CellPress
REVIEWS

Opinion

Using Biological Insight and Pragmatism When Thinking about Pseudoreplication

Nick Colegrave[1] and Graeme D. Ruxton[2,*]

Pseudoreplication is controversial across experimental biology. Researchers in the same field can disagree on whether a given study suffers from pseudoreplication and on to what extent any pseudoreplication undermines the value of a study. A recent survey indicated that concerns about pseudoreplication can strongly impact peer review of manuscripts submitted for publication. Here we explore controversies around pseudoreplication, identify issues requiring resolution, and in each case offer a resolution. We emphasise that having non-independence in data points and pseudoreplicating are not the same thing. Researchers should be able to demonstrate that in a given experiment they have minimised and controlled the risk of non-independence weakening their study. If they do that to the satisfaction of others, they have avoided pseudoreplication.

Pseudoreplication: Important but Controversial

In 1984 Stuart Hurlbert published a monograph called 'Pseudoreplication and the design of ecological field experiments' [1]. This publication has been tremendously influential in the design and analysis of experiments across the biological sciences (it has been cited over 5000 times). Nevertheless, Hurlbert's recommendations have not been universally accepted, and papers have also been published with titles such as 'Logic of experiments in ecology: is pseudoreplication a pseudoissue?' [2] and 'Pseudoreplication is a pseudoproblem' [3]. A recent review on the subject came to the conclusion that pseudoreplication remained a 'controversial issue' [4]. That study found that 50% of surveyed ecologists reported having manuscripts criticised by journal reviewers on grounds of pseudoreplication. Further, pseudoreplication has recently been described as one of the key factors underlying the publication of false-positive findings in the scientific literature [5]. Our aim is to explore controversies around the application of the concept of pseudoreplication, identify issues requiring resolution, and offer such resolutions.

What Is Pseudoreplication?

The key issue here is that replication helps us because, if we have done things properly, the individuals we have measured are a random sample of the population we are really interested in. So imagine that we want to estimate the characteristic mass of male mice in a large laboratory colony as part of an exploration of possible sex difference in mass. We select male mice by simple random sampling (without replacement) from the colony, weighing each individual. The actual characteristic mass of our male population is 45.2 g. Of course, when we draw male mice for our sample, each of their actual masses is likely to differ from the population average by some amount, with some being bigger than average and others smaller. Let us call this deviation of each individual's mass from the population average their residual mass. If our sample is random, the residual for one individual in our sample will be entirely uncorrelated with that of any other mouse in the sample. For example, if the first mouse has a small negative residual, this gives us no clue about whether the residual of the next mouse will be positive, negative, small, or large. This is important because it means that as we sample more and more individuals, the average size of the residual in our sample

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, UK
[2]School of Biology, University of St Andrews, St Andrews, UK

*Correspondence: graeme.ruxton@st-andrews.ac.uk (G.D. Ruxton).

© 2017 Elsevier Ltd. All rights reserved.

CrossMark

will approach zero and we will get an unbiased estimate of the population average mass. In a random sample, each individual measured provides an independent estimate of the thing we are interested in. However, now suppose for convenience that (rather than drawing a truly random sample of individual mice) we instead randomly sample cages each of which contains a number of mice and measure all of the individual males in each selected cage. If some of the variation in mass in our population (of all male mice that might have been in our sample) is due to environmental factors that show cage-to-cage variation, individuals in the same cage are more likely to have similar weights than two randomly chosen individuals from the population. This environmental variation might be, for example, because the cages that mice in this colony are housed in are of a range of different types with different dimensions. In this situation (if cage dimensions do affect mouse mass), the residual masses of the set of cage mates in our sample are likely to be positively correlated with each other. So if the first individual has a small negative residual, it is more likely that its cage mates will also have small negative residuals. This positive correlation of residuals means that the measures of two cage mates are not providing us with two independent measures of the average mass of the population. If we treat them as independent, our estimate of the mean mass of the population will be biased. Put another way, two non-independent individuals give us less information about the distribution of values of the trait of interest in our statistical population than two independent individuals. Of course, independence makes sense only relative to a specified population. In the example above, cage mates do not provide independent measures of the average mass of our laboratory population, but if we were actually interested in the population 'male from cage A15', provided we have measured random individuals from within that cage, our cage mates do provide independent information. An underlying assumption of many statistical tests is that all of the individuals in a sample can be considered independent of each other. If we apply such a test to samples where some individuals are not statistically independent, the test can become unreliable and we have committed pseudoreplication. This is essentially noncontroversial. Let us now look at some areas where there might be apparent disagreement between researchers. Before this, however, we should note that we do not consider replication as an essential component of every ecological study (Box 1) and that inferential statistics can still be of value in unreplicated studies (Box 2).

