Journal of Comparative Psychology
2009, Vol. 123, No. 4, 434–443

© 2009 American Psychological Association
0735-7036/09/$12.00 DOI: 10.1037/a0016221

COMMENTARIES

The Ancient Black Art and Transdisciplinary Extent of Pseudoreplication

Stuart H. Hurlbert
San Diego State University

The history, definitions, and transdisciplinary extent of pseudoreplication, as well as some key concepts of experimental design, are briefly reviewed. Pseudoreplication, sometimes also referred to as the 'unit of analysis error,' is one of the commonest errors of statistical analysis and interpretation. It is a simple albeit serious one. It persists in part because of the failure of statisticians and scientists to develop a clear, consistent terminology for the concepts of statistics, experimental design, and sampling design that is used across all disciplines, as well as a terminology for specific categories of the more common errors. Lack of a clear terminology, in turn, has fostered narrow, discipline-specific jargon, inconsistency among textbooks and reference works, and ineffective teaching. Reform of terminology is possible, and great improvement in statistical practice would follow.

Keywords: unit of analysis error, chi-square tests, pooling, experimental design, evaluation unit, experimental unit

"Dr. Box chuckled when I read him the letter – he has similar experience to yours, but in chemical experiments, and agrees that treating subsamples from a single experimental unit as if each represented an independent experimental unit is one of the commonest errors of analysis."

—Joan Fisher Box, in litt. to S. Hurlbert, 27 November 1981

Entreaty of a Youthful Offender

Said the student to professor:
Will you be my true confessor?

I have sinned a bit of late,
And I have to know my fate.

Last week sans *contemplation*
I did faulty replication;

All my data have been lumped –
Must my chi-squared tests be dumped?

Is my study superficial?
Was I blindly sacrificial?

What's the verdict – I can't wait –
Did I pseudoreplicate?

[*] *cf* Hurlbert (1981)

About those others who all did it –
Oh how cleverly they hid it!

Though you surely can expose them
It will not soon now depose them,

For they've got their modest fame.
Fending you off's just a game.

What of me? I'm a beginner,
You could help me be a winner.

I'm in such a great big hurry –
Don't have time for all this worry.

So please don't depilate[*] *it*
If I've pseudoreplicated.

– Joy Zedler, 1985

Stuart H. Hurlbert, Department of Biology, San Diego State University.

I dedicate this article to Celia Lombardi who invited me in the 1980s to collaborate in an examination of statistical problems in the literature of behavior and psychology and who has been a most stimulating colleague and pseudoreplication psleuth ever since.

This article has greatly benefited from suggestions by Emili García-Berthou, Mikhail Kozlov, Celia M. Lombardi, Kathryn L. Mier, Peter Petraitis, Susan J. Picquelle, Michael Riggs, N. Scott Urquhart, and Joy Zedler. Jeffrey C. Schank, Thomas J. Koehnle, and Ed. Gordon Burghardt are thanked for goading me into the enlightening task of reading dozens of books and articles that otherwise might have gone unread.

Correspondence concerning this article should be addressed to Stuart H. Hurlbert, Department of Biology, San Diego State University, San Diego CA 92182. E-mail: shurlbert@sunstroke.sdsu.edu

This paper represents an invited commentary on Schank & Koehnle (2009; hereinafter referred to as *SK*). The matters at hand are serious ones involving the implication by SK that my papers on pseudoreplication constitute a sort of statistician's version of *The Satanic Verses* (Rushdie, 1989), full of error, deception, and heretical "doctrine"; their injection of further terminological confusion into an arena already overloaded with it; and some misstatements of fact. But let me start with these chuckles from Wisconsin.

Joan Fisher Box, R.A. Fisher's second oldest daughter, included in her fine biography of her father (Box, 1978) an account of how Fisher had committed what I was calling pseudoreplication in his statistical analysis of a complex experiment on potatoes and fer-

tilizers (Cochran, 1980; Box, 1978, pp. 110–112, Hurlbert, 1984) but had removed that error, without comment, when he presented that experiment as an example of ANOVA in his *Statistical Methods for Research Workers* (Fisher, 1925). I enquired of Mrs. Box whether she had any further information on the history of her father's recognition of the error. She provided some quotes from letters to Fisher in 1923 from his wise mentor, William Sealy Gosset, suggesting that it was Gosset who gave Fisher some second thoughts on the matter. As I explained to Mrs. Box, my interest in summarizing this misstep by Fisher was, in part, to "soften the blow for these modern pseudoreplicators, whose works are listed and discussed in my paper; since they include several journal editors and at least three recent presidents of the *Ecological Society of America*, such softening may greatly increase the likelihood that my review will eventually be accepted for publication by an ecological journal" (S. Hurlbert *in litt.* to J. F. Box, 9 November 1981).

That apparently was what got the chuckle out of "Dr. [George E. P.] Box," who was Mrs. Box's husband, author of *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building* (Box, 1978) and more than a hundred other statistical publications, and the R. A. Fisher Professor of Statistics at the University of Wisconsin in Madison. "Dr. Box" perhaps would have thought the time of Mssrs Schank and Koehnle better invested in analysis of pseudoreplication in reports of industrial chemical experiments, where Box himself had earlier worked, than in an attempt to suppress a convenient label for a specific class of statistical error.

The poet, Joy Zedler, also is renowned as a wetland ecologist and is currently the Aldo Leopold Professor of Restoration Ecology at the University of Wisconsin. She composed the poem, however, when she was my colleague at San Diego State University, in the cauldron where we were instilling in our graduate students the discernment and blood lust required for competent pseudoreplication psleuthing. SK have much to say about the matter of "lumping" or "pooling" in analyses of continuous variables and its attempted justification on the basis of *F* tests. Joy's poem refers to the pooling of replicate tables of categorical data prior to a χ^2 or G-test for a treatment effect, where justification for pooling is sometimes attempted on the basis of a preliminary χ^2 or G- test for "heterogeneity." In either circumstance, as discussed later, sacrificial pseudoreplication is usually the result of such pooling. Joy, I should mention, has always been fastidious about her χ^2 tests, and nothing autobiographical is to be inferred from the poem. Not that she was a beginner even in 1985.

Preliminary Clearing of the Deck

SK contains many errors, erroneous attributions, and confusing statements. Space limitations preclude detailed rebuttal of them all, so let me simply state some corrections telegraphically in list form: pseudoreplication is not a "doctrine," nor a synonym or neologism for "experimental confounds" or "failure of statistical independence"; 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*); I have never argued "that locations close in space and time are by their very nature statistically dependent"; I have never stated or implied that "study [of] a particular lake or watershed..[can provide no] knowledge about lakes or watersheds in general . . ."; I have never stated or implied that "temporal pseudoreplication . . . occurs [*automatically,* is implied] when repeated measures are taken on an experimental unit over time;" in every discipline there *are* "satisfactory criteria for drawing boundaries around experimental units"; in no type of experimentation is it appropriate to "control physical conditions . . . as much as possible"; 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 [*non*-]essential for well-controlled experiments," as in the case of ecological field experiments; and if there is high potential for "interactions or contamination among units," then the experimental unit has been inappropriately defined or constructed, and results will be compromised and resources wasted. It should be easy to locate in SK where the corresponding problematic statements are found.

