pp 1-10 in P. McGregor (ed) Playback studies in animal communication. Plenum Press, New York & London (1992)

DESIGN OF PLAYBACK EXPERIMENTS:

THE THORNBRIDGE HALL NATO ARW CONSENSUS.

Peter K. McGregor, Clive K. Catchpole, Torben Dabelsteen, J. Bruce Falls, Leonida Fusani, H. Carl Gerhardt, Francis Gilbert, Andrew G. Horn, Georg M. Klump, Donald E. Kroodsma, Marcel M. Lambrechts, Karen E. McComb, Douglas A. Nelson, Irene M. Pepperberg, Laurene Ratcliffe, William A. Searcy, Daniel M. Weary [1]

Introduction to the Issues

Playback is an experimental technique commonly used to investigate the significance of signals in animal communication systems. It involves replaying recordings of naturally occurring or synthesised signals to animals and noting their response. Playback has made a major contribution to our understanding of animal communication, but like any other technique, it has its limitations and constraints.

This section of the workshop was intended to address two different issues. The first concerned the design of playback experiments and the analysis of the subsequent responses. The second issue was the range and type of practical pitfalls involved in actually carrying out playback experiments.

The First Issue

A paper by Hurlbert (1984) on the design of ecological field experiments stimulated an examination of the design of playback experiments (Kroodsma 1986; 1989a). Kroodsma suggested that the design (and analysis) of many playback experiments was inappropriate for the questions being investigated. The suggestion triggered a lively debate about such issues in the literature (Searcy 1989; Catchpole 1989; Kroodsma 1989b, 1990a, 1990b; Weary and Mountjoy in press). One of the purposes of this workshop was to bring together practitioners of playback with interests in diverse topics and animal groups in an effort to reach a consensus on this controversial area. The first section of this chapter attempts to identify clearly the nature of the problems of playback design and analysis that are at the root of the controversy, and then to assess the implications for playback experiments and make recommendations for future work.

[1] The details of authors' affiliations are given in the list of workshop participants.

The Second Issue

Most experimenters with experience of playback have a list of factors that they consider to be important in a well-executed playback study. By incorporating these factors into their design, experimenters try to ensure that the experiment presents the animals with stimuli that differ only (or at least mainly) in the signal feature of interest. Although journals are often reluctant to print such details in the methods section of papers, the information is needed by experimenters trying to replicate studies. Our workshop, which hosted researchers with expertise on different taxonomic groups and areas of interest, presented an obvious opportunity to collate a list of factors considered to be important in running playback experiments. Although specific questions and specific animal groups will require additional factors to be considered, the list presented at the end of the chapter (Table I) is a starting point of general features to which more specific factors can be added.

Appropriate Design and Analysis - the Pseudoreplication Debate

The issues of appropriate design and analysis have come to be referred to as *the pseudoreplication problem* in the literature, following Hurlbert's (1984) title. This section aims to explain what is meant by this expression, and how to avoid the problem, and how it relates to external validity and the limits on generalisation.

What is Pseudoreplication?

Hurlbert (1984) defines *pseudoreplication* 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." Hurlbert was mainly concerned with cases from field ecology, in which, for example, only one control field and only one experimental field would be compared statistically by using sub-samples drawn from each field, or in which experimental plots were spatially segregated from control plots. Pseudoreplication, however, is a problem in a great many areas of science.

In bioacoustics, the term has been applied most frequently to cases in which some general hypothesis is stated about response to general classes of stimuli, and the hypothesis is tested using insufficient numbers of exemplars from each class (Kroodsma 1989). The problem with the latter test is that the stimuli almost certainly vary within each class as well as between classes, so that any difference in response cannot necessarily be ascribed to the between-class difference in stimuli. An example of such a problem would be playing a number of birds one song from their own dialect and one song from a distant dialect. The two test tapes will vary in a number of features, only one of which is the feature of interest, that is, the signal structure that distinguishes own from distant dialect.

As there has been some debate over exactly how Hurlbert's (1984) definition applies to playback experiments, we wish to present our own definition, which we believe is clearer. We define pseudoreplication as the use of an *n* (sample size) in a statistical test that is not appropriate to the hypothesis being tested. Thus whether pseudoreplication can be said to occur in a given experiment depends on the hypothesis that is stated as being tested. Some hypotheses will dictate that we sample a sufficient number of stimuli from a particular class of stimuli, some that we sample a sufficient number of animals from a population of animals, some that we sample a sufficient number of groups of animals, and so on.

