ANIMAL BEHAVIOUR, 2001, **61**, 1029–1033
doi:10.1006/anbe.2000.1676, available online at http://www.idealibrary.com on IDEAL®

COMMENTARIES

Pseudoreplication in playback experiments, revisited a decade later

DONALD E. KROODSMA, BRUCE E. BYERS, EBEN GOODALE, STEVEN JOHNSON & WAN-CHUN LIU
Department of Biology, University of Massachusetts

(Received 30 May 2000; initial acceptance 15 August 2000; final acceptance 27 September 2000; MS. number: AS-1259)

About 10 years ago, several papers in *Animal Behaviour* addressed the quality of experimental designs in 'playback' experiments (Kroodsma 1989a, b, 1990a, 1992; Searcy 1989; Weary & Mountjoy 1992), and this debate culminated in a consensus report by McGregor et al. (1992). The key issue was 'pseudoreplication', defined by Hurlbert (1984, page 187) as 'the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent'. McGregor et al. (1992, page 2) offered their own simplified definition, 'the use of an n (sample size) in a statistical test that is not appropriate to the hypothesis being tested'. McGregor et al. agreed that pseudoreplication was a serious issue, and that designing and implementing good experimental designs was a worthy and attainable goal.

What effect did the debate and subsequent consensus report have on the quality of experimental designs used in animal behaviour? To answer that question, we surveyed the designs used in 50 papers published during the last several years. The papers were chosen by searching electronic databases for 25 papers that cited a key paper on pseudoreplication and for 25 other 'playback' papers that did not explicitly cite or address pseudoreplication issues. We reasoned that these two samples of papers should provide an index to the experimental designs and logic currently being used by investigators in animal behaviour. (Note: we chose not to review playback experiments that used synthetic stimuli. Although use of synthetic stimuli may solve some problems (e.g. see McGregor et al. 1992), we encountered a number of papers in which we felt that interpretations based on large sets of synthesized playback variants exceeded the permitted inferential space.)

Correspondence: D. E. Kroodsma, Department of Biology, University of Massachusetts, Amherst, MA 01003, U.S.A. (e-mail: kroodsma@bio.umass.edu).

Our reviews show that some progress has been made in eradicating the simplest of the pseudoreplication problems (Fig. 1). A decade ago, the most commonly used experimental designs involved 'simple pseudoreplication' (Hurlbert 1984), in which only a single exemplar from a class of stimuli was used to test hypotheses about the class itself (e.g. using only a_1 and b_1 to ask questions about Class A and Class B; see Table 1). In papers reviewed by Kroodsma (1990b), for example, 15 of 22 papers (68%) used simple pseudoreplication. In these studies, treatments were unreplicated, although users of such designs typically analysed results as if multiple replicates had been performed. Our sample of recently published papers revealed that this faulty design is used more rarely now (Fig. 1), indicating that authors are increasingly aware of the need for multiple stimuli to represent a class of stimuli.

Some authors who continue to use this simple, faulty design seem unaware of pseudoreplication issues, but others, in our opinion, misunderstand some of the basic problems. For example, one approach that some investigators now use to 'minimize' pseudoreplication is to combine several stimuli (e.g. a_1, a_2, and a_3; see Table 1) into one presentation, and then use that composite stimulus as if it were representative of all stimuli in the class. Although a composite stimulus may be better than a single stimulus, we believe this approach still constitutes simple pseudoreplication, because only one stimulus (albeit a composite stimulus) is used to represent each class. Another approach that some investigators use is to identify, sometimes with great care, a 'typical' stimulus, and then assume that stimulus is representative of the class. As discussed by McGregor et al. (1992), however, that approach is not appropriate either. Another argument we encountered is that, because stimuli vary less within a class than between classes, only one representative of each class is needed. Or that, because the experimenter can detect little or no variability in a particular class of signal, only one signal is needed to represent each class. All of these arguments, we believe, make

