ANIMAL BEHAVIOUR, 2003, **66**, 585–588
doi:10.1006/anbe.2003.2231

COMMENTARIES

Is there an ideal behavioural experiment?

R. HAVEN WILEY
Department of Biology, University of North Carolina, Chapel Hill

(Received 23 April 2002; initial acceptance 19 July 2002; final acceptance 4 October 2002; MS. number: AS-1275)

Much of the field of animal behaviour rests on experimental studies of the responses of animals to different classes of stimuli. Playback experiments, which compare responses of animals to tape recordings of different sounds, are a prime example. Consequently, the proper design of these experiments is central to the scientific study of animal behaviour. Over a decade ago, a discussion of the design of behavioural experiments focused on the problems of pseudoreplication (Kroodsma 1989a, b, 1990; Searcy 1989; McGregor et al. 1992; Weary & Mountjoy 1992), and a recent paper has reviewed subsequent progress in avoiding pseudoreplication in experimental studies of bird song (Kroodsma et al. 2001).

In this context, pseudoreplication consists of repeatedly presenting the same stimulus, repeatedly using the same subject, or pooling the results from presentations of similar stimuli, all problems identified by Hurlbert (1984) in some ecological experiments. Kroodsma et al. (2001) advocate a nested analysis of variance (ANOVA) to avoid these problems of pseudoreplication. In the proposed design, each subject receives a single presentation and each exemplar of a stimulus is used only once. This proposed design implies that there is only one experimental design ideally suited for comparisons of responses to different stimuli.

My objective here is not to challenge nested ANOVA but to expand the discussion of behavioural experiments. To this end, I identify some compromises any experimenter must make in justifying the biological independence of subjects, the external validity of conclusions, the multiple use of exemplars and subjects, and the effects of sample size on unsuspected bias. These compromises make it less clear that any one design is universally optimal. In some circumstances, it is appropriate to use each exemplar more than once and to test each subject more than once. Although there is no ideal experimental design, I propose that there is an ideal way to report a behavioural experiment, one that explicitly identifies the compromises involved.

The experimental design for a nested ANOVA derives from Fisher's pioneering work in the first half of the 20th century, which now forms the basis for innumerable textbooks on statistics and experimental design. A simple experiment might compare responses of crops to two different treatments of fertilizer. For instance, higher levels of fertilizer might be applied to one set of plots and lower levels to another set. Fisher emphasized that the plots for each treatment should be assigned at random, so that the plots prior to treatment constitute a single population in a statistical sense. An ANOVA could then compare the variance in responses within treatments to the variance between treatments.

This design is, of course, also used for many experimental studies of responses to two classes of stimuli. In such a behavioural study, treatments might be different categories of songs presented by playback of tape recordings. Individual subjects hearing these playbacks are analogous to the plots of an agricultural study.

Even such simple experiments require attention to some basic issues, especially independence of subjects' responses, external validity, multiple use of exemplars and subjects, and sample size. A reconsideration of these issues leads to the conclusion that any experiment involves some critical compromises. There is no single ideal experimental design. Instead, the objective should be to identify and to justify the compromises.

Biological Independence of Responses

The simple playback experiment just described assumes that each subject's response is biologically independent. In other words, the responses of subjects do not influence each other. The same concern might apply to the

Correspondence: R. H. Wiley, Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, U.S.A. (email: rhwiley@email.unc.edu).

0003–3472/03/$30.00/0

agricultural experiment. If rapid growth on one plot affected the growth of crops on other plots, for instance, bycompetition for light or water, then the assumption of biological independence fails, just as if vigorous response to playback by one subject influenced the responses of the next subject (or the same subject at a later time).

Biological interactions between the subjects of an experiment result in pseudoreplication in a strict sense of the word. When two subjects interact (either in a way that increases or decreases the differences in their responses), then the responses of the two are not biologically independent and, consequently, the apparent replication of treatments is compromised. Hurlbert's (1984) definition of psuedoreplication is broader than this one. He emphasizes 'statistical independence' of observations, a requirement that the values of any one observation not be a condition of the values of any other. It requires both biological independence of subjects and exclusion of the effects of extraneous variables, usually by randomization and interspersion of treatments.

An experiment could lack statistical independence of observations if one treatment was applied to one set of subjects first and then another treatment applied to a second set of subjects. Although the subjects might not interact biologically, the observations might lack statistical independence if seasonal changes affected subjects' responses. On the other hand, statistical independence might fail, even if treatments were randomized, if subjects interacted behaviourally as a result of the treatments. The first problem was clearly addressed by Fisher; the second problem is one that Hurlbert emphasized in his discussion of 'simple pseudoreplication'.

In a playback experiment with bird song, pseudoreplication in this strict sense occurs, for instance, when one subject's response influences another subject's response. This situation could arise when subjects are territorial neighbours. It might even occur when subjects hold territories within hearing of each other. It could also occur when subjects serve as their own controls, for use with paired statistical tests. When a subject receives two presentations, its response to the second might be influenced by its experience with the first.