The application of this definition can be made clearer with a specific example. We will first state this example in as simple a manner as possible, shorn of statistical terminology. Next, we restate the example using the language of analysis of variance (ANOVA), which we have found makes the example more understandable for some and less understandable for others. The example is based on the phenomenon of bird song dialects because here little specific background is needed to grasp the questions addressed by playback.

A Specific Example The question of interest is the difference in response shown by birds to playback of different dialects. If the hypothesis is framed very narrowly, for example that birds of dialect X respond differently to song X_1 of their own dialect than to song Z_1 of a specified foreign dialect (Z), then it can be admissible to use only single exemplars of the two dialects, in this case X_1 and Z_1. If the hypothesis is stated more broadly, i.e. that birds of dialect X respond differently to own dialect (X) than to a specified foreign dialect (Z), then using only two exemplars, one from X (X_1) and one from Z (Z_1), and using the number of subjects as the *n* in a statistical test, would be to pseudoreplicate. To avoid pseudoreplication in this case, one would have to use a sample of songs from each dialect (X_1 X_2 X_3 X_4 X_5 etc. and Z_1 Z_2 Z_3 Z_4 Z_5 etc.), using enough songs to be sure that a statistical test could be done with the number of songs as the sample size. If the hypothesis is stated even more broadly, for example that birds of dialect X respond differently to own dialect than to foreign dialects in general, then songs from several foreign dialects (U, V, W Z etc.) must be played to avoid pseudoreplication.

The Example Restated in ANOVA Terminology Suppose that the hypothesis is that response to songs of their own dialect (X) is different from response to songs of a specified foreign dialect (Z). The easiest way to visualise the design is as a diagram (Fig. 1):

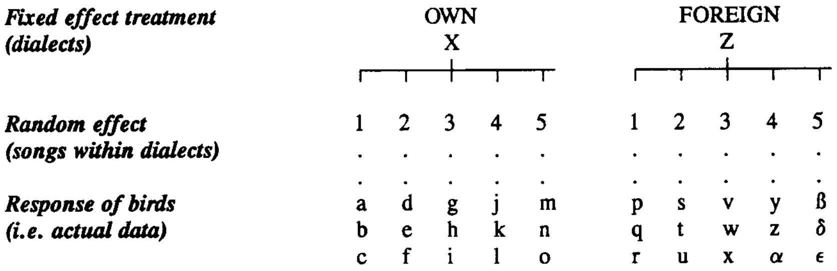

Figure 1. A diagrammatic representation of an ANOVA design to test the hypothesis that response to own song dialect (X) is different from a specified foreign dialect (Z). The figure shows five male birds' renditions of own dialect (X_1 to X_5) and five of foreign dialect (Z_1 to Z_5). Each letter (a to ϵ) indicates the data (such as approach, amount of song, etc.) collected from a single male subject.

The design is a two-level mixed-model nested ANOVA. The treatments (own *v.* foreign dialect, i.e X *v.* Z) represent fixed effects since they are determined by the experimenter and are repeatable. However, while dialects are fixed, songs within dialects (X_1 to X_5 and Z_1 to Z_5) vary unpredictably between male birds (for example there may be individual differences in rendition of the dialect) and possibly within males also (for example there may be song by song differences in the male's renditions, i.e. male 1's version of X (X_1) may vary X_{1i} X_{1ii} X_{1iii} etc.). Thus a mixed-model nested ANOVA is appropriate, since this design allows for variation in response to randomly different songs within fixed treatments. If a single exemplar for each dialect were to be used, we could never be certain that the observed differences in response were really due to the different dialects: it is possible that uncontrolled factors might cause differences in response to songs even if they had come from the same dialect. The only way to separate uncontrolled from dialect effects is to replicate songs within dialects. Using *responses* (a to c, and p to r if the single exemplars were X_1 and Z_1 respectively) as replicates for *dialect treatments* (X *v.* Z), rather than *songs within dialects* (X_1 to X_5), is to pseudoreplicate.

In this experimental design, nothing is really gained by having replicate responses to any given song (X_3, say), except that it provides a better estimate of the average response to that song. The vital component is the replication of songs within dialects, which allows a test of the hypothesis that the average response to own dialect is different from the average response to a foreign dialect.

