely "inductive spirit." But most were likely hybrid in nature. Science often benefits most from experiments that simultaneously allow both testing of our existing preconceptions or theories and ample opportunity for new observations and insights. The latter may lead to new theories and generalizations. Across all sciences the greatest value of an experiment often is provided not when it confirms a preconception but rather when it contradicts one, or when a response variable that initially was regarded as a minor focus of the study behaves in a surprising manner.

It does not seem useful to draw hard epistemological lines between deductive and inductive studies or objectives. At least it is not necessary to decisions as to how treatment effects should or should not be assessed statistically. Those decisions are mostly dictated by the experimental design or sampling design that has been employed.

Since Hurlbert (1984) does not deal with this issue, I offer no further comment on Oksanen's specific views on the matter. In good conscience, however, I must quickly pass on to Ford (2000) the crown of the Prince of Inductionism Restored. His excellent treatise on scientific method gives a cogent discussion of the roles of inductive and deductive reasoning and shows how they are equally critical to the advance of science. To quote a few lines:

> There is no single method of reasoning that scientists can, or do, follow. We reason in two general ways: (1) Deductively, when we use the logic of a theory to make a deduction that we then investigate. ... (2) Inductively, to extend a theory to explain more, where we consider an idea that applies in one situation will also apply in another. ... Most reasoning in scientific research is inductive. ... The hypothetico-deductive (H-D) method is wider in scope than empirical induction because it seeks support from the deductive consequences of existing theory – but it is essentially an inductive method and predictions should not be viewed as pieces of evidence completely independent of the theory used to produce them (Ford 2000:170, 183).

Definition of 'experiment'

Throughout their histories, all the natural and social sciences have used experiment and experimental primarily to refer to any sort of empirical study or observation that is carried out in order to answer a question or test an idea, prediction or hypothesis. In that sense, the terms stand in contrast to theory and theoretical, the other routes to knowledge.

Experiment continues to be used in this sense throughout all the sciences today. Starting sometime in the 19th century, however, it was given an alternative, much narrower, and more precise meaning by large sections of the scientific community, that of the controlled, comparative or manipulative experiment. This established a taxonomy for empirical studies that classified them into two types – experimental investigations and observational investigations. This terminology is commonly employed by statisticians. Most scientists, however, especially those who make little or no use of manipulative experiments, do not accept observational as an appropriate descriptor of the often complex, sophisticated investigations they carry out. For that reason and also because observational in this sense embraces such a tremendous variety of types of empirical investigations – from counting daisies in a field to working out the structure of DNA to determining the atmospheric composition of a planet – the term observational itself conveys little information.

Consequently scientists in general have continued to use experimental in its ancient sense as a synonym for empirical. At the same time, in particular fields – such as agriculture, medicine, physiology, ecology, psychology, industrial technology – experimental has tended to be used, over the last hundred or so years, often in its narrower sense of pertaining only to manipulative experiments. So, not surprisingly, there is a great deal of confusion in writings on the history, philosophy and methodologies of science where the origin, nature, and value of "experimental" science are discussed. Truesdell (1987) is the only work I am aware of that clearly

describes the origin of these two different senses of experiment. It is apparent from the literature and other scientific discourse, that very few scientists and statisticians are aware of this history and these distinctions.

This history is relevant only because Oksanen's (2001) critique reflects the long-standing confusion over the distinction between empirical studies in general and manipulative experiments in particular. In his title and the early part of his paper, he seems to be using experiment only in the sense of manipulative experiment; this is appropriate as the latter were the entire focus of Hurlbert (1984). At one point he even distinguishes "experimental" and "observational" as two types of "empirical study." But gradually he shifts to using experiment in the sense of any empirical study carried out in accordance "with the basic principles of hypothetical-deductive science." In one place he claims "neither replication nor control are necessary parts of a critical experiment," and in another he refers to an astrophysical observation as a "spontaneous experimental situation."