There is no way to predict from first principles how far apart in space or time two presentations must be to avoid biological interactions. To make these decisions, experimenters often rely on experience with similar situations. Then, to compensate for any possible biological interactions, they randomize or permute the order of presentations to nearby subjects or to the same subject. In a study of discrimination between neighbours and strangers, for instance, the order of presentation of neighbours' and strangers' songs is usually randomized or balanced across subjects.

These procedures do not, of course, eliminate the possible influences of one stimulus presentation on the subjects' responses to others. Instead, they distribute these influences more or less evenly across the experiment. The only way to evaluate possible spatial and temporal interactions is to include the separation of presentations as a variable in the experimental design. This expanded design, however, would require a compromise, as a larger sample of subjects would be needed to examine the expanded set of hypotheses.

External Validity

In the earlier caveats about the design of playback experiments, one concern was the range of situations in the real world to which an experiment applies, in other words, its external validity (Kroodsma 1989a, b; Searcy 1989). Suppose an experiment presents one exemplar of a song pattern to one set of subjects and an exemplar of another song pattern to a second set of subjects (with suitable attention to randomization of presentations and minimization of biological interactions between subjects). Such an experiment in itself raises no problems concerning the independence of observations. If a significant difference is found in responses to the two exemplars, one can safely conclude (assuming all other issues are resolved) that the two exemplars evoke different responses.

Nevertheless, a broader conclusion, that a class of stimuli represented by the first exemplar evokes a different response than another class represented by the second exemplar, has weak justification. If fertilizer from one source produces greater growth of crops, do fertilizers from other sources have the same effect?

How can we achieve greater external validity for an experiment? Clearly, using more exemplars of each class of stimulus is a step in the right direction. A complete solution is not so simple, however. The exemplars should represent the possibilities within each class. To do so requires (1) a strict definition of the class of stimuli and (2) an adequate sample of that class. Random samples, especially when small, are not necessarily representative. Small samples have high between-sample variance. Five exemplars, as in the proposed nested ANOVA (see Table 2 in Kroodsma et al. 2001), is a small sample for these purposes.

If the class of possible stimuli is strictly defined, an experimenter could consider stratified random sampling of the variation within the class, as an alternative to completely random sampling. For example, the total set of song patterns in one class might be divided into two or more levels for some variable identified by the experimenter; then, one or preferably two exemplars could be selected randomly at each level. Artificial synthesis of stimuli, rather than recording and presentation of natural stimuli, would also help to assure clear definitions of the classes of stimuli.

It might be objected that the experimenter, not the subject, has defined the classes of stimuli. It is the subjects' classification that we are interested in, of course. But, the point of an experiment is to try one possible classification. The experimenter selects the classification for testing; the subjects then indicate whether they also differentiate these classes. Further experiments can specify in progressive detail how the experimenter's classifications of stimuli map onto the subjects'. To succeed in this venture, experimenters must define their classes of stimuli clearly and choose exemplars that represent each class adequately. There is no prescription

for the number of exemplars needed for adequate representation of a class of stimuli, so again compromise is inevitable.

Single or Multiple Use of Exemplars and Subjects

Should each exemplar be presented more than once? Following the discussion by Hurlbert (1984) of 'sacrificial pseudoreplication', Kroodsma et al. (2001) criticize pooling the responses to different exemplars (subclasses of stimuli) within a class, even when a statistical test shows no differences in these responses. They rightly argue that accepting a null hypothesis, especially with small sample sizes, is problematic. An alternative to pooling, however, is a two-level nested ANOVA, which can correctly compare within- and between-class differences in responses.

Suppose we identify two classes of songs, then choose in some appropriate way (by stratified random sampling, for example) two exemplars (subclasses) of each class, and present each exemplar to three subjects each (12 subjects receive one presentation each). A test for a difference in mean responses to the two exemplars within each class would have $df=2(3-1)=4$. If we subsequently decided to pool responses to the two exemplars within each class, a test for a difference in mean responses to the two classes would have $df=2(6-1)=10$. Thus we would incorrectly use a less powerful test for the difference between subclasses within each class than we do for the difference between classes.

A two-level nested ANOVA, in contrast, correctly compares within-subclass, within-class and between-class variances (in this example, $df=8$, 2 and 1, respectively, on the assumption of equal variances across all subclasses). If we wish to perform a similar analysis, but without the assumptions of a parametric ANOVA, we can use an alternative strategy. Instead of testing for differences between exemplars within each class separately, we can combine the probabilities of such tests across both classes. In other words, we use all of the data to compute a single probability for overall differences between exemplars (Sokal & Rohlf 1995, Chapter 17). We would thus increase the power of the test for these differences in relation to that for differences between classes. A nonparametric ANOVA is another way to accomplish this objective.

A reason to use each exemplar more than once is to determine whether or not exemplars thought to be similar by the experimenter are also perceived to be similar by the subjects. If each exemplar is presented only once, the experimenter can never evaluate whether the exemplars chosen for each class of stimuli are heterogeneous or not. In this case, the experimenter must assume that multiple exemplars represent a single class of stimulus. To test for possible heterogeneity in exemplars within a class, rather than to assume its absence, an experimenter must present each exemplar more than once.