The rest of this paper is intended to be of value to a broad audience with minimal interest in fine debating points. It puts forward a coherent conceptual and terminological framework for discussing not only pseudoreplication, but the basics of experimental design as well. It recounts some classic cases of pseudoreplication in early education, agricultural, and genetic experiments that still have lessons for us. It discusses problems that arise when clear, universally accepted labels are lacking, both for specific types of statistical errors, such as pseudoreplication, and for specific concepts in experimental design. Finally, it documents the poor advice on pooling in textbooks responsible for much test-qualified sacrificial pseudoreplication. And it points out how multilevel modeling, valuable as it is, cannot prevent the commission of pseudoreplication by persons who do not yet understand *t* tests, χ^2 tests, and the basics of experimental design. My comments on further aspects of SK are embedded in these discussions.

A Conceptual and Terminological Framework for Experimental Design

SK ignore much of the conventional terminology of experimental design. This is not entirely their fault as clear definitions for key concepts are rare in books on experimental design or statistics and the jargon varies from discipline to discipline. On the other hand, I reviewed earlier versions of SK in 2001 and 2004 and recommended on each occasion that they consult some of the more cogent introductions to experimental design, mentioning Cox (1958); Steel and Torrie (1980), and Mead (1988). These works—or ones of comparable clarity—remain unreferenced in SK.

Before we proceed to a discussion of pseudoreplication, a clear terminology needs to be established covering the key concepts of experimental design. In the absence of a common language, useful discussion is difficult. The specific definitions below are taken from papers by myself and my colleagues, but in all cases represent slight reformulation or relabeling of ancient concepts most recently codified by authors such as Cox (1958); Kempthorne (1979); Urquhart (1981), and Milliken and Johnson (1984). They are applicable to all fields in the natural, social, and behavioral sciences and in engineering, where manipulative (or controlled or comparative or randomized) experiments are conducted. Pseudoreplication can also be committed in observational studies,

where alternative terminologies would apply but that topic will not be treated here.

So here are some definitions proposed for universal adoption. I note places where SK seem not to accept these definitions.

Experiment: "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." (Hurlbert, 2004).

In a radical departure from conventional terminology, SK label as a "controlled experiment" a sampling exercise "to determine which of two urns contain the greater proportion of red to blue marbles." While it is an ancient custom across most of the sciences to use "experiment" and "experimental" to refer to *any* set of empirical observations undertaken to test a hypothesis or answer a question (Truesdell, 1987; Hurlbert, 2004; Lombardi, 2007), confusion is introduced when authors' meanings jump back and forth between this ancient, broader sense of "experiment" and the more precise, modern, statistical definition.

Experimental unit: "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).

In their urn example, SK refer to the urns as "experimental units," but the term and concept of *experimental unit* have no application outside manipulative experiments. The same confusion is exhibited in their statement that "there are no satisfactory criteria for drawing boundaries around experimental units (see Burstein, 1980, for an excellent overview of this problem . . .)." Burstein's is indeed a fine review article, but its coverage was "restricted to multilevel issues in large-scale, *nonexperimental* [italics supplied] educational research and evaluation" (Burstein, 1980, p. 159), and so not surprisingly the term *experimental unit* is nowhere to be found in it.

The long-recognized, most fundamental criterion for the bounding of experimental units is that this be done in such a way that what transpires on or in one experimental unit will have no effect on other experimental units. Student (1923, p. 278) gave an early example of the conflict between the desirability of having "maximum contiguity" of experimental units (in what would later be called a block) and the possibility of biased treatment effects arising from shading of a short variety of barley by a taller one in an adjacent plot. Cox (1958, pp. 19–21) treated the topic nicely in a three-page section titled "Interference between different units." Wiley (2003) gives an excellent synopsis of the principle in his review of design and analysis issues in bird behavior studies. Mead (1988) throughout his text emphasizes that "to achieve good experimental design, the experimenter should think first about the experimental units" (p. 7) and gives an example (p. 120) of how a biased estimate of hormone effect could result if pigs receiving different hormone treatments are penned together. SK imply that Heffner, Butler, and Reilly (1996) did not understand this principle in advocating "separate rooms" (e.g., units with independent air circulation systems) as the kind of experimental units needed for a study involving mouse pheromones; yet, apart from possible ambiguity injected by their phrase "far apart," Heffner et al. (1996) got it exactly right.

Standing as a marked counterexample to Burstein's care with language is a review by Koch, Amara, Stokes, and Gillings (1980) published in the same year. Those authors say they are concerned only with "split-plot experiments" and "repeated measurements experiments," yet of the 12 hypothetical cases they construct for discussion, four are actually observational or sampling studies, not manipulative experiments; and in each of those four cases, Koch et al. explicitly synonymize "primary sampling unit" with "experimental unit." SK are not the originators of such careless terminology which has long been widespread, but they are perpetuating it.

SK present some "simulated contamination experiments," the main point of which seems to be to show the obvious—that problems arise when events on one experimental unit influence events on other experimental units. Though they present only systematic or completely randomized layouts and no randomized block designs, SK call the experimental unit a "block." They thus end up with confusing text that refers to "block effects," "control blocks," "treatment blocks," and "experimental blocks." For the better part of a century, *blocking* has been defined as the grouping of experimental units into sets, all units in a set being as alike as possible in some key respect(s), and assigning treatments in such a way that the replicates of each treatment are distributed as evenly as possible among the sets, with a *block* being defined as a set so created (e.g., Fisher, 1935; Cox, 1958; Kirk, 1982; Mead, 1988).

Evaluation unit: "The unit of research material on which a response is evaluated" (Urquhart, 1981), or "that element of an experimental unit on which an individual measurement [of a response variable] is made" (Hurlbert, 1990; Hurlbert & White, 1993).

The critical distinction between the *experimental unit* and *evaluation unit* and its implications for statistical analysis are recognized even in older literature (e.g., Neyman & Pearson, 1938; Lindquist, 1940), and Urquhart's (1981) distinct label of *evaluation unit* greatly assists definitions of pseudoreplication (e.g., Hurlbert & White, 1993; Lombardi & Hurlbert, 1996; Hurlbert & Meikle, 2003; Hurlbert, 2004; Hurlbert & Lombardi, 2004; Kozlov & Hurlbert, 2006) as indicated below. And it also nicely sets the stage for the long overdue project of banishing of the overused term *subject* from the design and statistical analysis literature. More on that later. Though *observational unit* and, sometimes, *sampling unit*, have long been used for what Urquhart (1981) calls *evaluation unit*, the former terms seem less desirable in experimental contexts as they potentially foster further confusion over the distinction between experimental and observational or purely sampling studies (Kozlov & Hurlbert, 2006).