Ecology is one field where at least over the last half century there has been an increasing tendency to apply the terms experiment and experimental only to manipulative experiments. Attempting to accelerate clarification of terms, I suggested that, given the long history of labeling as experimental many types of observational studies, one compromise might be to label such, especially the more complex of them, as mensurative experiments (Hurlbert 1984). Though this term is now used by many, the effect may not always have been positive. In the past, virtually all textbooks on experimental design have focused exclusively, or almost so, on the design of manipulative experiments. But three new books on experiments and experimental design by ecologists have reverted to using experiment in its more ancient sense. This implies a much broader range of topics than is usually covered in an experimental design text, though none of these new books follow through on the implicit promise. Scheiner (1993) defines experiment "as any test of a prediction." Underwood (1997):16) considers the distinction between manipulative and mensurative experiments a "distraction" and covers both. Quinn and Keough (2002):157) indicate that their "emphasis is on manipulative experiments," but in fact most of the examples they present are from observational studies (Hurlbert and Lombardi 2003). Thus on this matter the discipline of ecology may be headed back into semantic fog.

Surprisingly most books on statistics or experimental design, including many of the classics, offer no attempt to define the manipulative experiment. Of the few that do, their efforts seem inadequate. So here is an attempt to fill this vacuum:

> A manipulative experiment is an exercise designed to determine the effects of one or more experimenter-manipulated variables (= experimental variables or treatment factors) on one or more characteristics (= response variables) of some particular type of system (= the experimental unit). Its primary defining features are: (1) that the experimenter can assign treatments or levels of each experimental variable at random to the available experimental units; and (2) that there are two or more levels established for each experimental variable used.

Importance of treatment interspersion

A matter emphasized in the 1984 paper was the importance of spatial (or other sorts of) interspersion of treatments when there are replicate units under each treatment. At one point it states, "Perhaps experimental ecologists fall primarily into two groups: those who do not see *the need for any interspersion* [emphasis not in original], and those who do recognize its importance and take whatever measures are necessary to achieve a good dose of it." Oksanen (2001) does not indicate whether he agrees on the importance of such interspersion. But he creates some confusion, first by misquoting the statement (replacing the italicized portion above with "any need for dispersion") and then interpreting it as a statement concerning replication per se. Oksanen believes the statement unfairly implies that lack of treatment replication is often "a consequence of ignorance." The statement does not imply that, but Hurlbert (1984) taken as a whole does, and the implication is a fair one!

Valid tests for treatment effects in absence of treatment replication: special cases

There is a simple error in Hurlbert (1984) to which it is time to confess. The first line of that paper's abstract states that "Pseudoreplication is defined as the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated or....". Moral: don't rush the writing of abstracts! Oksanen (2001) approvingly reiterates the essence of that statement in saying, "If an experiment is not replicated [i.e. lacks replication of treatments], there is no possibility to strictly establish a connection between the treatment and the apparent effect."

There are, however, a few situations where these claims do not hold, and the situations are treated in most advanced statistics texts. One would be where the treatment factor is a continuous variable (e.g. fertilizer application rate) and the response variable (e.g. crop yield) is measured on only one experimental unit at each of several treatment levels. One can fit a regression model, e.g. a straight line, to such a data-set and estimate the appropriate error mean square from the deviations of observed values from model-predicted values. This mean square can then be used to test whether the slope of the regression line differs from zero. If the true functional relation between the treatment factor and the response

variable is not well described by the model used, then the mean square obtained will tend to overestimate the true error mean square, i.e. the one that could have been estimated directly had there been multiple experimental units under each treatment. Such an overestimate will reduce the power of the test and make the test for the slope of the line conservative. So if a low P is obtained in this test, one has grounds for concluding there was a treatment effect.

A second situation would be a factorial experiment where each treatment combination is applied to only a single experimental unit. Results from such an experiment could be analyzed with an ANOVA that uses the interaction mean square as an estimate of the true error mean square. If there is no interaction of the treatment factors, the interaction mean square is an unbiased estimate of the true error mean square and its use to test for treatment effects is valid. If there is factor interaction, this use of the interaction mean square will render such tests conservative, i.e. of low power, so that, again, if low P-values are obtained they constitute strong evidence for treatment effects.

