Pseudoreplication capstone: correction of 12 errors in Koehnle & Schank (2009)

Stuart H. Hurlbert
Department of Biology, San Diego State University, San Diego CA 92181
shurlbert@sunstroke.sdsu.edu
MARCH 4, 2010

Abstract: Many errors of fact or attribution are found in a commentary by T.J. Koehnle and J.C. Schank on pseudoreplication in the November 2009 issue of the *Journal of Comparative Psychology*. A sampling of twelve of these are identified and corrected here. Collectively these corrections strongly refute the claim that "the core ideas behind pseudoreplication are based on a misunderstanding of statistical independence, the nature of control groups in science, and contexts of statistical inference."

Preamble

In the most recent issue of the *Journal of Comparative Psychology*, Schank & Koehnle (2009; hereinafter referred to as SK) presented an article titled *Pseudoreplication is a pseudoproblem* that criticizes the concept of pseudoreplication and my writings on it. Accompanying SK were five commentaries solicited by *JCP* editor Gordon Burghardt. My own commentary (Hurlbert 2009; hereinafter referred to as SH) very briefly pointed out some errors in SK but mostly focused on providing further historical perspective on the topic and suggesting improvements in terminology. Of the three other commentaries provided (Coss, 2009; Freeberg & Lucas, 2009; Wiley, 2009), only that by Coss (2009) expressed general agreement with SK's critique. His main complaint, however, seemed aimed at "rigid thinking" on the part of editors and reviewers whose misunderstandings of pseudoreplication have led to unwarranted criticisms and manuscript rejections. Coss was acknowledged by SK (p. 421) as having encouraged them to write their article. Finally, a response to the above four commentaries was provided by Koehnle & Schank (2009: hereinafter referred to s KS).

The purpose of the present note is to document errors in KS which, if left uncorrected, seem likely to foster additional "noise" in the literature on this and related topics. The errors I address are merely a sample of all those present in KS, but hopefully a sufficient one to inspire readers to regard KS as a whole *cum grano salis*.

There are some benefits, especially with respect to the efficient use of editors' and reviewers' time, of egalitarian, minimally reviewed exchanges such as these in the November 2009 issue of *JCP*. But such exchanges can also be counterproductive if many incorrect statements are allowed to stand. Thus when I submitted this catalogue of errors to *JCP* I suggested that it be reviewed not only by Koehnle and Schank but also by several professional (Ph.D.) statisticians with expertise in experimental design. This would eliminate from my list any item a consensus of these professional statisticians found to be wrong. Of course many items in this catalogue concern not technical matters but simply attribution to me of claims I have not made. Unfortunately but not unreasonably, the American Psychological Association does not allow publication of commentaries on commentaries, and other journals are unlikely to be interested in doing housekeeping for *JCP*.

It may be noted that neither SK nor any of the five commentaries were reviewed pre-acceptance by professional statisticians (G. Burghardt, pers. comm.) though SH was critiqued pre-submission by four professional statisticians at my own request.

In 1834, the Statistical Society of London (later to become the Royal Statistical Society) was formed and adopted as its motto, *Aliis exterundum, Let others thrash it out* (Cochran, 1976). With typical British understatement they failed, however, to say whether the "thrashing out" was best done pre-publication or post-publication. It has always been done in both manners. In principle, pre-publication would be best and give us the highest quality scientific literature. Realistically, given the limited time of reviewers and editors, the inadequacies of most statistics textbooks and university statistics curricula, and the resultant generalized confusion over many statistical issues, we will be forced for a long time to rely heavily on post-publication wrestling matches. Guidelines for making them maximally useful and efficient are therefore of high importance. (The SSL motto admittedly actually was intended to refer to the idea that responsibility for interpretation of statistical information was best left to decision makers, subject matter specialists and so on, and should not be assumed by statisticians.)

The errors

Here then are some of the errors in KS that readers need to be aware of. KS persist in referring to a "doctrine of pseudoreplication," which they denote with "DP", while remaining unclear on the full set specific set of beliefs that includes. So where they

say "DP says this…" I have usually paraphrased that as something like "Hurlbert says this…."