Experimental design: "the logical structure" of a manipulative experiment (Fisher, 1935).

"An experimental design has four aspects: (1) treatment structure, (2) treatment replication, (3) design structure, and (4) re-

sponse structure . . . **Treatment structure** is the set of experimental treatments or treatment combinations used and how they relate to each other . . . **Treatment replication** refers to the number of experimental units that will be subjected to a treatment . . . **Design structure** refers to the manner in which treatments or treatment combinations are assigned to experimental units . . . There are three basic design structures . . . a completely randomized design, . . . a randomized block design, [and] . . . a split-unit design . . . The **response structure** consists of the list response variables to be measured and the sampling plan that specifies when, where, and on what components of the experimental unit one will make and record their observations and measurements [that is, measure those response variables]" (Hurlbert & Lombardi, 2004, after Urquhart, 1981).

To date only a few texts (e.g., Milliken & Johnson, 1984; Valiela, 2001) have adopted this conceptual framework, or some variation of it, for experimental design. It eventually will find more widespread adoption as it helps resolve conceptual and terminological problems that have long plagued many disciplines.

Pseudoreplication Is a Real Problem and a Useful Label

More Precise Definitions

SK in their abstract refer to "the growing criticism of the concept of pseudoreplication," but I am aware of only a single paper (Oksanen, 2001) that is as critical and disdainful of the term as are SK. The misunderstandings in Oksanen (2001) were corrected by Cottenie and DeMeester (2003) and Hurlbert (2004), and further assuaged in private correspondence. There now is controversy over the issue in Russia (Tatarnikov, 2005; Rosenberg & Gelashvili, 2008) provoked in part by Kozlov (2003) and Kozlov and Hurlbert (2006) and exacerbated by a lack of good Russian-English terminological concordances. The latter derives from structural differences in the languages as well as from political suppression in Russia of theoretical and applied statistics as "bourgeois" from the 1930s into the 1950s (Kotz, 1965), just the period when the principles of modern experimental design were being refined.

Misunderstanding of pseudoreplication by SK can be laid partly on my shoulders because my first article (Hurlbert, 1984) on the topic indeed offered *characterizations* of the nature of the problem more than it did precise *definitions*. That in turn led to many inaccurate characterizations and definitions of pseudoreplication in works of other authors. My description of pseudoreplication in terms of *replicates*, though consistent with much of the prior and current statistical literature, was also unfortunate. It led to others promoting or coining misleading terms such as *true replicates*, *false replicates*, and *pseudoreplicates*. I have tried to counter such language by strongly recommending that *replicate* be used *only* as a qualifier, as in replicate experiments, replicate experimental units, replicate blocks, replicate evaluation units, replicate samples, replicate subsamples, and so on (Hurlbert, 1990).

Nevertheless, the clarifications and numerous additional examples in my and my colleagues' post-1984 papers on the topic should have removed all confusion about the meaning of pseudoreplication by now. SK cite but seem not to have read Hurlbert and White (1993) wherein the first three of the following five definitions were offered:

Simple pseudoreplication. "There is a single experimental unit per treatment, but multiple measurements on each experimental unit . . . These multiple measurements are then treated statistically as if each represented a separate experimental unit."

Temporal pseudoreplication. " . . . multiple measurements on an experimental unit are taken successively in time [and] . . . are treated as if each represented a different experimental unit."

Sacrificial pseudoreplication. "When the number of experimental units (n) per treatment is 2 or more and . . . the number of evaluation units (k) measured per experimental unit is 2 or more . . . [and] the analyses ignore the structure in the set of nk measurements per treatment and treat each measurement as if it represented an independent replicate of the treatment."

Test-qualified sacrificial pseudoreplication. Sacrificial pseudoreplication committed on the ground that the multiple evaluation units within an experimental unit are supposedly validly treated as experimental units "when tests for differences among experimental units (within treatments) yield high p values." (Hurlbert, 2004).

In experimental contexts then, **pseudoreplication** "represents confusion between the experimental unit and the evaluation unit" (Hurlbert & White, 1993) and is broadly defined as "a serious type of statistical error that . . . occurs when measurements made on multiple evaluation units, or multiple times on a single evaluation unit, in each experimental unit are treated statistically as if each represented an independent experimental unit . . . The usual [but not universal] consequence of pseudoreplication is exaggeration of both the strength of the evidence for a true difference between treatments and of the precision with which any difference that does exist has been estimated" (Hurlbert & Lombardi, 2004). It is simply an error of statistical analysis and interpretation and is not merely a weak design or an inevitable consequence of such (Hurlbert, 1984, 2004; Hurlbert & White, 1993; Hurlbert & Lombardi, 2003, 2004; Hurlbert & Meikle, 2003; Kozlov & Hurlbert, 2006).

An Ancient Tradition

As the mathematical apparatus and recipes for modern statistical tests developed well ahead of clear, consistent conceptual and terminological frameworks for facilitating their proper application and interpretation, it is not surprising that both scientists and statisticians have long made errors in applying these tests. Pseudoreplication, a rather simple type of error, has been committed, recognized, and warned against for a long time. Brief review of some early and little known cases may be of interest.

At the beginning of the 20th century, agronomists conceived the reasonable idea of taking a limited number of samples (= evaluation units) from the large plots used as experimental units in grain experiments instead of harvesting the entire plots. This could save time with possibly only small losses in precision. One team of researchers in Minnesota, Arny and Steinmetz (1919), carried out several experiments using 1/10 acre plots and measuring yields for the entire plots as well as for 1 square yard subplots within them. When all yields were expressed on a per acre basis, they found, not surprisingly, higher variability among square yard-based estimates than among 1/10 acre-based estimates. They also determined that if field conditions were not too heterogeneous and if, in essence, the square yard subplots were treated as the experimental unit, then the standard error of a treatment based on $\sim 5n$ samples of one square yard (obtained by sampling 5 square yards in each of the n

experimental units) was about the same as the standard error obtained using only the *n* values for the total yields of 1/10 acre plots. They concluded that, "Under similar circumstances the yields from a greater number of square-yard areas may be considered more accurate than those from the entire [1/10 acre] plots." This was archetypical sacrificial pseudoreplication. Reflecting the spurious increased power that pseudoreplication usually confers, statistically significant differences among treatments were found more often with 5*n* to 10*n* measurements per treatment for square yard plots than with the *n* measurements for total yields of 1/10 acre plots. It was W.S. Gosset (Student, 1923, p. 286) who first caught this error, dryly remarking, "It is rather surprising that they did not realize that there are 484 square yards in 1/10th acre, so that by taking 484 yards they would be likely to be more accurate than if they took any lesser number . . . " When Fisher (Barbacki & Fisher, 1936) committed the same type of error several years later, it was once again Gosset (Student, 1938, p. 368) who had to point it out. There are grounds for considering Gosset the Original Pseudoreplication Psleuth.