Thus lack of treatment replication not only does not constitute pseudoreplication, it also does not necessarily preclude valid tests for treatment effects.

Overvaluation of experiments lacking treatment replication

Understanding of ecological and other natural phenomena that take place at large spatial and temporal scales can rarely be obtained by means of manipulative experiments. Creation or use of appropriately scaled experimental units is simply not feasible. In this regard, 'large scale' ecology is more similar to fields such as astronomy, geology, oceanography, epidemiology, and sociology than it is to fields such as medicine, agriculture, cell biology and industrial processes.

Nevertheless, it is occasionally possible for ecologists to set up manipulative experiments with a spatial extent (whole lakes, islands, small watersheds, large forest patches) much greater than that of the conventional agricultural plot, the archetype experimental unit for field biologists and statisticians. Often these large-scale manipulative experiments may lack replication of treatments. Nevertheless, partly because such studies are rare and especially when they have involved manipulations with dramatic apparent effects, some such experiments have led to significant new insights, corroborated particular theories, and advanced science. All this is acknowledged in Hurlbert (1984), where, contrary to Oksanen's (2001) implications, such studies (Likens et al. 1970, Schindler et al. 1971) were not referred to as "unrigorous" or "pseudoreplicated."

Aside from its critique of Hurlbert (1984) and of "inductionism", the main thrust of Oksanen (2001) is a strong defense of experiments lacking treatment replication, the importance of applying "inferential statistics" to them, their power to test theory, and the ability of meta-analysis to compensate for their deficiencies. Oksanen draws a hard line according to whether a "study is conducted in a deductive or inductive spirit." He claims in his abstract that

> If the experiment is based on deductive logic, the rules of the game are entirely different... and replication is not an essential part of the experimental design... The scope of a deductive experiment is... to allow the experimentalist to check 'yes' or 'no' boxes in a pre-existing test protocol.... For a strict advocate of the hypothetico-deductive method, replication is unnecessary even as a matter of principle, unless the predicted response is so weak that random background noise is a plausible excuse for a discrepancy between prediction and results.... Hence choosing two systems and assigning them randomly to a treatment and a control is normally an adequate design for a deductive experiment.... replication can always be obtained afterwards... by using meta-analysis.

There is much subjectivity to evaluation of these matters, but let me offer the following counterpoints.

First, though he says "normally," his recommendations actually seem intended for a very narrow class of situations, those where it is known beforehand that treatment effect will be so much greater than "background variation" that treatment replication can be dispensed with. Some might take this as advice to select a magnitude or level of the treatment factor that will function as a sledgehammer even if the ecological question or hypothesis would seem to call for a tack hammer. So his advice is not applicable to experimentation in general of either the "deductive" or "inductive" variety.

Second, though he insists on application of "inferential statistics" (obtaining estimates of "experimental error" from within-experimental unit variation) to results from unreplicated treatments, his 'box checking' protocol would seem to require only determination of whether the difference between the two sample means was positive or negative. He misreads Hurlbert (1984) in stating that "to require that inferential statistics should not be used in the context of unreplicated experiments is plain nonsense." My recommendation to editors suggested they "[dis]allow the use of inferential statistics where they are being misapplied." This can hardly be considered controversial advice. If an investigator gets a low P-value in a t-test applied to the results of a two-treatment-no-replication experiment and claims that the low P-value constitutes statistical evidence against the null hypothesis of no treatment effect, then that is a clear misapplication of inferential statistics. In the literature it has been rare that those who have carried out such experiments and tests have refrained from interpreting

their low P-values as definitive evidence of treatment effects.

