acta ethol (2000) 3:3–8

REVIEW

Peter K. McGregor

Playback experiments: design and analysis

Received: 23 September 1999 / Received in revised form: 24 February 2000 / Accepted: 25 February 2000

Abstract The scientific value of the outcome of an experiment is closely related to its design and analysis. This article deals with the design issues of pseudoreplication (whether the experimental design has the statistical features needed to answer the question as posed) and execution errors (problems arising from how the experiment was conducted). Three issues of analysis are also dealt with: the number and type of response measures to record; how measures should, and should not, be combined into a single response measure; and how to interpret an apparent lack of response. Interactive playback is considered separately because it raises its own specific design and analysis issues. Although the examples generally refer to video playback, these issues are common to all experiments in behaviour.

Key words Pseudoreplication · Experimental error · Response measures

Introduction

An experimental approach to scientific understanding relies critically on the design and analysis of experiments. Such a statement is so self-evidently true that perhaps it is worth emphasising that all experiments are, to a greater or lesser extent, limited in their explanatory power by aspects of their design and analysis. Playback is an experimental technique in which natural or synthetic signals are broadcast and the response of animals noted. Playback experiments in any sensory modality are simply a subset of all possible experiments; therefore the design and analysis considerations discussed in this article are most definitely *not* specific to video playback. Such considerations also apply to various alternative approaches to investigating visual stimuli (e.g. live stimuli, dummies, still photographic images) as well as to many other sorts of experiment. However, since the Lisbon Video Playback Workshop dealt with video playback, I shall use examples drawn from this method to explain general design and analysis considerations. In much the same way that the consensus paper in this issue (Oliveira et al. 2000) is not an exhaustive list of features to be considered, neither is this article a recipe for that unattainable goal – the perfect experiment. Rather it should be regarded as a starting point. It is based on general discussions of experimental design and analysis in biology (e.g. Sokal and Rolf 1981; Barnard et al. 1993) as well as more specific issues related to behaviour (Martin and Bateson 1993; Milinski 1997). It has been influenced strongly by experience with playback of acoustic stimuli.

Many biologists find experimental design and analysis to be considerably less exciting topics than the various stimulus representations and manipulations that can be performed on a personal computer. This does not alter the fact that experimental design and analysis cannot be ignored.

The starting point for this article is that adequate video stimuli are available, as is knowledge of the appropriate environment and context in which to play them back. It should be obvious from the rest of the articles in this issue that it takes considerable thought and expertise to get to such a starting point. Just in case it is not, let me emphasise that I consider stimulus design, context, and delivery every bit as critical to a meaningful experiment as the issues I shall discuss.

Issues of experimental design

In this section I shall deal with two issues: (1) pseudoreplication, that is, whether the experimental design has the statistical features needed to answer the question as posed, and (2) experiment execution, that is, possible problems arising from how the playback experiment is

Communicated by R.F. Oliveira

P.K. McGregor (✉)
Department of Animal Behaviour, Zoological Institute,
University of Copenhagen, Tagensvej 16,
2200 Copenhagen N, Denmark
e-mail: pkmcgregor@zi.ku.dk

carried out. Contrary to most peoples' expectations, pseudoreplication is a far easier issue to identify and avoid (at the experimental design stage) than issues of execution. In my experience, however, most biologists find pseudoreplication the more difficult issue to understand.

Pseudoreplication

Pseudoreplication arises when there is confusion between the number of measurements made and the number of statistically independent replicates available for a statistical test. The result of such confusion is that the sample size (*n*) used in a statistical test is not appropriate to the hypothesis being tested. To illustrate with an obvious example, imagine we wished to see whether the height of men differed from that of women. We measured the height of one man six times and of one woman six times. We then performed a t-test using the number of measurements made, *n*=6. This is pseudoreplication, as the true number of statistically independent replicates for the test is one and not six, since there was only one man and one woman. No one would make this obvious mistake, but most cases of pseudoreplication are much more subtle than this.