Pseudoreplication was the norm in early experimental research in education and remains very common in that field (McCall, 1923; Confrey & Stohl, 2004). Pittman (1921) wished to compare two supervisory systems on student performance. He applied each system to all the schools and students in a single county, the schools and pupils in schools representing two levels of nesting in the response structure for the single experimental unit (= county) under each treatment. Assessment of the effect of supervisory system was conducted with a primitive sort of *t* test that treated the individual student as the experimental unit. That analysis formally qualified as simple pseudoreplication, but Pittman did two things to minimize bias in his estimate of a treatment effect: he selected the two counties specifically for their similarity on a variety of relevant sociological and economic measures, and, at the end of the experiment, he paired students (one from each treatment) on the basis of their test scores before estimating a treatment effect. As usual, the conclusions of such a study require a careful, but mostly subjective, assessment of other possible sources of, or contributors to, the apparent effect of supervisory system, but the study is not without value.

Stigler (1986, p. 244) suggested that portions of Fechner's (1860) *Elemente der Psychophysik* "constituted the most comprehensive treatment on [experimental design] before R.A. Fisher's, 1935 *Design of Experiments*," but a good case can be made that the honor should go to McCall's (1923) *How to Experiment in Education*. That book summarizes the concepts and methods developed mostly in the United States in the first two decades of the 20th century by education researchers. Though McCall's pre-Fisherian terminology will be mostly unfamiliar it parallels modern terminology. It recognizes the various forms an experimental unit may have ("experimental subject, thing, or group"), emphasizes that prior to application of a treatment, the experimental units should be "equivalent . . . have like means and like variability of subjects within them" (e.g., pupils in classrooms), acknowledges the possibility of achieving this by randomized assignment of students to classrooms but says "measurement . . . is the best basis" for doing this, that is, using measured characteristics such as grades, test scores, sex, and so forth to assure a similar spectrum of pupils in each classroom. He discusses cross-over design structure ("rotation experimental method") and the potential problem of carryover effects. He states that "normally" (1923, p. 18) there is only a single experimental unit under each treatment ("equivalent groups experimental method"). But he does not even acknowledge the possibility of completely randomized designs with replicate experimental units (e.g., classes) under each treatment even though he claims to treat in his book "all the common varieties of experiments" (p. 140). Prior to Fisher (1935) there was independent but parallel evolution of experimental design principles in educational and agricultural research. The notion of educational researchers that they could avoid bias or confounding merely by assuring the pre-experiment "equivalence" of individual experimental units had its parallel in the notion of early agricultural researchers and statisticians like W. S. Gosset (Student, 1923, 1938; Box, 1978; Hurlbert, 1984) that systematic design structures (or, more precisely, randomized block designs with systematic allocation of treatments to experimental units within blocks, which were Gosset's focus) can do the same by favoring pre-experiment "equivalence" of different *sets* of experimental units. Eventually the value both of replicate experimental units and of randomized block designs to their common objective became clear to all. Nevertheless, it is understandable that the perspectives of educational and agricultural researchers still differ. To include more field plots in an experiment is "child's play" compared to the headaches, hassle, and expense of including more classrooms in an experiment that involves various fractious elements of a hierarchical *Homo sapiens* social structure.

While Fisher's commission of pseudoreplication in Fisher and Mackenzie (1923) and Barbacki and Fisher (1936), was, as earlier discussed, recognized long ago, it has not been recognized that he also provided us with one of the earliest examples of simple pseudoreplication in the context of a regression analysis (Fisher, 1925, p. 214, 1958, p. 252). This concerned an experiment to assess the relation between rearing temperature and the number facets in the eyes of fruit flies. One rearing bottle (= experimental unit) of flies was kept at each of nine different temperatures (15–31 °C), and facet number was determined for 53 to 137 flies from each bottle. No statistical analyses were presented in the original work (Hersh, 1924), but Fisher analyzed the data, testing for significance of the negative slope observed (7 error d.f., $p < .001$)) and for deviations from linearity (814 error d.f., $p < .001$). The first test is fine and used the mean facet number calculated for the flies of each bottle. The second test represented simple pseudoreplication in that the individual flies were being treated as independent *experimental units* rather than as the *evaluation units* they were. The deviation from linearity of the best-fit *line* could have been due to either a bottle effect or a temperature effect, and a mathematician might not care which. But a biologist conducts the test for deviation from linearity in order to examine how much evidence there is for curvilinearity in the *true functional relation* between temperature and facet number. There are different ways to do that, the most obvious involving use of two or more experimental units at each temperature; but it is not validly done as Fisher did it.

Joseph Berkson (1942), an early critic of significance testing, plotted the fruit fly data, observed that the nine means seemed to be closely and roughly randomly distributed about the regression line and intuited something was wrong. He concluded that, Fisher's finding of a "significant" deviation notwithstanding, "it appears as straight a line as one can expect to find in biological

material. What has betrayed the author [Fisher] is a faithful adherence to an unsound principle: to wit, reject the null hypothesis tested.. if the *P* of the test is small." That principle is, of course, *not* unsound, but any test does involve certain assumptions, and Berkson correctly guessed that Fisher's suspicious result could have resulted from strong violation of one or more of them. Berkson speculated that heterogeneity of within-bottle variances or fluctuations in the independent variable, temperature, might have been involved. It did not occur to Berkson that the most obvious, and perhaps only, violation on Fisher's part was his assumption that the "bottle effect" was zero, that is, that measurements on individual flies possessed the statistical independence to be treated as if they represented individual experimental units.

His stiletto pen ever at the ready and his paternal affection for significance tests raising the hairs on the back of his neck, Fisher (1943) shot back:

> "It is not my purpose to make Dr. Berkson seem ridiculous, nor, of course, to prevent him from providing innocent amusement. Had he looked up Hersh's original paper he would have been spared a blunder, but we should have lost an example of the dangers of authoritarian judgment, based on subjective impressions . . . It is very well worthwhile to be reminded that general condemnations of 'biological material' based on limited experience, as Dr. Berkson's judgment must be, may vastly underestimate the cogency of the evidence which careful and extensive work neatly provides."

Examination of Hersh's paper would have been of no help to Berkson, and the "cogency of the evidence" provided by Fisher was simply less cogent than Berkson's "subjective impressions," albeit just as weak as Berkson's critique of significance tests. Forty-four years after Fisher's analysis of this fruit fly experiment, Sokal & Rohlf (1969, p. 438) presented an exactly parallel experiment and statistical analysis involving four tanks of fish each established with a different density of fish. The ANOVA they applied to their data represented simple pseudoreplication, as discussed elsewhere (Hurlbert, 2004; Kozlov & Hurlbert, 2006), but so did their test for deviations from linearity (error d.f. = 96) when they conducted a regression analysis on the same data. This fish experiment example was retained in their second edition but deleted from the third (Sokal & Rohlf, 1981, p. 488, 1995).

