Design of Song Playback Experiments

DONALD E. KROODSMA[1]

A recent Workshop on Experimental Design led by Alan Kamil at the 1985 A.O.U. meeting in Tempe, Arizona, concluded not only that more laboratory and field experiments are necessary, but that experimental design is an extremely important feature of this work. These discussions on the characteristics of good designs for experiments rekindled my concerns for designs that are often used in studying the development and function of bird song. To illustrate my concerns, I here present a composite design for a playback experiment based on designs from several recent publications. I then discuss what I believe to be weaknesses in the design, and I offer possible solutions to those problems.

The biological question I propose to test in my hypothetical experimental design is drawn from my work with the Blue-winged Warbler (*Vermivora pinus*). In these warblers, each male has two song types. Analysis of spectrograms indicated that the Type I song, the "bee-bzzzz," is highly stereotyped throughout the geographic range of the species, but that the Type II song varies microgeographically (Kroodsma 1981). My playback experiment therefore is designed to test whether Blue-winged Warbler males also discriminate between local and foreign examples (i.e. recognize "dialects") of the Type II but not the Type I songs.

I designed the experiment as follows. From a study population of 10 color-banded birds in Amherst, Massachusetts, I obtain one recording of each song type. I also make a high-quality recording of a Type I and a Type II song at the Rockefeller University Field Research Center in Millbrook, New York. I then make two playback tapes, one with Type I songs and one with Type II songs, for two-speaker playback experiments. Each two-track tape is designed so that songs of the same type from Amherst and Millbrook are played alternately from two speakers placed 20 m apart. During the playback, I call to an assistant the estimated location of the responding territorial male. After a 5-min playback, speaker cables are switched, and the playback is repeated. The median position of the responding males during each of the 5-min playbacks is used in a two-tailed, matched-pairs, signed-ranks test [see Lanyon (1978) for the rationale of this two-speaker playback design].

In early May, I begin playbacks with the Type II tape within the banded Amherst population, and I continue for three weeks until the data reveal a statistically significant difference ($n = 20$, $P = 0.05$). I next test the birds with the Type I playback tape and find that the same number of playbacks ($n = 20$) to the same birds reveals no significant difference in response ($P = 0.50$). From these data I conclude that Blue-winged Warblers recognize dialects in one song type (II) but not the other (I).

I believe, however, that there are several weaknesses in the design of this experiment. Moreover, these design flaws preclude the general conclusion that was made from the data. After the weaknesses are discussed (1–6 below), I suggest ways to eliminate each of these problems.

(1) Because *I* formulated the hypothesis, recorded the birds, made the playback tape, *and* called to the assistant the location of the experimental bird, the playbacks were not done "blindly." The data were therefore subject to potential, even though unintentional, biases. I know what results would to me be most exciting (and publishable and fundable), and when I *estimate* the location of the bird, the data are subject to subtle biases. To eliminate this potential bias, I could have my naive (i.e. "blind") assistant, who does not know the hypothesis I am testing, call out the location of the bird to me. Because Type II songs from Amherst and Millbrook are strikingly different from one another, I believe this is the only possible solution. With Type I songs a second approach is possible; songs from Millbrook and Amherst are not distinguishable to the human ear, and I could have remained blind to the identification of the songs on the tape if the assistant had withheld that information from me. Precautions such as these help to remove investigator bias and are highly desirable in all experimental designs.

(2) I used only one exemplar per song type for each location, and a general statement about these particular song dialects cannot be made unless each exemplar is representative of the entire population. Background sounds or degraded songs can reduce the potency of a signal. Other features, such as duration, might increase the potency of a song. In addition, males at each location have some familiarity with the local test songs. In the design I presented, 1 of the 10 males is tested with his own song, and several birds are tested with the song of a territorial neighbor. Such familiarity with songs can significantly bias responsiveness (Falls 1982), and perhaps the best solution is to use test songs from non-neighboring males from the vicinity of the playback subjects. All local subjects are then about equally unfamiliar with the test song.

Ideally, I believe that a different playback tape should be made for each playback (in this case, 20 tapes for Type II playbacks and 20 for Type I playbacks). At the very least, several different tapes (perhaps 4–5 with $n = 20$) should be used so that one

[1] Department of Zoology, University of Massachusetts, Amherst, Massachusetts 01003-0027 USA.

potentially atypical song does not dominate the results.

(3) Even if the songs are representative of their respective populations, there remains a serious related problem: to make the *general* conclusion about dialect discrimination, I believe that each playback should be from a different Type II dialect. There are several reasons for this precaution. First, Type II songs from either Millbrook or Amherst may be especially potent or weak songs. Thus, if Amherst birds respond more strongly to local than to Millbrook songs, it is possible that Amherst songs would be especially potent, or that Millbrook songs would be especially weak, to all males *throughout* the range of the species. Second, Type I songs may vary geographically, but on a somewhat larger spatial scale than the Type II songs. Or Type I songs could vary as much as Type II songs, but with a different geographic pattern. In selecting the test areas, I may by chance have crossed a Type II boundary but not a Type I boundary. I realize that visiting a different dialect area for each playback is impractical, but my design should at the very least include playbacks in Millbrook to demonstrate that those birds also respond more strongly to their local Type II songs.