Third, Oksanen's arguments do not recognize that if such an experiment and test are carried out and if the null hypothesis of no treatment effect is true, then the probability of making a type I error will approach 100 percent and the probability of 'confirming' the substantive hypothesis or prediction will approach 50 percent, as the number of measurements made in each experimental unit becomes large. This is because two experimental units are, in reality, always different, and a test, with large sample sizes, of the null hypothesis that they are not is thus practically guaranteed to yield a low P-value. And because if treatments are assigned randomly, the observed difference in the response variable will have a 50 percent chance of being in the direction predicted by the hypothesis or theory being tested. Therefore, when such an experiment "confirms" such a prediction and bolsters a theory, it represents the weakest, least rigorous sort of confirmation imaginable. On the other hand, if a proper test of the same prediction is carried out, the probability of making a type I error that appeared to "confirm" our prediction would be only 0.5 times alpha, again on the assumption the null hypothesis were true.

Finally, one must question the notion that "our collective rate of progress" in large scale ecology will be maximized by allocating resources to large numbers of experiments lacking treatment replication and relying on meta-analysis rather than allocating the same resources to a smaller number of more expensive experiments with modestly replicated treatments. Meta-analysis is far from a methodological panacea that can compensate for the weaknesses of studies fed into it. When for lack of treatment replication, estimates of effect size contain large amounts of 'noise' or random error, the output of a meta-analysis will also be 'noisy.' More than usually would usually be the case, meta-analysis will be unlikely to lead to any greater understanding than provided by simpler, less pretentious, more direct reviews of published studies. Many subjective decisions are involved in the conduct of meta-analyses; we should not be deceived by the statistical apparatus involved into thinking them a powerful, objective and rigorous tool. Much of their quantitative output is artifactual and tells us more about experimenters and meta-analysts than it does about nature. Meta-analyses can function as convenient and condensed summaries of what is already known from the best well-designed studies, but at least in ecology I am not aware of any meta-analysis that has provided significant new insights into the literature or natural phenomena it describes. It must be doubted that one conducted primarily for weakly designed experiments would break the trend.

Another resource allocation consideration also argues against a relative increase in support for experiments with unreplicated treatments. When costs of setting up and maintaining an experiment are very high, it is an unaffordable economic luxury to restrict ourselves to a rigid hypothetico-deductive framework and measure only one or a few response variables about which our theory makes firm predictions. We will maximize the value of the experiment by monitoring large numbers of other variables. This will be possible at a relatively small incremental cost to the project. Some of these other variables may serve only to define the conditions of the experiment, others may provide insight into the mechanisms by which the treatment factor exerts its effects, and others may provide insights into new phenomena or relations marginally related to the phenomena and theories of prime interest. But if treatments are not replicated then our information on these other variables will be inconclusive indeed. Not that this will prevent some fancy story-telling.

In the last analysis, every proposed experiment must be judged by its own objectives, design, possibilities, and costs. There should be no automatic rejection of experiments where no treatment replication is proposed. Nor should there be automatic rejection of more powerful experiments having treatment replication simply on the basis of their costs.

But let us not fear to call a spade a spade. Pseudoreplication continues to be one of the most common statistical errors in ecology and many other social and natural sciences. Its commission has nothing to do with the distinction between deductive and inductive modes of reasoning. Scientists who are familiar with the most common varieties of pseudoreplication – simple, temporal, and sacrificial – will find it easy to avoid them. Editors and referees who are not familiar with them will continue to misdiagnose manuscripts and foster confusion in the journals. Viva stigmatization!

Acknowledgements – I thank Marie Coe and Emili García-Berthou who first called my attention to Oksanen (2001) and demanded that I rebut it; and to them, Celia Lombardi, and Heikki Hirvonen for reviewing and critiquing the manuscript that eventually emerged.

References

Cottenie, K. and De Meester, L. 2003. Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). – Oikos 100: 394–396.

Ford, E. D. 2000. Scientific method for ecological research. – Cambridge Univ. Press

García-Berthou, E. and Hurlbert, S. H. 1999. Pseudoreplication in hermit crab shell selection experiments: comment to Wilber. – Bull. Mar. Sci. 65: 893–895.