This list of errors was submitted to Koehnle and Schank, to the four other discussants of SK (Richard G. Coss, Todd M. Freeberg, Jeffrey R. Lucas, R. Haven Wiley) earlier selected by *JCP* editor Burghardt, and to two professional statisticians. The latter were: Lyman McDonald (Senior Biometrician, WEST, Inc., associate editor of *Journal of Agricultural, Biological and Environmental Statistics*, former chairman of the Statistics Department at the University of Wyoming, and Fellow of the American Statistical Association) and N. Scott Urquhart (former professor of statistics, New Mexico State University & Colorado State University, associate editor of *American Statistician*, editor of *Environmental and Ecological Statistics*, and Fellow of the American Statistical Association). Given copies of SK, KS, and SH, each of the above eight persons was asked if there were any errors or inaccuracies in any of the twelve corrections. presented here.

Koehnle and Schank simply replied, "We see no error on our part" (T. Koehnle, email message to S. Hurlbert, 18 December 2009). Coss declined to comment, but the other three discussants and the two statisticians all replied that they could find no errors in my corrections.

1. KS, p. 452, 454, 456, 458 : "the relevance of Hurlbert's experimental units;" "the problem with the notion of experimental units;" Kozlov & Hurlbert (2006) are "offering a new definition of experimental unit;" "we can explicitly test whether there are experimental units by…;" "there are no *a priori* criteria for demarcating a given level as the level of experimental units."

Correction: The claim that I have advocated some new or controversial conceptualization of the experimental unit is false. Kozlov & Hurlbert (2006) acknowledged that their definition was a distillation of the definition put forward by Cox (1958, pp. 2, 155), which has been implicitly accepted by many researchers and statisticians including R.A. Fisher and W.S. Gossett, for at least half a century before Cox's book and by virtually all books on experimental design published since then that use the term *experimental unit*. Likewise, it is false that for a given study one cannot demarcate the criteria for defining the experimental unit *a priori*. Indeed, until one has done so, one cannot determine what materials and facilities will be needed for the experiment, nor can any decisions be made as to the specific procedures to be employed in setting up, treating, managing, and monitoring the experimental units during the course of the experiment. The implication of KS that the existence of experimental units in an experiment is something that one can "test" for is highly curious.

2. KS, pp. 452, 453, 456: "We [SK] found that averaging within experimental units decreases statistical power…Presumably [Hurlbert now believes] averaging among subjects within a given unit is no longer an essential step in formal data analysis;" Hurlbert says "data from subjects within experimental units is [sic] to be averaged;" "Hurlbert (1984), however, urged us to average measurements within experimental units to get a better estimate of error variance."

Correction: When multiple evaluation units are measured per experimental unit, I have never "urged" or said it was "essential" to use only the mean for each experimental unit in a significance test regardless of the objectives and nature of the full data set. I have pointed out that in the majority of the types of experiments my critiques have focused on, where covariates are not involved and interest is only in treatment effects, it is valid and sufficient to do analyses using only the mean value for each experimental unit. Whether that approach or a full nested ANOVA is carried out, the F and *P* values for the test for a treatment effect will be unchanged. By definition, the power of the test will also be unchanged, not "decreased".

3. KS, p. 452, 453: "Hurlbert claims that the simulated layout experiments [in SK] are not a block design… We fail to see how [these] experiments are not block designs."

Correction: This is the fourth time, since I first reviewed SK for *JCP* in 2001, that I have pointed out to the authors that SK, and now KS, have ignored the conventional definitions of *block* and *blocking* as used in the field of experimental design. They have persisted in synonymizing *block* and *experimental unit* (SK, p.425, col. 1, bottom) and in refusing to read (to judge from their Reference sections) books repeatedly recommended to them that would explain the difference. As described and figured in SK, their simulated experiments definitively do not have block designs, and they analyzed those experiments with ANOVAs appropriate to completely randomized designs, not randomized block designs.

4. KS, p. 453: "as [Kreft & de Leeuw (1999)] show, pooling across levels is at the heart of multilevel

modeling. One cannot reject pooling and accept multilevel modeling."

Correction: The citation for "Kreft & de Leeuw (1999)" in SK (and copied into SH and KS) is incorrect with respect to title and date (J. de Leeuw, pers. comm.). Presumably Kreft & de Leeuw (1998) was intended.