A more realistic case, and one that is still surprisingly common in the literature of most experimental fields, can be illustrated with the following example. Imagine that we present each of ten subjects (male fiddler crabs) with the same two stimuli, balanced for the order of presentation over the course of the experiment. One stimulus is a video of a male crab waving and the other stimulus is the same male resting (i.e. not waving) (Fig. 1). We test if the subjects' response to one stimulus differs from the other with a paired t-test, *n*=10. Statistically this is a valid comparison; if we find a significant difference we have shown a difference in response elicited by the two stimuli. There is no pseudoreplication. The problem is that the question we wished to answer was probably not whether our two stimuli in particular elicited different responses, but the more general question of whether waving (in general) elicited a different response from resting (in general). The main difference we think there is between the stimuli we used (and presumably the reason for choosing them) is that one shows a waving male and the other shows a male resting. Therefore, it is natural to claim that we have shown that male crabs respond differently to waving compared with resting males. It may be natural, but it is also wrong. In statistical terms we have committed pseudoreplication. In the same way that our single man and woman give *n*=1 regardless of how many times we measure them, so too do our single examples of waving and resting, regardless of the fact we have ten subjects. This over-generalisation of the result is so natural that most experimenters do not realise that they have done it. However, a moment's thought makes it clear that to test whether waving elicits a different response from resting we need to use several

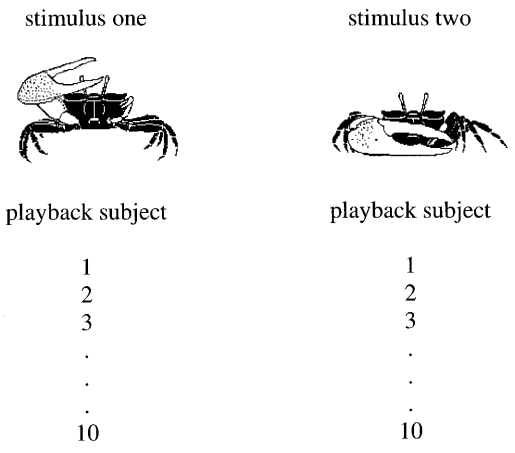

Fig. 1 Two stimuli were played back to each of ten subjects (male fiddler crabs 1–10). Stimulus one was a video of a male waving and stimulus two was a video of the same male resting. This is an example of an experimental design that does not exhibit pseudoreplication if the question is "does stimulus one elicit different responses from stimulus two?". However, such a design cannot answer the question of whether waving males elicit a different response from resting males because there are no replicates of waving and resting stimuli

randomly chosen exemplars of both waving and resting (e.g. Fig. 2). We may think that the only obvious difference between our two stimuli is the presence or absence of waving, but our subjects may see other differences that elicit a difference in response, such as a difference in overall brightness. The reason that we need multiple, randomly chosen exemplars is to try and ensure that the only difference in common between the two sets of stimuli is that one shows waving males (in all their variable forms) and the other shows resting males (in all *their* variable forms).

It is commonly thought that if the stimuli result from a manipulation (e.g. animating a crab) then there is no need to replicate stimuli. It is true that in most cases producing stimuli in such ways will reduce the likelihood of other response-eliciting differences being present, but it does not address the issue of pseudoreplication (because still *n*=1). It makes it easier to convince readers that our special pleading (that the only difference between the two stimuli is whether the male is waving or resting) is reasonable. But it is still special pleading, therefore suitable for the Discussion section, rather than a statistical result (suitable for the Results section). If the stimuli are sections of video recordings from the wild (sometimes called natural exemplars) then despite the best efforts of the video-recordist, there are likely to be several differences between stimuli, and the special pleading is weakened.