0003–3472/01/051024+05 $35.00/0

© 2001 The Association for the Study of Animal Behaviour

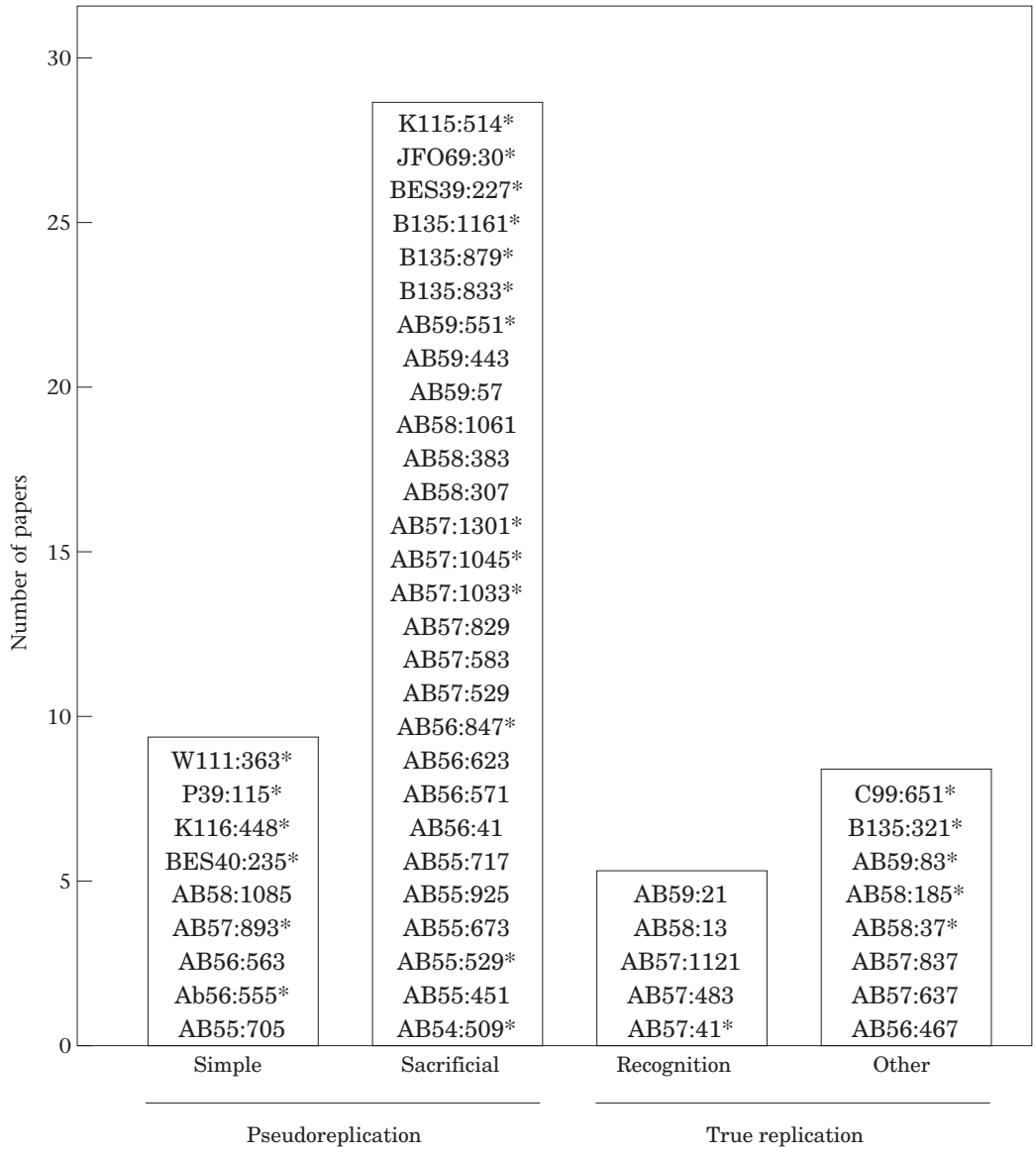

Figure 1. A survey of 50 recently published 'playback' papers, showing few cases of 'simple' pseudoreplication, much 'sacrificial' pseudoreplication (data pooled, thus biasing *P*), and about one in four papers with true replication. *Studies citing a key paper on pseudoreplication did not improve the quality of the experimental design, as pseudoreplication occurred in 76% (19 of 25) of papers that addressed the issue and in 72% (18 of 25) of papers that did not. Figure entries consist of a code for the journal name, volume and number of first page of surveyed articles. AB=*Animal Behaviour*, B=*Behaviour*, BES=*Behavioral Ecology and Sociobiology*, C=*Condor*, JFO=*Journal of Field Ornithology*, K=*Auk*, P=*Primates*, W=*Wilson Bulletin*.

unwarranted assumptions about how the animals themselves perceive the signals, and it is, after all, the animals' perception that we wish to test. Because the goal in our experimental designs is not to 'reduce' or 'minimize' pseudoreplication, but to eliminate it, these experimental designs are unacceptable because they still falsely replicate.