Sokal and Rohlf (1981, Table 10.2) run through the calculation of an exactly similar design: the only decision to make is whether to pool the within-group and songs-within-dialect mean squares before testing the significance of the between-dialect mean square. Sokal and Rohlf give the criteria for making this decision.

If own and foreign dialects are both played to an individual bird (i.e. response of birds is a-o for X and a-o for Z in Fig. 1), the design is different. If the hypothesis is that birds respond differently to own dialect than to a foreign dialect, then formally the test is that the average of (own *minus* foreign) response $< >$ 0, and can be tested with a t-test.

While there may be only one level for any given explicitly stated question or hypothesis where replication is mandatory, it may be desirable to estimate variance in lower level variables using a model II nested ANOVA. These estimates may be interesting in their own right and could help to explain a lack of a treatment effect. For example, no difference in response to dialects may be due to variation among exemplars within a dialect and/or variation among responses of individual subjects.

Avoiding Pseudoreplication

The two key features in avoiding problems of pseudoreplication are being explicit about the question being addressed by the playback experiment and deciding on the number of exemplars.

How Many Exemplars are Adequate? There is no simple answer to this question. Indeed, statistics texts will say that no answer is possible at all unless there is an estimate of the variability of the items of interest (Sokal and Rohlf 1981, p.262).

In the past it has been argued that the variability of the signal gives an indication of the extent of replication necessary. For example, if the signal appears to an observer to be stereotyped then fewer replicates would be required to represent the variation adequately than if a signal were more variable. There are two problems with this approach. First, there is no good *a priori* reason for the variation that is apparent to the human visual system when inspecting sound spectrograms or oscillograms (the same is true for measures taken from spectrograms) to be the same as the variation perceived by the study animal. Second, the feature of interest is the variation in *response* to the signal (playback); such variation may not be directly related to the variation perceived in the signal by both humans and the study animal (see chapters by Weary and by Ratcliffe and Weisman in this volume). There are three possible reasons for this difference: first, the animal may not perceive the variation (e.g. lack of perceptual ability, artifact of presentation); second, the animal may perceive the variation but it elicits no difference in response because this would be biologically inappropriate (e.g. there are competing behaviours such as mate guarding and feeding); finally, the measures of response taken by the experimenter may simply be too crude to show a difference between stimuli.

Workshop participants suggested a solution to the problem of determining the appropriate number of replicates; that is, for a two-stage approach. The first stage is to use previous experience with, and other work on, the study species to make an informed guess at the level of replication necessary and to use this level to examine the question. The second stage is to use the information from the first stage to refine the number of replicates needed. This procedure may involve formal measurement of the variation of response, possibly with principal components analysis to reduce complex features of the original signals to a manageable number, and the application of the standard formulae available to estimate the sample size necessary to show a difference of the required magnitude and probability level (e.g. Sokal and Rohlf 1981, Box 9.13; see also the section on Bayesian statistics in the chapter by Gerhardt in this volume). Once again the discussion during the workshop emphasised the importance of being clear about the question that playback was being designed to answer, as this will obviously affect the level of replication necessary.

Can Use of Synthetic Calls Help to Avoid Pseudoreplication?

Pseudoreplication typically arises in a playback experiment because of lack of control over the differences in our treatments. For example, in the dialect case, if we take song A_1 from one dialect and song B_1 from another, they must differ in dialect features, but they may also differ in all sorts of other features, e.g. motivation of the male when singing, quality of recording, etc. Our only hope of controlling these other differences is to use several examples of each dialect, so that those differences will average out. In contrast, if we use artificially modified stimuli we have better control over the differences between our stimuli. For example, if the experiment tests response to one natural song versus exactly the same song in manipulated form, then a statistical test can be done with *n* as the number of subjects to test the hypothesis that this manipulation of this particular song affects response. It may not be clear to what aspect of the changed stimulus the subjects are attending - for example if the manipulation is to halve a song, the subjects could be attending to the changed duration or the missing acoustic elements etc. - but still it is clear that the manipulation is responsible for the difference in response.

The proper use of synthetic sounds avoids many of the design problems arising from the multidimensional nature of natural signals and variations in recording quality. In principle, an investigator can explore the behavioural relevance of the entire perceptual or preference space that is delimited by variation in a set or sets of acoustic signals of particular interest: within-population, between-population (dialects), between different signals in the repertoire, and between species. There will be considerable practical difficulties of studying systems in which there are very many acoustic properties of potential significance, but we are hopeful that in many of these the set of relevant properties will be some relatively small subset of the possible properties. In fact, a few studies have begun to tackle the problem of varying two properties of known pertinence at the same time (Nelson 1988; Date et al. 1991; Gerhardt and Doherty 1988; Dooling 1982). As the number of simultaneously varying properties being examined increases, the design, execution and interpretation of such experiments will probably warrant another workshop like this one.