The examples discussed above should have made it clear that for any well-defined question it is straightforward to avoid pseudoreplication by including sufficient replicates at the experimental design stage to give an appropriate *n* for statistical testing. However, all experiments have limited generality (external validity in statisti-

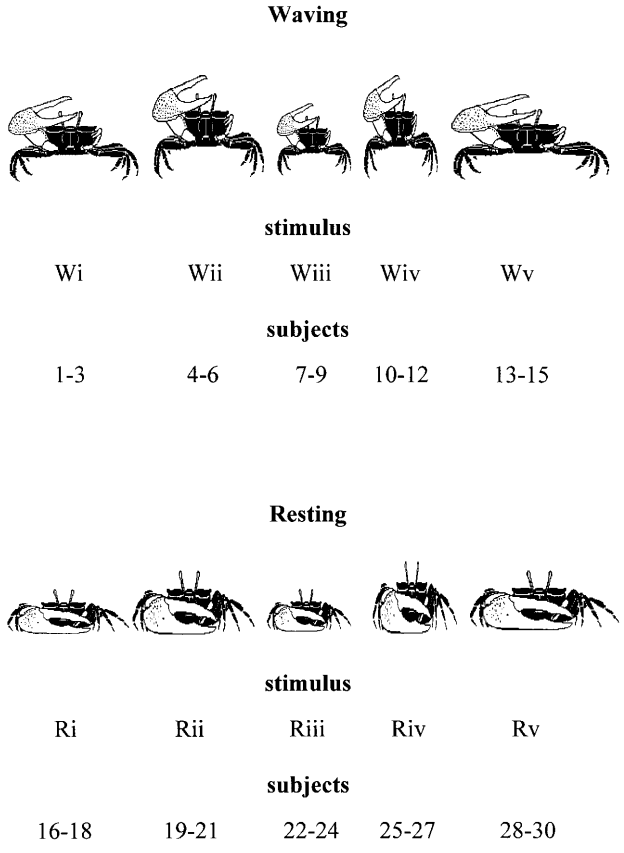

Fig. 2 Ten stimuli were played back to 30 subjects (male fiddler crabs 1–30). Each of the stimuli was played to three different subjects. Five of the stimuli (Wi–Wv) showed different males waving and five stimuli (Ri–Rv) showed different males resting. This is an example of a two-level, mixed-model, nested analysis of variance. It can be used to answer the question of whether a waving male elicits a different response from a resting male (cf. Fig. 1)

cal terms). In our fiddler crab example, if the stimuli and subjects came from the same population, then strictly our result (from the design shown in Fig. 2) only applies to this population. To claim that the experiment applies to different populations is special pleading, with no statistical back-up. If we wanted to test whether the difference in responses was affected by population, we would have to have sufficient replicates at the population level. This should make clear the link between pseudoreplication and external validity. One can design an experiment to avoid pseudoreplication at one level (e.g. within population in Fig. 2), but to generalise to the level above this (all populations of the species) is to commit pseudoreplication. However, one must stop somewhere! Of course we would like our experiments to have world-wide validity and for all time, but we actually perform experiments in a limited number of places, for a limited period of time, and with practical limitations on the numbers of experimental subjects and replicate stimuli. All experiments, therefore, contain limitations in the degree to which their results can be generalised, but we can choose where those limitations arise by juggling with experimental design.

The issue of suitable designs of playback experiments to avoid pseudoreplication at various levels has been quite extensively aired in the literature (e.g. Kroodsma 1989, 1990; McGregor et al. 1992a). A suitable design for any of the questions likely to be addressed by video playback almost certainly exists in the literature.

The take-home message of this section is that pseudoreplication has to be considered at the design stage of a playback experiment (after this, it is too late) and the question addressed by the experiment must be clearly defined. This is an easy aspect of design to get right and together with care in interpretation, in particular recognising the limits to interpretation and the special pleading involved in generalising, will avoid the pseudoreplication problem in playback experiments.