Even if each exemplar is used only once, an experimenter might choose to group them into subclasses. For example, in the experiment above, if we had chosen 12 different exemplars (6 in each of two classes) and presented each one once, we might have grouped them into subclasses of two or three exemplars each to allow an evaluation of possible heterogeneity within classes.

The same justification could be made for using each subject more than once. Multiple presentations to each subject allow tests for heterogeneity among subjects in a two-level nested ANOVA or, even better, paired statistical tests. When the examples of 'sacrificial pseudoreplication' cited by Kroodsma et al. (2001) are examined, at least some of them seem to have used appropriate two-level nested or repeated measures ANOVA for designs in which subjects or exemplars were used more than once.

In such studies, a decision to include tests for heterogeneity among exemplars or subjects requires a compromise. If an experiment uses each subject and each exemplar more than once, the sample size for the main test decreases. On the other hand, such an experiment avoids complete blindness to unsuspected heterogeneity among subjects or exemplars. Any experiment has a finite sample of subjects and thus must face this compromise.

Sample Size and Unsuspected Bias

So far, we might conclude that a clear behavioural experiment should (1) satisfy all of the usual conditions for randomization and interspersion of treatments and, in addition, (2) determine (or at least justify) that biological interactions between subjects do not affect their responses, (3) clearly define the classes of treatments and include exemplars that adequately represent variation within each class, and if possible (4) include replicate presentations of each exemplar. This minimal experiment involves several exemplars of each of several classes of stimuli, with each exemplar presented to several subjects. Often this level of complexity is as much as an experiment in the field can manage.

The number of trials (the sample size of an experiment) of course affects the power of the associated statistical test to detect a significant difference between treatments. In this sense, a big experiment is better than a small one. Furthermore, there seems to be no limit to the size of an 'ideal' experiment. The larger the sample size, the smaller the possible difference in mean responses it is possible to detect (provided variation in responses remains constant).

Sample size, however, has a rarely noticed influence on the possibility for erroneous conclusions. Just as small samples only permit detection of relatively large differences in the responses to two treatments, they also reveal only relatively large systematic biases in treatments. Large samples permit detection of smaller differences between treatments but can also reveal correspondingly small systematic biases.

Every experiment is performed by humans, or by apparatus constructed by humans, who are notoriously prone to error and influenced by expectations. Randomizing and interspersing treatments are the ways we can reduce these unwanted influences on responses. Thus, when we suspect a possible bias, we can distribute its influence evenly across treatments. Randomization can minimize

other biases. Yet interspersion and randomization as applied by real humans to real situations, as opposed to abstractly defined situations, always involves assigning treatments or choosing exemplars by particular features. We randomize locations, orders, individuals and song patterns, but we cannot in the real world randomize in general. Consequently, unsuspected biases cannot be eliminated in an absolute sense with any actual experimental design.

The problem with large experiments is thus insidious: the possibilities for finding small unanticipated biases increase, for the same reason that the possibilities for finding small differences between treatments increase. An experiment with a large sample runs the risk of small but statistically significant systematic biases. Because these small effects are difficult to identify, either by the experimenter or by a reader, large experiments that report statistically significant, but small, differences always require special attention. An experiment with a relatively small sample might also include systematic bias. In this case, however, the bias would have to be relatively large to reach statistical significance and thus is more easily identified.

Choice of sample size for an experiment thus requires compromises. For experiments in highly controlled environments with minimal intervention of humans, larger samples are plausible. When the possibilities of inadvertent bias and extrinsic variation are greater, for instance, in experiments in the field with natural stimuli, smaller samples are optimal. In the latter situation, searching for small effects with large samples cannot be so easily justified, because of the difficulty of identifying small systematic biases. Field experiments, in general, should seek large effects with simple designs. Otherwise, a report of the experiment should devote special attention to the possibilities of small unsuspected biases.

Ideal Reporting Instead of Ideal Experiments

In conclusion, because every experiment involves several kinds of compromises, there is no single ideal experiment. There might well be, however, an ideal form for reporting experiments. A report of any behavioural experiment should go beyond the traditional explication of the experimental design and the associated statistical procedures. It should also (1) describe reasons for rejecting the possibility of biological interactions between subjects, (2) explicitly define the classes of stimuli compared, (3) describe procedures for adequate sampling of the variation within each class of exemplars, (4) specify features used in randomizing treatments, and (5) justify the sample size in relation to the experimental circumstances.

I thank members of the Triangle Behavior Seminar, the Animal Behavior group at Chapel Hill, and several refreshingly frank colleagues, especially Steve Nowicki and Bill Searcy, for much advice.

References

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

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

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

Kroodsma, D. E. 1990. 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., Byers, B. E., Goodale, E., Johnson, S. & Liu, W. 2001. Pseudoreplication in playback experiments, revisited a decade later. *Animal Behaviour*, **67,** 1029–1033.

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

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

Sokal, R. R. & Rohlf, F. J. 1995. *Biometry*. 3rd edn. New York: W. H. Freeman.

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