The sort of pooling being referred to would include, e.g., situations where, in a simple experiment with measurements made on multiple evaluation units in each experimental unit, "among experimental unit" and "among evaluation unit" sums of squares are pooled, if after "testing" for differences among experimental units with some arbitrarily specified alpha, e.g. 0.05 or 0.25) a $P >$ alpha is found. The new error mean square, with its increased degrees of freedom, is then used to test for the treatment effect. This has been termed "test-qualified sacrificial pseudoreplication" and its propensity for producing biased P values discussed (Hurlbert 1997; SH). Multilevel modeling, including simple nested ANOVAs, does not require that one consider such pooling procedures to be a valid option. Multilevel modeling remains a valuable tool even if one rejects "test qualified pseudoreplication" as acceptable procedure. Nothing in Kreft & de Leeuw (1998) contradicts this. They do claim that the "reliability of results" will not be much affected by pooling unless the "intra-class correlation" (e.g. true differences among experimental units under the same treatment) is "significant and substantial" (p. 4), two quite subjective criteria. But they do not claim or imply such pooling is "at the heart of multilevel modeling."

5. KS, pp. 453, 454, 455: Hurlbert's writings reflect "a misunderstanding of independence in statistical inference;" "Consider…fish sharing the same pond. At any given time, sampling one of the fish and measuring its body size will not provide any information about the body size of other fish. These measures are independent;" "Hurlbert's definition of independence (Kozlov & Hurlbert, 2006) is not consistent with the definition of independence in probability theory;" "We see no way to argue that two fish sharing the same aquarium violate the assumption of statistical independence…"

Correction: Kozlov & Hurlbert (2006) attempted *no* definition of statistical independence but only described how experimental units had to be defined and "dealt with independently" during the experiment if measurements made on separate experimental units were to possess statistical independence. KS's comments on the fish examples reflects their misunderstanding of these concepts. As pointed out in a critique (Hurlbert 1997) of another work reflecting the same misunderstanding: "statistical independence…can be evaluated only in reference to both a data set and a specified hypothesis. If we take a random sample of bug density from each of two plots, the 'errors' (epsilons) will possess the statistical independence needed for testing the H_0: *no difference between plots*. But, in the case where one plot has been sprayed with an herbicide and the other kept as a control, these errors will *not* possess the statistical independence required for testing the H_0: *no difference between treatments*." The same applies to fish in ponds or aquaria.

6. KS, p. 454: Hurlbert claims his "definition of experimental unit…does not apply to urns because urns by definition cannot be experimental units… We do not see how the newer definition would exclude an urn, or a ladle of balls taken from the urn…"

Correction: I nowhere say or imply that urns cannot be experimental units. It is easy to conceive of them being such if, for example, each contains mice or a bacterial culture and half the urns get some experimental treatment and the other half are kept as controls. The urn exercise in SK, however, is *not* a manipulative experiment, and so by definition experimental units are not involved in that exercise.

7. KS, p. 454: "…after several pages of describing why designs of type B are pseudoreplicated…"

Correction: I have never described any design as "pseudoreplicated," and have pointed out in multiple publications that pseudoreplication "is simply an error of statistical analysis and interpretation and is not merely a weak design or an inevitable consequence of such" (SH, p. 437).

8. KS, p. 455: "Hurlbert also believes that his concerns about spatiotemporal proximity and thus his definition of experimental units eliminates subjects, or more generally, individual organisms, as experimental units."

Correction: Completely false. First, my "concerns" clearly have not been about spatiotemporal proximity *per se*, but only about potential interactions among experimental units and potential segregation of treatments. Second, I have never said or implied individual organisms cannot serve as experimental units, and well-designed experiments using them as such are common. Indeed my own first scientific paper (Hurlbert 1961) reported experiments in which the experimental units were individual chickadees

housed in their individual cages -- with the cages in close "proximity" to each other -- 10-15 cm apart, as I recall.

9. KS, p. 455: "Hurlbert claims there are 'satisfactory criteria for drawing boundaries around experimental units' in every discipline…, but never gives any criteria for any discipline…He offers examples, such as rooms, pens, aquaria and bottles, but there is no logical argument for why they are so special….[W]e are not typically interested in studying rooms, pens, aquaria, bottles, and so forth. We are interested in studying the things we put into them such as rats, goats, fish, and fruitflies."