Experimental execution

The huge range of subtle, unreported, but critical issues associated with how an experiment was conducted (also called execution errors) is much more likely to mislead understanding than pseudoreplication, because pseudoreplication is much easier to detect in published papers. This point is made strongly by Hurlbert (1984) (who stimulated much of the pseudoreplication debate), but it has been largely ignored, overshadowed by the pseudoreplication issue. In an attempt to redress the balance, a consensus paper on the pseudoreplication debate included a section on features considered to be important during the execution of bird song playback (Table I in McGregor et al. 1992a). Whether these features do have an important effect on playback outcome has rather rarely been investigated directly; they are simply considered so likely to have an effect that they are controlled for in playback designs. Unlike bird song playbacks, rather few video playbacks are conducted in the field [but see Clark et al. (1997) and Burford et al. (2000) in this issue]. Nevertheless, many of the factors considered by practitioners of bird song playback are also relevant to video playback. Several more are mentioned in the consensus article in this issue (Oliveira et al. 2000). If authors include more detail on procedural aspects of experiments in their manuscript and if they are encouraged to do so by editors (currently, the reverse is generally true), then a body of best practise for video playback will become established in the literature.

Issues of experimental analysis

This section deals with three questions: (1) How many and what type of measure of response to playback should be recorded and how should they be collected? (2) How should several measures of response to playback be treated and can they be combined? (3) How should an apparent lack of response to playback be interpreted? [Note that Schlupp (2000) deals with this question in a complementary way in this issue.]

Which responses to measure

The issue often seems to be decided by what others working on the species or topic have measured, or have been able to measure. While there is nothing wrong with such criteria, it is worth giving some thought to how these measures relate to what you really want to know. In most instances playback experiments are designed to address questions that ultimately relate to genetic fitness (e.g. mate choice or resource defence) but it is rarely possible to measure the effect on fitness directly. The further from the real effect of interest the responses measured are, the more likely it is that alternative explanations exist for any difference found. For example, mate choice may be inferred from measuring the time a female spends near a male. In some cases experiments have been done to see whether such apparent indicators of mate choice translate into actual mating with the male (Bischoff et al. 1985; Ryan et al. 1990), but in most cases such a relationship remains inferred rather than demonstrated. It is worth remembering that time spent in one place may result from spending time avoiding an alternative place and therefore close association does not necessarily indicate choice to associate. Given that some video stimuli appear to be aversive (see Fleishman and Endler (2000) in this issue), care should be taken when inferring choice of a stimulus from time spent in its presence, as the subjects may be avoiding the other stimulus. This problem can be overcome in some cases by combining time spent close to a stimulus with another, more direct measure of the behaviour of interest. For example, female fighting fish *Betta splendens* develop a pattern of vertical bars when ready to mate (Simpson 1968). In this species, time spent in close proximity while displaying this reproductive coloration is therefore likely to be a more relevant indication of mate choice than a measure of proximity alone (e.g. Doutrelant and McGregor, submitted).

Is there any merit in measuring many, rather than few, response behaviours? Except in the instance of a pilot experiment, this approach often generates more problems than it solves. Even with post hoc correction of significance probabilities, a spurious result is more likely. A good criterion to follow is that there should be a priori justifications of the measure taken. In some cases this may mean measures that will allow data to be compared with those obtained by others, even if there are reasons to think that such measures are flawed (e.g. Lafleur et al. 1997).

The question of whether to use qualitative (e.g. occurrence of a display) or quantitative (e.g. time spent close) measures of response is to some extent pre-empted by considerations of which is more closely related to the behaviour of interest (see above). Appropriate statistical techniques exist to handle qualitative data (e.g. re-sampling statistics, Manly 1997), so statistical tractability as a justification for the use of indirect but quantitative measures is no longer valid.

The advantages and disadvantages of video recording responses are fairly obvious and often discussed in books about behavioural design and analysis (e.g. Martin and Bateson 1993). Its worth pointing out here that if the measure of response relies on a qualitative assessment (e.g. level of display) then video recordings can be used to verify observer judgements post hoc or to eliminate observer bias. Video recordings can also be a useful tool when collecting data in real time, rather than subsequently extracting it from recordings. A video camera can provide different views of the subjects' responses, eliminating problems noting colours or displays caused by the restricted view and angle of view of the observer. The close proximity of a video camera is also less likely to modify subjects' behaviour than the close proximity of a human observer. However, it may be more difficult to extract data that relies on good spatial resolution and 3D cues from a monitor than from the real scene (G.G. Rosenthal, personal communication). Therefore, it is a good idea to check that the behaviours of interest can be adequately extracted from a monitor. If there are other good reasons for video taping the experiment, but if some behaviours cannot be adequately noted from the monitor, it will be necessary to have an observer note these behaviours from the real scene.