Although this simple form of pseudoreplication is now used infrequently, another form of pseudoreplication has become increasingly common (Fig. 1). Over half of the authors in our survey laudably used multiple stimuli to represent a class, but then pooled data within each class ('sacrificial pseudoreplication,' sensu Hurlbert 1984), an approach we find troublesome (see also Machlis et al. 1985). An example of such inappropriate pooling would be to use a_1, a_2, a_3, a_4 and a_5 to represent Class A, and b_1, b_2, b_3, b_4 and b_5 to represent Class B (see Table 1), test five animals with each replicate, and then pool data for all replicates within a class (perhaps after a statistical test shows 'no difference' among the replicates). Then, hypotheses comparing Class A to Class B would be tested as if the experiment contained 25 independent replicates of the comparison. We believe that this kind of data pooling is ill-advised.

Our objection to pooling the data is perhaps best explained with a hypothetical example of inappropriate pooling (see also Hurlbert 1997). Suppose that an investigator uses stimuli a_1–a_5 and b_1–b_5 to test for differences

Table 1. A sampling design for testing responses to two classes of stimuli, most appropriately analysed as a nested analysis of variance (ANOVA), without pooling of data from replicates within a class

Class A					Class B				
a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5
1	6	11	16	21	26	31	36	41	46
2	7	12	17	22	27	32	37	42	47
3	8	13	18	23	28	33	38	43	48
4	9	14	19	24	29	34	39	44	49
5	10	15	20	25	30	35	40	45	50

Numbers 1–50 refer to different subjects.

between Class A and Class B (see Table 1). With, say, five responses to each stimulus, he next tests for differences in responses within the stimuli a_1–a_5 and for differences in responses within the stimuli b_1–b_5. Finding a high *P* value for each test, he accepts the null hypothesis that the different stimuli within each class evoke identical responses, and therefore pools the data within each class. The problem lies, we believe, in interpreting this high *P* value as evidence 'confirming' the null hypothesis. A high *P* value could reflect low *N*, high variances, and/or small (but not zero) effect size. Typically, in playback studies (e.g. see Fig. 1), the sample sizes for these statistical tests are small and the power of the tests therefore low; as a result, these weak tests rarely (never in the references in Fig. 1) show differences in responses to the different replicates.

If our significance tests show 'no significant difference' among the a or b replicates, we have shown only that we have no strong evidence that they differ. We have not demonstrated that they are identical, and pooling of data would be appropriate only if responses are identical, as only then would our nominal alpha correspond to the true alpha in our test of the difference between the two stimulus classes. If we pool response data from nonidentical replicates, we no longer know which alpha we are using, and the *P* values are biased (see also Machlis et al. 1985). As a result, our confidence in the reported results is eroded.

Given the wide variety of difficulties and compromises faced by experimenters wishing to devise effective, valid experimental designs (Hurlbert 1984; McGregor et al. 1992), is it reasonable for us to advocate that extra effort be devoted to eliminating inappropriate pooling? We think so. To us, pooling data (with or without the weak statistical tests used to justify pooling) compromises our overall experimental efforts. This practice still constitutes pseudoreplication, because the final statistical test is inappropriate for the hypothesis being tested. It is inappropriate because, to the (unknown) extent that pooled stimuli are not identical, the alpha or *P* levels for the test have been altered. Accepting the pooled stimuli as identical after a weak statistical test casts only a 'veneer of rigor' (Hurlbert 1984) on sacrificial pseudoreplication. Using this veneer of statistics to justify inappropriate pooling serves only to preserve the common delusion of 10 years ago, that multiple stimuli are unnecessary because all signals within a class are essentially identical.

Although the majority (37 of 50, or 74%) of papers in our survey committed pseudoreplication, we were pleased to find that a number of authors had eliminated pseudoreplication from their experimental designs (see Fig. 1). One such group of papers focused on questions about topics related to individual recognition. When addressing these kinds of questions, investigators must choose playback stimuli that are specific to a subject's experience, because each subject in a neighbour-recognition study has different neighbours and therefore necessarily requires different playback stimuli to represent the neighbour category. Hence, the nature of the questions tends to force experimental designs that avoid pseudoreplication. Typical papers were those by O'Loghlen & Beecher (1999), Price (1999), Reby et al. (1999), Sayigh et al. (1999) and Beecher et al. (2000).