(4) Because I did 20 playbacks to 10 birds in each half of my experiment, each of the 20 sample points used in the statistical test was not an independent replicate. Only 10 birds were tested, and the sample size is therefore 10, not 20. The only way to increase the sample size is to use more birds.

One method often used in finding additional birds is to listen for a singing male. Yet, singing birds are more likely to be unpaired, and perhaps even to be first-year birds, than are nonsinging males. Unpaired males, then, are not representative of the entire population. In the Blue-winged Warbler, Type II songs seem to be used more during male-male interactions, but Type I songs predominate when a male is unpaired. The use of these two songs changes during the season. Mating status therefore may be an important variable that should be controlled rather than ignored, or playback subjects should at least be chosen by a method that yields an adequate cross section of the population.

(5) Males use the two song types in different circumstances, and use of the songs and response to them may shift during the breeding season. More importantly, because I did playbacks with Type II songs in May and with Type I songs during late May and early June, either (a) a declining motivation or interest in discriminating songs through the season or (b) habituation to the playbacks could account for my results. To correct this design flaw, I should do playbacks with both song types during the same portion of the breeding season. Using different birds for each playback, I might alternate Type I and Type II songs each morning throughout the season and actually search for seasonal differences in song discrimination. Alternatively, I might present Type I and Type II playbacks to the same male on successive mornings. I would control for the sequence of exposure to the two playbacks by exposing half the males to Type I and the other half to Type II playbacks first. Thus, the time of the breeding season is related to mating status and it, too, should be controlled, not ignored, as a secondary variable.

(6) My last concern is again with the sample size. Sample sizes should be determined at the onset or by some (valid) statistical approach in the early stages of the experiment (e.g. Sokal and Rohlf 1969, James and McCulloch 1985). Occasionally testing for significance (i.e. $P = 0.05$) and doing just enough playbacks to obtain significance is a misuse of statistical testing procedures.

Those who have done playback experiments will be sensitive to other issues as well. The use of a two-speaker as opposed to a one-speaker playback experiment could be debated. Interpreting whether flight from a speaker is a strong or weak response may not be a straightforward matter. A playback to one male also may affect his neighbor, so that tests on adjacent males probably should not be done on the same day. In addition, placement of the speakers, distance between speakers, duration of song stimuli, sequence of song stimuli (i.e. which of the two songs plays first), rates of song delivery, amplitude levels, concealment of the observer, weather, density of conspecifics, and other extraneous variables will all be considered by the careful experimenter. These factors are certainly important, even though I have not chosen to stress them here.

These precautions do require extra investment of money and effort. I must now (1) make sure I am accompanied by a blind observer, (2) construct extra playback tapes, (3) test birds at additional locations, and (4) find additional experimental birds. Being sensitive to (5) the sequence of playback experiments and (6) proper ways to establish sample sizes requires forethought. To ignore these precautions may be convenient, but this convenience will also compromise the strength of the conclusions that can be drawn from the data.

These same precautions are also pertinent to experiments in vocal ontogeny and function in the laboratory. For example, blind observers should compare spectrograms of pupils with those of tutors, and they also should rate the "copulation solicitation" displays of female songbirds responding to different song variants. Balanced experimental designs (points 3 and 5), independent replicates (4), appropriately determined sample sizes (6), and care to ensure that the specific stimuli are representative of the general population (2) are fundamental factors in robust experimental designs.

I thank Alan Kamil for an inspiring Workshop; R. Tod Highsmith, Edward H. Miller, Cynthia A. Staicer, and David A. Spector for constructive comments;

and the National Science Foundation (BNS-8506996) for research support.

Literature Cited

Falls, J. B. 1982. Individual recognition by sound in birds. Pp. 237–278 *in* Acoustic communication in birds, vol. 2 (D. E. Kroodsma and E. H. Miller, Eds.). New York, Academic Press.

James, F. C., & C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Current Ornithol. 2: 1–63.

Kroodsma, D. E. 1981. Geographical variation and functions of song types in warblers (Parulidae). Auk 98: 743–751.

Lanyon, W. E. 1978. Revision of *Myiarchus* flycatchers of South America. Bull. Amer. Mus. Nat. Hist. 161: 429–627.

Sokal, R. R., & F. J. Rohlf. 1969. Biometry. San Francisco, W. H. Freeman and Co.

Received 30 December 1985, accepted 17 January 1986.