Combining measures of response

The usual reason for combining several measures of response into one is to include different types of response (e.g. some females may approach, others may display, but both have responded in some way to the playback). Other reasons include making the interpretation of responses to playback easier and avoiding the statistical issues concerned with correlated measures (see McGregor 1992 for further details). The best way to combine measures is to use a principal components analysis and to perform statistical analysis on the statistically independent measures that are the result of such an analysis. Good descriptions of the procedure exist in the literature [Manly (1986) is particularly clear] and the technique has been discussed in relation to playback (McGregor 1992). One practice that has become common in the bird song playback literature is to carry out statistical analyses on combined measures of response derived by principal components, but to present summary statistics of the original measures as an aid to interpretation (e.g. Naguib 1996).

There is no merit in producing a single measure of response (often called an index) by combining arbitrarily weighted original measures (e.g. aggression index = $2 \times$ time fighting $- 0.75 \times$ time feeding). As the weights and signs of the combination are arbitrary, the measure is as likely to generate a spurious result as not.

No difference in response to stimuli?

In the sense that science proceeds through the falsification of hypotheses, at least in a Popperian view, the lack

of a significant difference in response is the strongest result that an experiment can obtain. In practice, it is difficult to exclude other reasons for the lack of a difference in response, reasons that have nothing to do with the falsification of the hypothesis. There are three broad classes of these reasons.

First, the measures of response may not be sensitive enough to detect the difference shown by the animal. For example, measuring the time that a female spends within a certain distance of video images of potential mates may fail to detect a change in behaviour elicited by video playback, whereas the time the female spent in reproductive colour while looking at the different images might have detected the difference. One way to be sure that the measure of response is sensitive enough and, perhaps more to the point, bears some relationship to the question of interest is to carry out pilot experiments.

A second reason for a lack of a difference in response is that although the animals can detect a difference between the stimuli, the same response is elicited by the stimuli. For example, in the context of a male defending his territory, any visual stimulus representing the presence of a rival within his territory boundary may elicit a similar, high level of territory defence. This would be an example of the ceiling effect (Martin and Bateson 1993). A floor effect (a uniformly low response) can also produce a lack of a difference in response. The common aspect of these effects is that a failure to find a difference in response occurs because a difference in response would be inappropriate to the context simulated by playback even though the difference (e.g. a manipulation of fiddler crab claw size) can be perceived by the subjects. In many cases, with well-known systems, it is possible to design the context and nature of the presentation to avoid ceiling and floor effects.

The final reason for a lack of a difference in response is that the subjects do not perceive the difference. Many of the articles in this issue deal with the question of whether video systems represent visual stimuli sufficiently adequately for the subjects to perceive the same difference that we humans see. The only point to add to their careful treatment of the subject is that whether the subjects can perceive a difference in natural stimuli and whether they can perceive a difference in the video stimulus presented are somewhat different issues.

Special design and analysis issues in interactive playback

Most playback experiments to date have involved the presentation of fixed stimuli, in the sense that the response of the subject has no influence on the playback stimulus. By contrast, interactive playback generally refers to the ability of investigators to alter the playback stimulus (e.g. the timing of delivery and type of behaviour) in response to the behaviour of the subject. I am not aware of any interactive video playback experiments. However, interactive video playback is technically feasible; therefore it seems appropriate to examine the design and analysis issues that interactive acoustic playback has raised. Dabelsteen and McGregor (1996) review the application of interactive playback to investigations of bird acoustic communication, and Peake et al. (2000) discuss how interactive playback of sounds is achieved. Although interactive playback raises the same issues of design and analysis as any playback experiment (see above), it also raises additional issues as a consequence of the experimental interaction with the subject. Such issues are the same as those raised by any experiment in which stimulus presentation changes in response to the behaviour of the subject, for example, the successive presentation of models or robots in different display postures as the subject's displays change. Note that playback of a stimulus in which the image varies (e.g. a sequence of courtship displays) independently of the subject's behaviour is not interactive playback, and there are no additional considerations.

