Animal Behaviour 86 (2013) 483–488

Contents lists available at SciVerse ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Commentary

Pseudoreplication: a widespread problem in primate communication research

B. M. Waller [a,*], L. Warmelink [a], K. Liebal [a,b], J. Micheletta [a], K. E. Slocombe [c]

[a] Centre for Comparative and Evolutionary Psychology, Department of Psychology, University of Portsmouth, Portsmouth, U.K.
[b] Cluster Languages of Emotion, Evolutionary Psychology, Freie Universitat Berlin, Berlin, Germany
[c] Department of Psychology, University of York, York, U.K.

ARTICLE INFO

Article history:
Received 27 February 2013
Initial acceptance 29 March 2013
Final acceptance 15 May 2013
Available online 28 June 2013
MS. number: 13-00188

Keywords:
ape
facial expression
gesture
monkey
pooling fallacy
pseudoreplication
statistics
vocalization

Pseudoreplication (the pooling fallacy) is a widely acknowledged statistical error in the behavioural sciences. Taking a large number of data points from a small number of animals creates a false impression of a better representation of the population. Studies of communication may be particularly prone to artificially inflating the data set in this way, as the unit of interest (the facial expression, the call or the gesture) is a tempting unit of analysis. Primate communication studies (551) published in scientific journals from 1960 to 2008 were examined for the simplest form of pseudoreplication (taking more than one data point from each individual). Of the studies that used inferential statistics, 38% presented at least one case of pseudoreplicated data. An additional 16% did not provide enough information to rule out pseudoreplication. Generalized linear mixed models determined that one variable significantly increased the likelihood of pseudoreplication: using observational methods. Actual sample size (number of animals) and year of publication were not associated with pseudoreplication. The high prevalence of pseudoreplication in the primate communication research articles, and the fact that there has been no decline since key papers warned against pseudoreplication, demonstrates that the problem needs to be more actively addressed.

> [pooling multiple observations from each individual] reflects a fundamental error in the logic underlying random sampling since it implicitly assumes that the purpose of data gathering in ethology is to obtain large 'samples of behaviour' rather than samples of behavior from a large number of individuals. **(Machlis et al. 1985, page 201)**

The goal of the majority of behavioural scientists is to draw conclusions about a specific population of individuals (usually a species) by examining a sample of individuals from this group. Scientists seek the largest samples they can achieve, in order to increase the reliability of extrapolating findings from their sample to the population, and thus increase the accuracy of their conclusions. Reliability, however, can be increased only by increasing sample size in terms of the number of individuals that make up the sample, not by taking multiple samples from each individual and pooling them to create a larger data set. In the 1980s, two key papers highlighted the problem of drawing conclusions from these artificially inflated samples, terming it 'pseudoreplication' or the 'pooling fallacy' (Hurlbert 1984; Machlis et al. 1985), and it is now widely acknowledged as an error to be carefully avoided.

The simplest sampling error that can lead to pseudoreplication is extracting more than one data point per individual, and then adding these data to the main data set without using appropriate repeated measures statistical techniques. The data points are thus not independent. The pooling procedure then results in an artificially inflated sample size, which falsely raises statistical power and increases the chances of making a type I error (a false positive: rejecting the null hypothesis when it is true). Mundry & Sommer (2007) demonstrated (using mock analyses on real data: contact calls of the Arabian babbler, *Turdoides squamiceps*) that conducting discriminant function analysis (DFA) on nonindependent data increases the chance of rejecting the null hypothesis. Specifically, the discriminability of groups (species, sexes, contexts, etc.) was overestimated when multiple samples were taken per animal (10 calls per individual), and the factor 'subject' was not taken into account.

Pseudoreplication can also occur when data points result from the same stimulus treatment and/or temporal, spatial or social

* Correspondence: B. M. Waller, Department of Psychology, University of Portsmouth, King Henry Building, Portsmouth PO1 2DY, U.K.
E-mail address: bridget.waller@port.ac.uk (B. M. Waller).

0003-3472/$38.00
http://dx.doi.org/10.1016/j.anbehav.2013.05.038

group, and so are not statistically independent from each other. In animal behaviour, experimental playback designs may be particularly prone to this form of pseudoreplication, as a single exemplar (e.g. one example of a type of call) is often used, but then multiple responses to this exemplar are analysed (Kroodsma 1990; Kroodsma et al. 2001). This latter type of pseudoreplication can be more difficult to identify (and avoid, but see Zuur et al. (2009) for methods to avoid temporal nonindependence), but the same principle applies: the replicates are not independent from each other and/or they do not increase the reliability of generalizations from the sample to the wider population.