One lingering concern that we had for many of these neighbour-recognition papers in particular and for playback studies in general, however, was repeated use of some of the playback stimuli. A single 'stranger' stimulus could be used repeatedly, for example, if that stimulus was unknown to most of the subjects, but repeated use of such a stimulus would constitute pseudoreplication if each use were treated as an independent data point. In all experimental designs, we encourage investigators to maximize independence of samples by using stimuli a minimum number of times, or if signals are reused, to average the responses to obtain a single data point for response to that particular stimulus. We also encourage authors to explicitly inform readers how a given stimulus is used throughout an experiment; without that important information, critical readers cannot accept results at face value (we were forced to exclude several papers from this review because of insufficient information).

For us, the most satisfying (exciting, even) papers were those in which authors used true replication to address questions other than neighbour-recognition questions (right column in Fig. 1). The authors of these papers chose experimental designs that avoided pseudoreplication, and these designs were rare in the literature only a decade ago. Thus, we especially applaud these authors, for two reasons: (1) they provide experimental designs that can serve as models for others to use, and, most importantly, (2) they provide research results that are less likely

to be an artefact of some unacceptable statistical manoeuvre and are more likely to reveal the true perceptual worlds of animals. One such exemplary paper was that by Searcy et al. (1997), which also includes an excellent discussion of pseudoreplication issues; in one experiment, for example, Searcy et al. used 10 pairs of playback stimuli, with statistical tests based on the number of stimuli in the two classes, not the number of individuals tested with those stimuli. A similar approach was used by Searcy et al. (1999). In other papers, D. A. Nelson (1998) avoided pseudoreplication by testing 44 birds with 44 different stimuli, Trainor & Basolo (2000) by presenting 18 different video clips to 18 different females, B. S. Nelson & Stoddard (1998) by testing 10 birds with 10 different stimuli, Houx & ten Cate (1999) by testing eight pairs of birds with eight different stimuli. Butchart et al. (1999) obtained multiple responses from each of six females, but in their statistical tests used an average of the responses from each female, basing their statistics on the number of birds tested, not the total number of responses obtained. One study (Poole 1999) used only one stimulus to represent a class, but wisely only tabulated the results, foregoing statistical inference; we laud that practice here because we believe that, when treatments have not been replicated, fewer or no statistics are appropriate. More authors, we feel, should strive for such restraint when the data do not permit statistical inference (see also Hurlbert 1984).

Overall, our survey shows that some progress has been made in the quality of experimental playback designs over the last decade. Most investigators now use multiple stimuli to represent a class of stimuli; as a result, simple pseudoreplication is rarer, and true replication more common. About half of all papers, however, now pool data within stimulus classes, an approach that we believe alters the *P* value for a given statistical test, so that our confidence in the result is eroded (see also Machlis et al. 1985). We encourage investigators to abandon this practice, and instead use 'true replication' in pursuit of understanding the perceptual worlds of animals.

During the course of preparing this commentary, we had the opportunity to discuss our views on pseudoreplication with a number of colleagues. Some of these colleagues questioned our focus on what seemed to them to be a relatively narrow issue of statistical analysis. They worried that a preoccupation with problems in data analysis might deflect attention from more fundamental questions about experiments, such as whether a given design is adequate for testing an experimenter's stated hypothesis. After all, it is theoretically possible for an investigator to commit pseudoreplication during analysis of a powerful, well-designed experiment, or for the results of a weak, poorly designed experiment to be analysed without statistical errors. One might therefore be tempted to conclude that our emphasis on pseudoreplication will be viewed by readers as encouraging reviewers and journal editors to overlook serious errors in experimental design.

We emphatically do not wish for concern about pseudoreplication to preempt or supersede attention to any other aspect of experimental design and analysis. We cannot, however, agree that the problem of pseudoreplication is too narrow or trivial to justify our concern. Performing an appropriate, valid analysis of experimental data is one of the most basic components of sound science, and it is difficult to see how we can have any confidence in the results reported by an investigator who fails to execute this elementary task. Indeed, in our review of papers for this commentary, we found that pseudoreplication was frequently tied to weak experimental designs that made sound analysis impossible.

We believe that good investigators who are aware of the issue will find that eliminating pseudoreplication from their experiments is a relatively straightforward matter. The problems of pseudoreplication that we typically encounter in published studies are easily solvable, and we see no reason that reviewers and editors should accept studies that fail to eliminate pseudoreplication. We certainly do not believe that absence of pseudoreplication should be the sole, or even the main, criterion for evaluating experiments. Rather, we see the absence of pseudoreplication as representing a minimum requirement that should be met before the merits of an experiment are evaluated.