Experience with interactive acoustic playback to birds, in which the subjects were territorial males, has consistently found two features that should be considered at the stage of experimental design (McGregor et al. 1992b; Dabelsteen and McGregor 1996; Dabelsteen et al. 1996, 1997). First, each trial with an individual is unique because the subject determines the detail of interactive aspects of the playback. This has the potentially unfortunate consequence that the "same" treatment may vary widely between subjects in presentation details that are likely to affect response measures (e.g. the total duration of stimulus presented). An experimental design in which all subjects receive all treatments (balanced for presentation order) can overcome this effect to some extent, but not if an individual responds differently to the same stimulus on different occasions. In most published interactive playback experiments the different treatments were designed to differ greatly (e.g. no song overlap vs. total song overlap) and post hoc analysis was unable to find a significant effect on response strength of other features that differed between treatments (e.g. Dabelsteen et al. 1997). The second feature common to interactive playback experiments with birds was that subjects appeared to habituate more rapidly than to non-interactive playback. This effect makes it more difficult to reduce the effect of between-subjects variation in response by presenting all playback treatments to all subjects. This is particularly true if there are more than two treatments and, as a consequence, recent interactive playbacks have tended to restrict the design to two treatments per subject (e.g. Otter et al. 1999).

It is also important to note that interactive playback does not automatically avoid issues of pseudoreplication. As always, whether pseudoreplication is a problem depends on whether there is a match between the question addressed and a suitable sample size for the statistical test employed to answer it (see above).

In conclusion, interactive playback has particular problems of design and analysis. These problems are unlikely to be outweighed by benefits associated with the greater "realism" of interactive playback. However, as

there is rather little knowledge about what constitutes a natural range of variation for most interactive features (but see Grafe 1999), this may be an unduly pessimistic view. At present it seems prudent to restrict interactive techniques to studies of the signal value of interactive features of communication.

Overall conclusion

The production of stimuli for use in video playback is a complicated and time-consuming business, as the other articles in this issue amply demonstrate. Once these difficulties have been overcome, the temptation to go and try them must be almost overwhelming. However, the point of this article is that there is a final stage to go through before the experiment can be run. As the outcome, or lack of outcome, of an experiment is only as good as the design and analysis of the experiment, this last stage is no less important than those preceding it.

Acknowledgements I thank ISPA for the support that enabled the workshop to be held; SNF grant 9801928 for personal financial support; Francis Gilbert for valuable comments on pseudoreplication; and Tom Peake, Fiona Burford, and Gil Rosenthal for helpful comments on previous drafts.

References

Barnard CJ, Gilbert FS, McGregor PK (1993) Asking questions in biology: design, analysis and presentation in practical work. Longmans, London

Bischoff RJ, Gould JL, Rubenstein DI (1985) Tail size and female choice in the guppy (*Poecilia reticulata*). Behav Ecol Sociobiol 17:253–255

Burford FRL, McGregor PK, Oliveira RF (2000) Response of fiddler crabs (*Uca tangeri*) to video playback in the field. Acta Ethol 3:55–59

Clark D, Macedonia J, Rosenthal GG (1997) Testing video playback to lizards in the field. Copeia 1997:421–423

Dabelsteen T, McGregor PK (1996) Dynamic acoustic communication and interactive playback. In: Kroodsma DE, Miller EH (eds) Ecology and evolution of acoustic communication in birds. Cornell University Press, Ithaca, N.Y., pp 398–408

Dabelsteen T, McGregor PK, Shepherd M, Whittaker X, Pedersen SB (1996) Is the signal value of overlapping different from that of alternating during matched singing in great tits? J Avian Biol 27:189–194