Value of Clear Labeling Versus Redundant, Imprecise Terms

Pseudoreplication can be found in the literature of any discipline from the moment that discipline began using statistical analyses, and discussion of the phenomenon and dissections of individual cases, such as those in the preceding section, have long abounded. For many, general awareness of the problem was given a sharp boost by reviews in the 1980s that focused on the problem in education (Barcikowski, 1981; Hopkins, 1982), medicine (Whiting-O'Keefe, Henke, & Simborg, 1984; Andersen, 1990, pp. 147–156), ecology (Hurlbert, 1984), animal behavior (Machlis, Dodd, & Fentress, 1985), and social sciences (Wolins, 1982, pp. 39–52). Surveys of its frequency are now numerous, with it often being found in 30%–80% of experimental studies examined (e.g., Gøtzsche, 1988; Kroodsma, 1990; Divine, Brown, & Frazier, 1992; Hurlbert & White, 1993; Heffner et al., 1996; Kroodsma, Byers, Goodale, Johnson, & Liu, 2001; Chuang, Hripcsak, & Heitjan, 2002; Kozlov, 2003; Thomas Ramsay, McAuley, & Grimshaw, 2003; Confrey & Stohl, 2004). Of four introductory statistics texts using real data sets, Alf & Lohr (2007) found that three recommended analyses that represented sacrificial pseudoreplication.

The term *pseudoreplication* is not one that was used prior to 1984, though it is now common in various biological and environmental sciences. The phenomenon has been and continues to be discussed under a variety of other labels, including *design effect, analytical mismatch, pooling fallacy, spurious replication, trial inflation, wrong sampling unit problem, confounding, variable and observational unit mistake*, and *unit of analysis error*. Typical of the terminological chaos in experimental design and statistics, none of these terms have ever been given specific, precise definitions. None except the somewhat awkward last term listed is widely used, and Murray (1998, pp. 104–107) discusses at length how "The phrase *unit of analysis* is the source of much confusion in the context of group-randomized trials."

Without doubt this chaos has impeded communication among statisticians serving different disciplines and between those statisticians and experimenters, and thus played a large role in fomenting erroneous analyses. SK seem unperturbed by the terminological chaos, have no well-defined alternative label for the phenomenon, and simply want editors and reviewers—and presumably others—to be barred from using the term *pseudoreplication*. An alternative view is that adoption by all disciplines of the error labels defined above, as well as the classical terminology of experimental design, would quickly improve education, communication, and statistical analysis. In their excellent review, Boruch & Foley (2000:213) emphasize how lack of a standardized terminology greatly complicates the searching and interpretation of the methodological literature—though they themselves seem happy to replace *experimental unit* with "primary unit of allocation and analysis"!

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 (e.g., Cox, 1958; Gill, 1978; Kempthorne, 1979; Kirk, 1982; Glass & Hopkins, 1984; Mead, 1988; Neter, Wasserman, & Kutner, 1990; Steel, Torrie, & Dickey, 1997). However, there also 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 (e.g., Altman, 1991; Winer, Brown, & Michels, 1991; Kantowitz, Roediger, & Elmes, 1997; Murray, 1998; Donner & Klar, 2000; Mitchell & Jolley, 2001). Consider the terminological tar pit that Winer et al. (1991) create for themselves and their readers. Instead of *experimental unit* they use "element assigned to a treatment class" (p. 80), they define "subject" as "the basic unit of observation" and then coin labels such as "subject-by-treatment designs" and "subjects-by-trial designs" (p. 220). They present a hypothetical example of two drugs being compared, one being given to patients in three randomly selected hospitals and the other being given to patients in three other hospitals (p. 359). Winer et al., (1991) label

this a "two-factor experiment [with a] hierarchical design," but in conventional terms it is nothing other than a simple unifactorial experiment with a completely randomized design, with multiple evaluation units in each experimental unit; other books use similar terminology or talk of "between-subjects designs" and "within-subjects designs" (e.g., Lindquist, 1940; Keppel, 1991; Kantowitz et al., 1997), rejecting the fully adequate classical terminology.

Nowadays, medical researchers might label Winer et al.'s (1991) drug experiment a "cluster randomized" or "group randomized" trial (e.g., Campbell & Grimshaw, 1998; Murray, 1998; Donner & Klar, 2000), but, like 'subject,' the term *cluster* (or *group*) is unneeded and ultimately confusion-generating even if it has helped wake up medical researchers to the fact that in many of their experiments the individual patient has *not* been the experimental unit and cannot be treated as such in a statistical analysis. The basic problem with *subject* and *cluster* is that *they do not represent statistical concepts*, lack clear definitions, have very general connotations, and hence transport all this baggage with them wherever used. An individual person, patient, pupil, or animal—or a group of them—in a given experiment could be an *experimental unit*, an *evaluation unit* within an experimental unit, or a *block* (as in crossover designs). So, in this regard, SK are correct when they say Hurlbert, like most *real* statisticians, believes "the individual organism holds no special status in experimental research."

Cluster sampling remains a useful term in *sampling design* where it labels one of the commonest types of sampling. Otherwise, continued use of labels such as subject, group, and cluster seems to be driven by a well-intentioned desire on the part of many authors of introductory statistics texts to employ vague or compromise terminology as a way of making clear that the statistical methods discussed have utility for both experimental and observational studies.

The term subject perhaps also appeals to researchers in medicine, psychology and education, because the ultimate objective of their experiments, however structured, usually is to understand and assist the individual person, pupil, or patient. In teaching I have long used an informal label, *material of prime interest* (MOPI), to denote the 'experimental material' or type of entity (or entities) a study is focused on without implying anything about what that entity may correspond to in the structure of an experimental design. Perhaps others can find *MOPI* useful.

Pooling in Simple Unifactorial Nested Designs

As it has historically been used to label a wide variety of procedures each of which has different costs and benefits, pooling is a large and complex topic. Here let me address only the situation where a researcher is considering whether to treat individual evaluation units as independent experimental units in order to increase power. Basic principles to consider are (Hurlbert, 1997): doing this will bias p values and confidence intervals unless there is no "experimental unit effect," that is, unless experimental units are *identical,* a matter that cannot be demonstrated with the often recommended "preliminary" significance tests; the bias will usually be downward for both p values and confidence interval widths; good power is normally obtained by increasing replication, increasing homogeneity of experimental units, and/or using design techniques such as blocking, and statistical controls such as covariates; and pseudoreplication may seem to increase power but the increase is spurious, not real.