We thank Stuart Hurlbert and especially Mike Sutherland for invaluable help in thinking about statistical issues, as well as Dan Albano, Curtis Marantz, Jean-Claude Razafimahaimodison and Nick Thompson. We thank an anonymous referee and especially Mike Beecher for constructive thoughts on improving this manuscript.

References

Beecher, M. D., Campbell, S. E., Burt, J. M., Hill, C. E. & Nordby, J. C. 2000. Song-type matching between neighbouring song sparrows. *Animal Behaviour*, **59**, 29–37.

Butchart, S. H., Seddon, N. & Ekstrom, J. M. M. 1999. Yelling for sex: harem males compete for female access in bronze-winged Jacanas. *Animal Behaviour*, **57**, 637–646.

Houx, B. B. & ten Cate, C. 1999. Song learning from playback in zebra finches: is there an effect of operant contingency? *Animal Behaviour*, **57**, 837–845.

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

Hurlbert, S. H. 1997. Book Review: *Experiments in Ecology*, by A. J. Underwood. *Endeavour*, **21**, 172–173.

Kroodsma, D. E. 1989a. Inappropriate experimental designs impede progress in bioacoustic research: a reply. *Animal Behaviour*, **38**, 717–719.

Kroodsma, D. E. 1989b. Suggested experimental designs for song playbacks. *Animal Behaviour*, **37**, 600–609.

Kroodsma, D. E. 1990a. Using appropriate experimental designs for intended hypotheses in song playbacks, with examples for testing effects of song repertoire sizes. *Animal Behaviour*, **40**, 1138–1150.

Kroodsma, D. E. 1990b. 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: *Interpretation and Explanation in the Study of Animal Behavior* (Ed. by M. Bekoff & D. Jamieson), pp. 226–245. Boulder, Colorado: Westview Press.

Kroodsma, D. E. 1992. Much ado creates flaws. *Animal Behaviour*, **44**, 580–582.

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 für Tierpsychologie*, **68,** 201–214.

McGregor, P. K., Catchpole, C. K., Dabelsteen, T., Falls, J. B., Fusani, L., Gerhardt, H. C., Gilbert, F., Horn, A. G., Klump, G. M., Kroodsma, D. E., Lambrechts, M. M., McComb, K. E., Nelson, D. A., Pepperberg, I. M., Ratcliffe, L., Searcy, W. A. & Weary, D. M. 1992. Design of playback experiments: the Thornbridge Hall NATO ARW Consensus. In: *Playback and Studies of Animal Communication* (Ed. by P. K. McGregor), pp. 1–9. New York: Plenum Press.

Nelson, B. S. & Stoddard, P. K. 1998. Accuracy of auditory distance and azimuth perception by a passerine bird in natural habitat. *Animal Behaviour*, **56,** 467–477.

Nelson, D. A. 1998. Geographic variation in song of Gambel's white-crowned sparrow. *Behaviour*, **135,** 321–342.

O'Loghlen, A. L. & Beecher, M. D. 1999. Mate, neighbour and stranger songs: a female song sparrow perspective. *Animal Behaviour*, **58,** 13–20.

Poole, J. H. 1999. Signals and assessment in African elephants: evidence from playback experiments. *Animal Behaviour*, **58,** 185–193.

Price, J. J. 1999. Recognition of family-specific calls in stripe-backed wrens. *Animal Behaviour*, **57,** 483–492.

Reby, D., Cargnelutti, B. & Hewison, A. J. M. 1999. Contexts and possible functions of barking in roe deer. *Animal Behaviour*, **57,** 1121–1128.

Sayigh, L. S., Tyack, P. L., Wells, R. S., Solow, A. R., Scott, M. D. & Irvine, A. B. 1999. Individual recognition in wild bottlenose dolphins: a field test using playback experiments. *Animal Behaviour*, **57,** 41–50.

Searcy, W. A. 1989. Pseudoreplication, external validity and the design of playback experiments. *Animal Behaviour*, **38,** 715–717.

Searcy, W. A., Nowicki, S. & Hughes, M. 1997. The response of male and female song sparrows to geographic variation in song. *Condor*, **99,** 651–657.

Searcy, W. A., Nowicki, S. & Peters, S. 1999. Song types as fundamental units in vocal repertoires. *Animal Behaviour*, **58,** 37–44.

Trainor, B. C. & Basolo, A. L. 2000. An evaluation of video playback using *Xiphophorus helleri*. *Animal Behaviour*, **59,** 83–89.

Weary, D. M. & Mountjoy, D. J. 1992. On designs for testing the effect of song repertoire size. *Animal Behaviour*, **44,** 577–579.