Dabelsteen T, McGregor PK, Holland J, Tobias JA, Pedersen SB (1997) The signal function of overlapping singing in male robins (*Erithacus rubecula*). Anim Behav 53:249–256

Fleishman LJ, Endler JA (2000) Some comments on visual perception and the use of video playback in animal behavior studies. Acta Ethol 3:15–27

Grafe TU (1999) A function of synchronous chorusing and a novel female preference shift in an anuran. Proc R Soc Lond B 266: 2331–2336

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211

Kroodsma DE (1989) Suggested experimental designs for song playbacks. Anim Behav 37:600–609

Kroodsma DE (1990) Using appropriate experimental designs for intended hypotheses in 'song' playbacks, with examples for testing effects of song repertoire size. Anim Behav 40:1138–1150

Lafleur DL, Lozano GA, Sclafani M (1997) Female mate-choice copying in guppies, *Peocilia reticulata*: a re-evaluation. Anim Behav 54:579–586

Manly BFJ (1986) Multivariate statistical methods: a primer. Chapman and Hall, London

Manly BFJ (1997) Randomization, bootstrap and Monte Carlo methods in biology, 2nd edn. Chapman and Hall, London

Martin P, Bateson P (1993) Measuring behaviour, 2nd edn. Cambridge University Press, Cambridge

McGregor PK (1992) Quantifying responses to playback: one, many or multivariate composite measures? In: McGregor PK (ed) Playback and studies of animal communication. Plenum Press, New York, pp 79–96

McGregor PK, Catchpole CK, Dabelsteen T, Falls JB, Fusani L, Gerhardt HC, Gilbert F, Horn AG, Klump GM, Kroodsma DE, Lambrechts MM, McComb KE, Nelson DA, Pepperberg IM, Ratcliffe L, Searcy WA, Weary DM (1992a) Design and interpretation of playback: the Thornbridge Hall NATO ARW consensus. In: McGregor PK (ed) Playback and studies of animal communication. Plenum Press, New York, pp 1–9

McGregor PK, Dabelsteen T, Shepherd M, Pedersen SB (1992b) The signal value of matched singing in great tits: evidence from interactive playback experiments. Anim Behav 43:987–998

Milinski M (1997) How to avoid seven deadly sins in the study of behaviour. Adv Study Behav 26:159–180

Naguib M (1996) Ranging by song in Carolina wrens *Thryothorus ludovicianus*: effects of environmental acoustics and strength of song degradation. Behaviour 133:541–559

Oliveira RF, Rosenthal GG, Schlupp I, McGregor PK, Cuthill IC, Endler JA, Fleishman LJ, Zeil J, Barata E, Burford F, Gonçalves D, Haley M, Jakobsson J, Jennions MD, Körner KE, Lindström L, Peake T, Pilastro A, Pope DS, Roberts A, Rowe C, Smith J, Wass JR (2000) Considerations on the use of video playbacks as visual stimuli: the Lisbon workshop consensus. Acta Ethol 3:61–65

Otter KA, McGregor PK, Terry AMR, Burford FRL, Peake TM, Dabelsteen T (1999) Do female great tits (*Parus major*) assess males by eavesdropping? A field study using interactive song playback. Proc R Soc Lond B 266:1305–1309

PeakeTM, Otter KA, Terry AMR, McGregor PK (2000) Screech: an interactive playback program for PCs. Bioacoustics (in press)

Ryan MJ, Hews DK, Wagner WE Jr (1990) Sexual selection on alleles that determine body size in the swordtail *Xiphophorus nigrensis*. Behav Ecol Sociobiol 6:231–237

Schlupp I (2000) Are there lessons from negative results in studies using video playback? Acta Ethol 3:9–13

Simpson MJA (1968) The display of the Siamese fighting fish, *Betta splendens*. Anim Behav Monogr 1:1–73

Sokal RR, Rolf FJ (1981) Biometry, 2nd edn. Freeman, San Francisco