SK, along with many others (e.g., Bancroft, 1964; Hairston, 1989, p. 33, Sokal & Rohlf, 1995, pp. 715–730, Underwood, 1997, p. 268, Zar, 1999, pp. 500–505, Quinn & Keough, 2002, p. 260), find test-qualified sacrificial pseudoreplication acceptable, saying "If there is no statistically significant effect among cages or pens within conditions, then it may be justified to pool individuals within conditions and ignore cages or pens as a unit of analysis." SK believe it is possible to select "an alpha level for concluding there is no [experimental unit] effect," but in fact even very high p values, for example, >0.25, give no grounds for preferring the null over the alternative hypothesis, that is, for concluding the true variability among experimental units to be zero. SK can and do claim that, in some sense, setting alpha at 0.25 for these purposes is "objective," but that does not eliminate bias in the p value from a consequent test for a treatment effect. The alpha selected for the preliminary test only sets a limit to the magnitude of the bias, a limit that will vary according to the specifics of an experimental design. Statisticians (e.g., Barcikowski, 1981; Zucker, 1990; Hines, 1996; Janky, 2000; Jenkins, 2002; Kromrey & Dickinson, 2007; Picquelle & Mier, 2009) are much less accepting of this sort of "pooling" or test-qualified pseudoreplication than are biologists, psychologists, and others. The general advice of the former is, if you want more power, design better. Even Sokal & Rohlf (1995), though they present Bancroft's (1964) recipe in detail, give it no strong recommendation, saying "One might well pool . . . [but] the experimenter cannot go wrong by not pooling . . ."

χ^2 and G-tests have long been misused in analyses of categorical data, and sacrificial pseudoreplication resulting from "pooling" – or "lumping," as Joy the Poet put it—is one of the commonest errors (Lewis & Burke, 1949; Wolins, 1982; Hurlbert, 1984; Kramer & Schmidhammer, 1992; Hurlbert & White, 1993; Wickens, 1993; Lombardi & Hurlbert, 1996; Hurlbert & Meikle, 2003). In their survey of zooplanktivore experiments, Hurlbert & White (1993) found that of 23 papers using χ^2 or G-tests, **78%** committed sacrificial pseudoreplication using those tests, 22% committed simple pseudoreplication, and only 13% committed neither of those errors. In contrast, of the 48 papers using methods for continuous variables (t test, ANOVA, U test, Kruskal-Wallis test), only **17%** committed sacrificial pseudoreplication, 10% committed simple, and 75% committed neither. The literature of experimental population genetics is rife with this problem, an Augean stable waiting for some young Hercules (or Hercula) with energy to spare.

This strong association of sacrificial pseudoreplication with χ^2 and G-tests, which is a widespread phenomenon, has an obvious source: the bad or misleadingly incomplete advice long given on pooling of replicate sets of "two-cell samples" of categorical data in widely used statistics texts (e.g., Snedecor & Cochran, 1989, pp. 202–206, Sokal & Rohlf, 1995, pp. 715–730, Zar, 1999, pp. 500–505, Steel et al., 1997, pp. 516–518). The problems reflect confusion of evaluation units with experimental units in situations where the response variable is a *continuous* one, such as a percentage or ratio that is measured by assessing the state of a *binary categorical* variable (e.g., green peas/yellow peas; male/female; alive/dead; responded/did not respond) for multiple evaluation units in each experimental unit. Such response variables can be analyzed with the standard statistical methods for continuous vari-

ables. However, at some point in some influential minds, these cases got mixed up with situations where the entity being categorized is indeed a *bona fide* experimental unit and not just an evaluation unit, for example, as when at a medical center 30 patients (= experimental units) are randomly assigned to one medication and 30 to another, and the response variable is survival (yes/no) at the end of 2 years.

The bad advice is fairly standard. A "heterogeneity" χ^2 or G-test, with alpha set at 0.05, is required by these books to see whether the experimental units are "homogeneous." If $p > .05$, it is concluded they are and their separate sets of evaluation units are pooled together. Via another χ^2 or G-test this table now can be tested for goodness-of-fit to some specified percentage (e.g., 50% males) or ratio (e.g., 3:1 in a genetics crossing trial), or it can be tested for a difference from a similarly constructed table yielded by another experimental treatment. If the preliminary test, however, does find, with $p < .05$, "heterogeneity" among experimental units, the next procedural step is left unclear. Sokal & Rohlf (1995, p. 721), for example, suggest we need to discover "Precisely which samples are homogeneous and which samples are different from the rest significantly enough to cause the heterogeneity?" They offer a convoluted way to attempt that quixotic quest, but they never get around to saying what to do with "deviant" samples [= experimental units]. Similarly, Snedecor & Cochran (1989) recommend upward adjustment of the standard error for the estimated mean proportion when "heterogeneity" is "significant." That replicate experimental units never need be presumed identical or "homogeneous" and that a simple *t* test would be adequate to compare two sets of proportions, or to compare one set with a hypothesized value, are facts completely missing from these accounts. These issues were discussed with clarity by Lewis & Burke (1949, pp. 447, 471) long ago and by Hurlbert (1984, p. 206) with a simple hypothetical example concerning effects of fox predation on vole sex ratio. Wickens (1993) gives an excellent discussion of how to treat replicate 2×2 contingency tables without falling into another variety of test-qualified sacrificial pseudoreplication.

The reader should be struck by the conspicuous lack of consistency between the textbook pooling "recommendations" in the two situations we have considered. With a nested ANOVA situation, the test for differences among experimental units is *optional* and pooling is also said to be an option but perhaps only when the test yields a Bancroftian $p \geq .25$. With the categorical data situation, the test for differences ("heterogeneity") among experimental units is *obligatory* and pooling is an option so long as the test yields $p \geq .05$; but if you don't like that option or if $p < .05$, you are left to stumble along on your own. Obviously many textbook authors have not yet fully thought these things out. You cannot go wrong, however, if you operate with the principles in the first paragraph of this section.

Multilevel Modeling

SK are correct in that the more sophisticated multilevel modeling procedures now available were not readily available when I was writing my 1984 paper. This was particularly true of methods for highly hierarchical data sets where one wished to use as explanatory variables, variables defined at levels in the hierarchy (e.g., student test score, teacher experience, mean family income for school district; or, tissue lipid content, fish length, tank ammonia concentration) other than that of the experimental unit. Every scientist new to this area would benefit from reading the first chapter or two of books on this topic cited by SK (e.g., Hox, 1995; Sniders & Bosker, 1999; Kreft & de Leeuw, 1999), as well as more recent ones (e.g., Raudenbush & Bryk, 2002; Goldstein, 2003; Bickel, 2007).