A first step is to generate a synthetic signal that is comparable in its behavioural effectiveness to a typical natural signal. The signal can be synthesised *de novo* or produced by modifying a naturally occurring sound. Normally, such a signal would have properties with values equal to the estimated mean values in the set of signals of interest. Ideally, in comparative tests, the synthetic standard call is neither more, nor less, attractive than a series of natural exemplars.

The second step is to develop criteria for choosing the amount of change in the value of a given parameter. In our view any of the three following criteria are appropriate, depending on the question being asked:

1) Change the value of a parameter by units equal to the standard deviation (or some other measure of variance) of the property in the natural set of interest. This procedure would be appropriate for estimating the proportion of signallers in a population that might be favoured by mate choice based on the property in question;

2) Vary the value of the parameter by some constant percentage, guided perhaps by any existing psychoacoustical data;

3) Third, if the distributions of a parameter in two classes of natural signals do not overlap, choose a difference between the property that corresponds to the minimum observed difference between the two sets (species, populations, dialects, signals within a repertoire). If the animals discriminate, then the difference in that property at least is adequate for recognition of the two natural classes of signals. If the animals do not respond selectively, then the difference can be increased systematically until there is a differential response. This information could then be related to the natural variation in the two sets of signals to provide an estimate of the proportion of individuals that would be likely to be distinguished in natural situations.

External Validity

External validity refers to the degree to which we can generalise from the results of a specific experiment. Limited external validity is a problem for all fields of science. An experiment can avoid pseudoreplication and still have limited external validity. In other words, no matter how good the design and execution of our experiments, there is always a logical danger in generalising from our results.

Although there is no theoretical reason for internal and external validity to be linked, practical field constraints may mean that adequate controls for the many features of the stimulus signal and its presentation will limit the range of the question posed and therefore the extent of external validity. For example, the logistics of carrying out experiments at two widely separated field sites and the constraints to carry out the playback experiments at the same time of day, breeding season or year may restrict the experiments to one field site, which will in turn limit the external validity.

Generalising

If all practical playback experiments are limited in their external validity, to what extent is it possible and desirable to generalise from these experiments? An extreme stance is that comments in the discussion section of a paper or manuscript should be restricted to the specific effect found, for example, "neighbour/stranger discrimination was shown for 16 first year males in the northeastern corner of the study site." It seems more reasonable to view each specific experiment as a step on the road to more general explanations and unless there is something unusual about the data set that could affect neighbour/stranger discrimination, such as that particular corner of the study population is the only one where males are territorial as first year animals, then the discussion would seem the obvious place for a consideration of the advantages of neighbour/stranger discrimination in general. Implicit in this approach is that results from the first experiment form the basis for a general hypothesis that will be tested in subsequent experiments. The hypothesis is only modified when the predictions are not supported by subsequent experiments; then a new hypothesis consistent with the literature is proposed for subsequent testing. Provided that the language in the discussion makes it quite clear what is being proposed as an hypothesis derived from the experiment as opposed to a result, then the distinction between test and idea will be maintained.

A variety of simple and complex problems can be examined using playback. Testing broad hypotheses and replicating treatments at a correspondingly broad level is certainly one valuable method for making progress in science. However, we also recognise the value of testing narrower hypotheses, with replication at a lower level (e.g. at the level of subjects rather than stimuli). Through a series of simple experiments, one can then approach the larger question. Both approaches are valuable and both demand explicit statement of hypotheses and appropriate replication.

Conducting Playback - Important Features to Consider

The recognition that experimental execution is critical to any investigation is the second issue addressed by this chapter. This point is also stressed by the paper that triggered the pseudoreplication debate (Hurlbert 1984). To quote Hurlbert (1984, p.189): "Yet in a practical sense, execution is a more critical aspect of experimentation than is design." The reasoning underlying his statement is that errors of execution are more common, more variable, more subtle and more difficult to detect at all stages of an experiment than design errors.