Correction: These comments seem disingenuous. It is quite normal in every discipline to use a simple label for the experimental unit as the latter is always a complex, multi-component entity even if measurements are actually made on one or a limited number of those components. Thus the label "aquarium" is shorthand for the water, fish, heater, filter, bacteria, and everything else in it, just as the label "plot" is reasonable shorthand for the grain, weeds, soil, soil arthropods, and everything else in it. None of the entities listed by KS is "so special" -- except when experiments are so designed that names of those entities are reasonable labels for the experimental units in those experiments. And the "criteria for drawing boundaries around experimental units" are the same in "any discipline" as they are in "every discipline" and pretty universally accepted (see SH, p. 436).

10. KS, p. 455: "…from a strict reading of the original 1984 article, a pseudoreplicationist must conclude that no inferential knowledge can be properly drawn from within a single lake or watershed. We do not attribute this view to Hurlbert personally…"

Correction: KS nowhere define "pseudoreplicationist," but apparently it is intended as a pejorative label for persons who believe that pseudorepliction is an error and should be avoided. In that case, I certainly am one, their disclaimer excusing me notwithstanding. But nothing in Hurlbert (1984) can justify any careful reader reaching the conclusion that KS state. Many published ecosystem studies, both observational and experimental, have been focused on a single lake or watershed and yet achieved valid inferences.

11. KS, p. 456: "Hurlbert states that he views physical control as "often non-essential" in ecological experiments…It is not clear how he came to this view, but his argument is implicitly undermined by the importance of interspersion [in his conception of good experimental design]."

Correction: My original article (Hurlbert, 1984, p. 191) dedicated a whole page to sorting out the confusing, multiple ways in which the word "control" has been used in the context of experimental design. SK and now KS now are reinjecting more confusion. My original reference was to "regulation of the physical environment in which the experiment is conducted." SK (p. 427) inaccurately paraphrased this as "physical control or regulation of the environment *for the control of possible intervening variables* [italics supplied]," which in turn inspired KS's comment on interspersion of treatments. Interspersion and design structure in general (*sensu* Urquhart, 1981, Hurlbert & Lombardi, 2004, and SH) are completely separate matters from, for example, whether temperature in the room in which all your aquaria are maintained is kept constant or allowed to fluctuate, or whether you do your experiment indoors or outdoors.

12. KS, p. 456: Hurlbert "prizes the calculation of a *p* value for a null test: 'if there are no design or statistical errors, the confidence with which we can reject the null hypothesis is indicated by the value of P alone' (Hurlbert 1984, p. 191). Thinking of this sort has elevated the Type I error rate into sort of a bronze bull. ….It is a simple mistake, however, to assume that the results of a single experiment can be generalized based merely on rejection of the null hypothesis or the *p* value…"

Correction: It is a simple mistake for KS to think readers will know what they mean here by "thinking of this sort." One can infer they believe "bronze bulls" are bad but not much more about what these really are. SK imply that I have "assumed" the false proposition they state, but they present no evidence of that and could not find any in anything I have written. An up-to-date discussion of *P* values and significance testing may be found in Hurlbert & Lombardi (2009).

Acknowledgments

Thanks to the discussants and reviewers named at the beginning of this article who were willing to take the time to put a microscope to this manuscript and related ones and thereby help clear the air.

References

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (ed.), *On the history of statistics and probability* (p. 8). New York: Marcel Dekker.

Coss, R. G. (2009). Pseudoreplication conventions are testable hypotheses. *Journal of Comparative Psychology 123*, 444-446.

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

Freeberg, T. M. & Lucas, J. R. (2009). Pseudoreplication is (still) a problem. *Journal of Comparative Psychology, 123*, 450-451.

Hurlbert, S. H. (1961). Further evidence in support of the searching image hypothesis. Honors thesis (B.A.), Department of Biology, Amherst College, Amherst, Massachusetts.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187-211.

Hurlbert, S. H. (1997). Experiments in ecology [Review of book by same title by A.J. Underwood]. *Endeavour, 21*, 172-173.

Hurlbert, S. H. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology, 123*, 434-443.

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. & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*, 311-349.

Koehnle, T.J. & Schank, J. C. (2009). An ancient black art. *Journal of Comparative Psychology 123,* 452-458.

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].

Kreft, I. & de Leeuw, J. (1998). *Introducing multilevel modeling*. Los Angeles, California: Sage.

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

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

Wiley, R.H. (2009). Trade-offs in the design of experiments. *Journal of Comparative Psychology, 123*, 447-449.