On the other hand, the mere *availability* of those methods is *not* "solving these problems" of pseudoreplication, and SK present no survey evidence indicating they have "largely disappeared . . . in sociology and education research," over, say, the last decade. None of the cases of pseudoreplication reported in the 251 experimental papers examined by Hurlbert (1984) and Hurlbert and White (1993) would have been prevented by the new multilevel methods. At least I can recall no experiment there where, for example, the evaluation unit was treated as the experimental unit because of a desire to use as a covariate a characteristic of the evaluation unit. In the cases reported, avoidance of pseudoreplication required understanding only such elementary matters as when a *t* test, not a χ^2 or G-test, was required, or the distinction between evaluation unit and experimental unit. With such understanding, statistical methods widely available for many decades would have sufficed for all but the cases of simple pseudoreplication. And where no covariates are involved and the only objective is assessment of treatment effects on the experimental unit, as it often is, averaging over evaluation units and using, in a simple ANOVA, the mean for each experimental unit sacrifices no information and gives exactly the same result for the treatment effect test as would a nested ANOVA. Where there are additional objectives, clearly other approaches can be employed.

In conclusion, reform and standardization of terminology in statistics, experimental design, and sampling design is badly needed, is possible, and would improve statistical practice. Clear labels for specific types of statistical error, like pseudoreplication, can play an important part in such a reform.

References

Alf, C., & Lohr, S. (2007). Sampling assumptions in introductory statistics classes. *American Statistician, 61,* 71–77.

Altman, D. G. (1991). *Practical statistics for medical research.* London: Chapman and Hall.

Andersen, B. (1990). *Methodological errors in medical research.* Oxford: Blackwell.

Arny, A. C., & Steinmetz, F. H. (1919). Field technic in determining yields of experimental plots by the square yard method. *Journal of the American Society of Agronomy, 11,* 81–106.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics, 20,* 427–442.

Barbacki, S., & Fisher, R. A. (1936). A test of the supposed precision of systematic arrangements. *Annals of Eugenics, 7,* 189–193.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267–285.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325–335.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press.

Boruch, R. F. & Foley, E. (2000). The honestly experimental society: Sites and other entities as the units of allocation and analysis in randomized trials. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy, vol.* 1 (pp. 193–238). Thousand Oaks, CA: Sage.

Box, G. E. P. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building.* New York: Wiley.

Box, J. F. (1978). *R. A. Fisher: The life of a scientist.* New York: Wiley.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8,* 158–233.

Campbell, M. K., & Grimshaw, J. M. (1998). Cluster randomized trials: Time for improvement. *British Medical Journal, 317,* 1171–1172.

Chuang, J.-H., Hripcsak, G., & Heitjan, D. F. (2002). Design and analysis of controlled trials in naturally clustered environments. *Journal of the American Medical Informatics Association, 9,* 230–238.

Cochran, W. G. (1980). Fisher and the analysis of variance. In S. E. Fienberg & D. V. Hinckley (Eds.), *R. A. Fisher: An appreciation (Lecture notes in statistics, Vol. 1)* (pp. 17–34). New York: Springer.

Confrey, J., & Stohl, V. (Eds.). (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations.* Washington DC: National Academies Press.

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

Cox, D. R. (1958). *Planning of experiments.* New York: Wiley.

Divine, G. W., Brown, J. T., & Frazier, L. M. (1992). The unit of analysis error in studies about physicians' patient care behavior. *Journal of General and Internal Medicine, 7,* 623–629.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold.

Fechner, G. T. (1860). *Elemente der Psychophysik, 2 vols.* [Vol. 1 translated by H. E. Adler in 1966 as *Elements of psychophysics,* D. H. Howes & E. G. Boring (Eds.), New York: Holt, Rinehart & Winston].

Fisher, R. A. (1925, 1958). *Statistical methods for research workers, 1st, 13th edns.* London: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* London: Oliver & Boyd.

Fisher, R. A. (1943). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association, 38,* 103–104.

Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manorial response of different potato varieties. *Journal of Agricultural Science, 13,* 311–320.

Gill, J. L. (1978). *Design and analysis of experiments in the animal and medical sciences, 3 vols.* Ames, IA: Iowa State University Press.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology, 2nd Ed.* Boston: Allyn & Bacon.

Goldstein, H. (2003). *Multilevel statistical models, 3rd edn.* London: Arnold.

Gøtzsche, P. C. (1988). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials, 10,* 31–56.

Hairston Sr., N. G. (1989). *Ecological experiments.* Cambridge, U.K.: Cambridge University Press.

Heffner, R. A., Butler, M. J., & Reilly, C. K. (1996). Pseudoreplication revisited. *Ecology, 77,* 2558–2562.

Hersh, R. K. (1924). The effect of temperature upon the full-eyed race of Drosophila. *Journal of Experimental Zoology, 39,* 43–53.

Hines, W. G. S. (1996). Pragmatics of pooling in ANOVA tables. *American Statistician, 50,* 127–139.

Hopkins, K. D. (1982). The unit of analysis: Group means vs. individual observations. *American Educational Research Journal, 19,* 5–18.

Hox, J. J. (1995). *Applied multilevel analysis.* Amsterdam: TT-Publikaties.

Hurlbert, S. H. (1981). A gentle depilation of the niche: Dicean resource sets in resource hyperspace. *Evolutionary Theory, 5,* 177–184.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs, 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. (1997). Experiments in ecology [Review of book of same title by A. J. Underwood]. *Endeavour, 21,* 172–173.

Hurlbert, S. H. (2004). On misinterpretations of pseudoreplication and related matters. *Oikos, 104,* 591–597.

Hurlbert, S. H., & Lombardi, C. M. (2003). Design and analysis: Uncertain intent, uncertain result [Review of *Experimental design and data analysis for biologists,* by G. P. Quinn & M. J. Keough]. *Ecology, 83,* 810–812.

Hurlbert, S. H., & Lombardi, C. M. (2004). Research methodology: Experimental design sampling design, statistical analysis. In M. M. Bekoff (Ed.), *Encyclopedia of Animal Behavior, 2,* 755–762. London: Greenwood Press.

Hurlbert, S. H., & Meikle, W. G. (2003). Pseudoreplication, fungi, and locusts. *Journal of Economic Entomology, 96,* 533–535.

Hurlbert, S. H., & White, M. D. (1993). Experiments with freshwater invertebrate zooplanktivores: Quality of statistical analyses. *Bulletin of Marine Science, 53,* 128–153.

Janky, D. G. (2000). Sometimes pooling for analysis of variance hypothesis tests: A review and study of a split-plot model. *American Statistician, 54, 269–*279.

Jenkins, S. H. (2002). Data pooling and type I errors: A comment on Leger & Didrichsons. *Animal Behaviour, 63,* F9–F11.

Kantowitz, B. H., Roediger III, H. L., & Elmes, D. G. (1997). *Experimental psychology: Understanding psychological research, 6th edn.* New York: West Publishing.