We have compiled a list of features (Table I) that should be considered by any playback experimenter in order that execution errors may be minimised or at least recognised. This is not a list of recommended procedures; neither is it an exhaustive list. Rather,it is a list of factors that can be important in the execution of playback experiments and whose importance must be judged by the experimenter in the context of his particular experimental situation. The methods section of any publication resulting from a playback experiment ought to state which of these features were judged to be important by the experimenter and how their effects were controlled.

Summary

Pseudoreplication as it applies to playback studies is a consequence of a lack of rigour in specifying the question being addressed by the study. Commonly a lack of replicates at the level of interest results in an inability to answer the question as posed. However, pseudoreplication is not ubiquitous in playback studies, nor does it impose a constraint on their usefulness. Pseudoreplication is a potential problem not only in playback, but in all areas of research. Remedying the problem is straightforward. Care has to be taken when specifying the hypothesis to be tested and an appropriate number of replicates must be used when conducting the experiment.

Equal care must be taken to appreciate the factors that can influence an animal's response to playback. Some important sources of execution error are identified in Table I, but we urge playback experimenters to include information on how execution errors were minimised in the methods section of resulting publications.

There is no reason to doubt that playback, with the continuing development of new techniques and experimental designs, will remain one of the most powerful tools available for the investigation of animal communication.

Table I. A list of some of the features affecting execution errors in playback experiments. SPL = sound pressure level, s/n = signal to noise.

Test tapes and test sounds
Sound per unit time, total amount of sound.
Degradation (distortion), SPL, s/n ratio of source sounds.
Encoded information on status, motivation, identity etc.
Level and type of background noise.
Filtering and editing to remove background noise.
Environmental conditions
Time of year (influences background noise, vegetation, activity of subjects and other species).
Weather conditions (same influences as time of year).
Time of day (degradation effects, see also time of year).
Test Animals
Subject location in relation to territory boundaries.
Effects of stage of breeding cycle.
Time of day effects.
Proximity of resources (mates, food etc.).
Activity of conspecifics (neighbours, intruders etc.).
Predator activity.
Playback equipment
Speaker directionality.
Fidelity of equipment (s/n ratio, frequency range, etc.).
Procedure
Position and behaviour of observer.
Use of blind experiments (observer bias likely?).
Loudspeaker position.
Information on failed tests.
Response measures (single, multiple).

References

Catchpole, C.K. 1989. Pseudoreplication and external validity: playback experiments in avian bioacoustics. *Trends in Ecology & Evolution*, **4**, 286-287.

Date, E.M., Lemon, E.R., Weary, D.M. and Richter, A.K. 1991. Species identity by birdsong: discrete or additive information. *Anim. Behav.*, **41**, 111-120.

Dooling, R.J. 1982. Auditory perception in birds. In: *Evolution and Ecology of Acoustic Communication in Birds. Vol.II.* (Ed. by D.E. Kroodsma, E.H. Miller & H. Ouellet), pp. 95-130. Academic Press, New York.

Kroodsma, D.E. 1986. Design of song playback experiments. *Auk*, **103**, 640-642.

Kroodsma, D.E. 1989a. Suggested experimental designs for song playbacks. *Anim. Behav.*, **37**, 600-609.

Kroodsma, D.E. 1989b. Inappropriate experimental designs impede progress in bioacoustic research: a reply. *Anim. Behav.*, **38**, 717-719.

Kroodsma, D.E. 1990a. Using appropriate experimental designs for intended hypotheses in 'song' playbacks, with examples for testing effects of song repertoire size. *Anim. Behav.*, **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 Behaviour. Vol. II.* (Ed. by M. Bekoff & D. Jamieson), pp. 226-245. Westview Press, Boulder, Colorado.

Gerhardt, H.C. and Doherty, J.A. 1988. Acoustic communication in the gray treefrog, *Hyla versicolor*: evolutionary and neurobiological implications. *J. comp. Physiol. A.*, **162**, 261-278.

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

Nelson, D.A. 1988. Feature weighting in species song recognition by the field sparrow (*Spizella pusilla*). *Behaviour*, **106**, 158-181.

Searcy, W.A. 1989. Pseudoreplication, external validity and the design of playback experiments. *Anim. Behav.*, **38**, 715-717.

Sokal, R.R. and Rohlf, F.J. 1981. *Biometry*. 2nd Edition. W.H. Freeman & Co., New York.

Weary, D.M. and Mountjoy, *in press*. On designs for testing the effect of song repertoire size. *Anim. Behav.*,