Hairston, N. G., Sr. 1989. Ecological experiments: purpose, design, and execution. – Cambridge Univ. Press.

Heffner, R. A., Butler, M. J. and Reilly, C. K. 1996. Pseudoreplication revisited. – Ecology 77: 2558–2562.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. – Ecol. Monogr. 54: 187–211.

Hurlbert, S. H. 1990. Pastor binocularis: now we have no excuse [review of Design of Experiments by R. Mead]. – Ecology 71: 1222–1228.
Hurlbert, S. H. and White, M. D. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. – Bull. Mar. Sci. 53: 128–153.
Hurlbert, S. H. 1997. Experiments in ecology [Review of book by same title by A.J. Underwood]. – Endeavour 21: 172–173.
Hurlbert, S. H. and Lombardi, C. M. 2003. Design and analysis: uncertain intent, uncertain result [review of Experimental design and data analysis for biologists, by G. P. Quinn and M. J. Keough], Ecology (in press).
Hurlbert, S. H. and Meikle, W. G. 2003. Pseudoreplication, fungi, and locusts. – J. Econ. Ent. 96: 533–535.
Jenkins, S. H. 2002. Data pooling and type I errors: a comment on Leger & Didrichsons. – Anim. Behav. 63: F9–F11.
Krebs, C. J. 1989. Ecological methodology. – Harper and Row.
Kroodsma, D. E. 1989a. Suggested experimental designs for song playbacks. – Anim. Behav. 37: 600–609.
Kroodsma, D. E. 1989b. Inappropriate experimental designs impede progress in bioacoustic research: a reply. – Anim. Behav. 38: 717–719.
Kroodsma, D. E. 1990. How the mismatch between the experimental design and the intended hypothesis limits confidence in knowledge, as illustrated by an example from bird-song dialects. – In: Bekoff, M. and Jamieson, D. (eds), Interpretation and explanation in the study of animal behavior. Vol. II. Westview Press, pp. 226–245.
Likens, G. E., Bormann, F. H., Johnson, N. M. et al. 1970. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed ecosystem. – Ecol. Monogr. 40: 23–47.
Lombardi, C. M. and Hurlbert, S. H. 1996. Sunfish cognition and pseudoreplication. – Anim. Behav. 52: 419–422.
Machlis, L., Dodd, P. W. D. and Fentress, J. C. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. – Z. Tierpsychol. 68: 201–214.
Mead, R. R. 1988. The design of experiments. – Cambridge Univ. Press.
Oksanen, L. 2001. Logic of experiments in ecology: is pseudoreplication a pseudoissue? – Oikos 94: 27–38.
Quinn, G. P. and Keough., M. J. 2002. Experimental design and data analysis for biologists. – Cambridge Univ. Press.
Scheiner, S. M. 1993. Introduction: theories, hypotheses, and statistics. – In: Scheiner, S. M. and Gurevitch, J. (eds), Design and analysis of ecological experiments. Chapman & Hall, pp. 1–13.
Schindler, D. W., Armstrong, F. A. J., Holmgren, S. K. et al. 1971. Eutrophication of lake 227, Experimental Lakes Area, northwestern Ontario, by addition of phosphate and nitrate. – J. Fish. Res. Bd Can. 28: 1763–1782.
Sokal, R. R. and Rohlf, F. J. 1969, 1981,1995. Biometry. 1st, 2d and 3d ed. – W. H. Freeman.
Truesdell, C. 1987. Great scientists of old as heretics in 'the scientific method'. – Univ. Press of Virginia, 96 pp.
Underwood, A. J. 1997. Experiments in ecology. – Cambridge Univ. Press.
Underwood, A. J. 1998. Design, implementation, and analysis of ecological and environmental experiments. – In: Resetarits, W. and Bernardo, J. (eds), Experimental ecology, issues and perspectives. Oxford Univ. Press, pp. 325–349.