Kempthorne, O. (1979). *The design and analysis of experiments, corrected edn.* New York: Krieger.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* Englewood Cliffs: Prentice Hall.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences, 2nd edn.* Pacific Grove, CA: Brooks/Cole.

Koch, G. G., Amara, I. A., Stokes, M. E., & Gillings, D. B. (1980). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *International Statistical Review, 48,* 249–265.

Kotz, S. (1965). Statistical terminology - Russian vs. English - in the light of the development of statistics in the U.S.S.R. *American Statistician, 19,* 16–22.

Kozlov, M. V. (2003). Pseudoreplication in Russian ecological publications. *Bulletin of the Ecological Society of America, 84,* 45–47. [Condensation of original article published in Russian in *Zhurnal Obstchei Biologii* [Journal of Fundamental Biology], *64,* 292–397].

Kozlov, M. V., & Hurlbert, S. H. (2006). Pseudoreplication, chatter, and the international nature of science: A response to D. V. Tatarnikov. *Zhurnal Obstchei Biologii* [Journal of Fundamental Biology], *67*(2), 128–135 [In Russian; English translation available as a pdf].

Kramer, M., & Schmidhammer, J. (1992). The chi-squared statistic in ethology: Use and misuse. *Animal Behaviour, 44,* 833–841.

Kreft, I., & de Leeuw, J. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage.

Kromrey, J. D., & Dickinson, W. B. (2007). Detecting unit of analysis problems in nested designs: Statistical power and type I error rates of the F test for groups-within-treatments effects. *Educational and Psychological Measurement, 56,* 215–231.

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 M. Bekoff & D. Jamieson (Eds.), *Interpretation and explanation in the study of animal behavior, vol. II* (pp. 226–245). Boulder, CO: Westview Press.

Kroodsma, D. E., Byers, B. E., Goodale, E., Johnson, S., & Liu, W.-C. (2001). Pseudoreplication in playback experiments, revisited a decade later. *Animal Behaviour, 61,* 1029–1033.

Lewis, D., & Burke, C. J. (1949). The use and misuse of the chi-square test. *Psychological Bulletin, 46,* 433–489.

Lindquist, E. F. (1940). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Lombardi, C. M. (2007). Animal use in experimental psychology. In M. Bekoff (Ed.), *Encyclopedia of human-animal relationships, vol. 4,* (pp. 1269–1274). Westport, CT: Greenwood Press.

Lombardi, C. M., & Hurlbert, S. H. (1996). Sunfish cognition and pseudoreplication. *Animal Behaviour, 52,* 419–422.

Machlis, L., Dodd, P. W. D., & Fentress, J. C. (1985). The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Zeitschrift fur Tierpsychologie, 68,* 201–214.

McCall, W. A. (1923). *How to experiment in education.* New York: MacMillan.

Mead, R. R. (1988). *The design of experiments.* New York: Cambridge University Press.

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data, vol. I: Designed experiments.* New York: Van Nostrand Reinhold.

Mitchell, M., & Jolley, J. (2001). *Research design explained, 4th edn.* New York: Harcourt College.

Murray, D. M. (1998). Design and analysis of group-randomized trials. Oxford: Oxford University Press.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models.* Boston: Irwin.

Neyman, J., & Pearson, E. S. (1938). Note on some points in "Student's" paper on "Comparison between balanced and random arrangements of field plots." *Biometrika, 29,* 380–388.

Oksanen, L. (2001). Logic of experiments in ecology: Is pseudoreplication a pseudoissue? *Oikos, 94,* 27–38.

Picquelle, S. J., & Mier, K. L. (2009). Avoiding pseudoreplication in observational studies: Statistical methods for comparing means of subsampled populations. *Canadian Journal of Fisheries and Aquatic Sciences* (under review).

Pittman, M. S. (1921). *The value of school supervision.* Baltimore: Warwick and York.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists.* Cambridge, U.K.: Cambridge University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, 2nd edn.* Thousand Oaks, CA: Sage.

Rosenberg, G. S., & Gelashvili, D. B. (2008). *Problems of ecological experiments: Planning and analysis of observations.* Tolyatti, Russia: Cassandra. [In Russian].

Rushdie, S. (1989). *The satanic verses.* New York: Viking.

Schank, J. C., & Koehnle, T. J. (2009). Pseudoreplication is a pseudoproblem. *Journal of Comparative Psychology, 123,* 421–433.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods, 8th edn.* Ames: Iowa State University Press.

Sniders, T., & Bosker, R. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage.

Sokal, R. R., & Rohlf, F. J. (1969, 1981, 1995). *Biometry: The principles and practice of statistics in biological research, 1st, 2nd, 3rd edns.* New York: Freeman.

Steel, R. G. D., & Torrie, J. H. (1980). *Principles and procedures of statistics: A biometrical approach, 2nd edn.* New York: McGraw-Hill.

Steel, R. G. D., Torrie, J. H., & Dickey, D. A. (1997). *Principles and procedures of statistics: A biometrical approach, 3rd edn.* New York: McGraw-Hill.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Student. (1923). On testing varieties of cereals. *Biometrika, 15,* 271–293.

Student. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika, 29,* 363–378.

Tatarnikov, D. V. (2005). On methodological aspects of ecological experiments (comments on M. V. Kozlov publication). *Zhurnal Obstchei Biologii* [Journal of Fundamental Biology] *66,* 90–93 [in Russian with English summary].

Thomas, R. E., Ramsay, C. R., McAuley, L. & Grimshaw, J. M. (2003). *British Medical Journal, 326,* 397–398.

Truesdell, C. (1987). *Great scientists of old as heretics in 'the scientific method'.* Charlottesville: University Press of Virginia.

Underwood, A. J. (1997). *Experiments in ecology.* Cambridge, U.K.: Cambridge University Press.

Urquhart, N. S. (1981). The anatomy of a study. *HortScience, 16,* 100–116.

Valiela, I. (2001). *Doing science: Design, analysis and communication of scientific research.* New York: Oxford University Press.

Whiting-O'Keefe, Q. E., Henke, C., & Simborg, D. W. (1984). Choosing the correct unit of analysis in medical care experiments. *Medical Care, 22,* 1101–1114.

Wickens, T. D. (1993). Analysis of contingency tables with between subjects-variability. *Psychological Bulletin, 113,* 191–204.

Wiley, R. H. (2003). Is there an ideal behavioural experiment? *Animal Behaviour, 66,* 585–588.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design, 3rd edn.* New York: McGraw-Hill.

Wolins, L. (1982). *Research mistakes in the social and behavioral sciences.* Ames: Iowa State University Press.

Zar, J. H. (1999). *Biostatistical analysis, 4th edn.* Upper Saddle River New Jersey: Prentice Hall.

Zucker, R. M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement, 50,* 731–738.

Received February 26, 2009
Revision received March 9, 2009
Accepted April 21, 2009 ■