Some scientists argue that by identifying anything connected through spatiotemporal proximity or physical boundaries as nonindependent, Hurlbert (1984) sets impossible standards for the design of experiments (Schank & Koehnle 2009). The authors argue that drawing lines around the experimental unit in this way can be arbitrary, and that the decision about whether units are independent or not needs to be made in light of the specific research question (and, if possible, through empirical analysis). Blanket acceptance of the relationship between statistical independence and spatiotemporal proximity may render pseudoreplication as a 'pseudoproblem' (Schank & Koehnle 2009, page 421). Others have argued that the need for replicates within an experiment is unnecessary, unless the predicted response is weak or highly prone to noise (Oksanen 2001). However, even the harshest critics of pseudoreplication tend to concur that taking multiple data points from one animal and treating them as independent is incorrect: 'If the point is that treating repeated measures, say from the same animals, as independent data points is an error, then we completely agree' (Schank & Koehnle 2009, page 422). Others predict that this simple pooling form of pseudoreplication must be rare given the wide dissemination of the classic papers that urged avoidance: 'An important result of Hurlbert's article (and others, e.g. Machlis et al. 1985) is that authors today are unlikely to publish articles with obviously pseudoreplicated data' (Freeberg & Lucas 2009, page 450).

Several reviews have documented the incidence of pseudoreplication in specific subfields of ecological and behavioural science. Hurlbert's original paper (Hurlbert 1984) reported an incidence of 48% among 101 ecological studies (that used inferential statistics) published between 1960 and 1980. In a review of invertebrate field experiments nearly a decade later, Hurlbert & White (1993) reported a lower pseudoreplication incidence of 32%. Later still, Heffner et al. (1996) found 12% pseudoreplication in a sample of articles akin to Hurlbert's original review. Thus, at least in ecology, the frequency of pseudoreplication seems to be on the decline. However, in other fields pseudoreplication is only starting to be highlighted, and so the problem may still occur at a greater frequency. A more recent review of pseudoreplication in zoo biology studies, for example, found an incidence of 40% (Kuhar 2006).

It is possible that studies of animal communication are particularly prone to artificially inflating the data set, as the unit of communication (the call, the facial expression, etc.) is a tempting unit of analysis. Using the unit of communication as the unit of analysis could often lead to pseudoreplication, unless: (1) each individual contributes only one data point averaged from their sampled communication; (2) appropriate within-subject statistical analyses are used (e.g. paired *t* test, Wilcoxon signed-ranks test, repeated measures ANOVA); or (3) individual-level membership is taken into account by the statistical test (e.g. by using hierarchical modelling techniques; Pinheiro & Bates 2009). The pervasiveness of this issue within communication research, however, has not been explored. Here, we focused on primate communication. The aim of this study was (1) to examine the prevalence of pseudoreplication in a systematic review of peer-reviewed articles published in the primate communication field and (2) to determine which factors (if any) were associated with pseudoreplication.

METHODS

Article Database

We reviewed a database of articles systematically collected for a previous study (see Slocombe et al. 2011 for a detailed description of the search criteria). The database contained 551 empirical, peer-reviewed research articles published in English and conducted on naturalistic, conspecific primate communication (excluding studies on communication with humans) from 1960 to 2008. Each article was coded for the primary modality of communication under investigation (vocal, gestural, facial or multimodal), method (whether the study used observational methods), taxa (great ape, lesser ape, monkey or prosimian), research environment (wild or not), impact factor of the journal (in 2011) and citations per year (at time of search: April 2013).

Coding for Pseudoreplication

Each article was read and the statistical analysis coded for the presence of pseudoreplication. Each article was classified as: (1) presenting no statistics; (2) undeterminable; (3) not including pseudoreplicated data; or (4) including pseudoreplicated data. Articles coded as presenting no statistics did not use inferential statistics (i.e. did not contain tests that yield *P* values). They may still have used descriptive statistics such as frequencies, percentages or similar, so although it is possible that these data were pseudoreplicated, any pseudoreplication of this sort was not counted in the coding system. Undeterminable articles either (1) did not present enough data for us to determine whether pseudoreplication had taken place, (2) stated that information was not available to the researcher (e.g. the researcher could not reliably track the number of animals) or (3) focused on a level below the individual (e.g. the neuron). The appropriate unit of analysis has also been questioned in neuroscientific studies (Lazic 2010), but we felt that comprehensive treatment of these papers was beyond this review.

To classify each article as pseudoreplicating or not pseudoreplicating data, the use of statistics was examined. For each statistic, the reported sample size (number of animals), the statistical test used and the degrees of freedom were noted. If this information was not mentioned in the text, the coder checked figures, tables, captions, footnotes or additional published material. This information was not always stated specifically in numbers, but if the text made clear how the data had been treated, this information was also accepted. A statistic was classified as pseudoreplicating data if the stated degrees of freedom were higher than the stated sample size, or if the analysis was explicitly conducted on the number of observations, and actual sample size was not included. Some tests (e.g. repeated measures ANOVAs) create higher degrees of freedom than the sample size. If such a test was used, the coder checked whether a subject contributed more than one data point per condition. If so, the statistic also qualified as pseudoreplication. For some statistics, the sample size, the test used or the degrees of freedom could not be determined. These statistics were listed as undeterminable. An article was classified as having pseudoreplicated if at least one of the statistic tests presented included pseudoreplicated data, regardless of the presence of any other statistics that did not pseudoreplicate. An article was classified as undeterminable only if all statistics used in the article were undeterminable.