Box 1. Are Studies with a Sample Size of One of Any Value Scientifically?

This issue was the subject of a flurry of papers early this century [2,6–8] but is now less controversial: everyone essentially agrees that just one counterexample can be sufficient to falsify a hypothesis. Similarly, if we are interested in the entity we are measuring in its own right (rather than using it to generalise to a wider population), study of that single entity is enough. Everyone also essentially agrees that in situations where replication is impractical, studies with a single experimental unit can still be of value. This practical limitation is likely to occur most often in situations where the mechanisms of interest occur on a large spatial scale [9]; this might involve, for example, the responses of an island ecosystem to the eradication of introduced mammals. The key issue here is that if it were possible to replicate and observe ecosystem change across a representative sample of islands, it is appropriate to make inferences on the wider population of all islands on the basis of that sample. We cannot do this on the basis of a study with a single experimental unit. However, if (on the basis of previous empirical work and/or understanding of mechanisms) we can make predictions about how we would expect island ecosystems to behave in response to this perturbation, we can use a one-unit study to test whether those general predictions hold in that one specific case. In our island example, if we predict that removal of mammals will lead to an increase in tree cover on islands, we can test whether this general prediction holds for one particular island. If it does we have evidence in support of the general understanding (it has been supported in at least one test case); if not we have the potential to explore whether current understanding requires modification and/or whether we can understand why this particular island does not conform to our *a priori* prediction. Clearly, the more detailed the prediction we can make (e.g., if we predicted that tree cover would increase fivefold over a decade) the more powerful the support for the existing understanding if the prediction is met and the more potential we have for understanding any mismatch between the existing understanding and observation on that one unit. Hence, providing authors are careful about how they interpret their study (along the lines we describe above), studies without replication can be valuable in their own right; they can also contribute to later meta-analyses.

Box 2. Is Any Use of Inferential Statistics in Unreplicated Studies a Form of Pseudoreplication?

It is highly likely that a large-scale unreplicated study will involve subsampling: it would not be practical to assess tree cover accurately across a remote island 20 km^2 in area, so the natural thing to do would be to subsample – to carefully assess vegetative cover over time in a range of smaller-scale (say 25 m^2) sampling sites across the island. Hurlbert [7] differs from some other authors (e.g., [6,8]) and from ourselves in whether it would be appropriate to use inferential statistics (e.g., calculation of effect sizes, confidence intervals, and/or *P* values) on the basis of these subsamples. Our view is that such use of statistics can aid the reader and should not mislead the reader provided the authors stick to interpreting their data appropriately, essentially remembering that they are seeking to understand one specific island and not islands generally. Hurlbert's point is that the 25-m^2 sample sites are non-independent examples of island vegetation (since they all come from the same island). He is absolutely correct, but the authors should not be seeking to make predictions about islands in general. They should be seeking to improve understanding of this particular island, and in that context (if carefully chosen) these sites could be considered as independent samples of vegetative cover on that particular island. Hence we believe that inferential statistics do have a place in studies with one (or only a few) replicates, providing authors are careful in their interpretation.

Does an Experiment Become Valueless If Data Collected from It Will Inevitably Be Strongly Pseudoreplicated?

It has been suggested [3] that readers of Hurlbert's original paper could be led to believe that this is his position, although Hurlbert stated that it was not his intention to give this impression [10]. It should be clear that this position is too extreme. In our hypothetical example, two males from the same cage give less information about the distribution of weights of male mice in this laboratory population than two males from different cages. However, weighing both of them still gives more information than weighing one of them and, provided their non-independence is appropriately dealt with (Box 3) when drawing inference from the measures, should not bias our estimate of male mouse mass. So we do not think it is true that non-independence of measurement units necessarily makes some experimental designs valueless. However, non-independence (even if handled appropriately from a statistical standpoint) will generally reduce the power of an experiment to detect the effect of interest relative to a redesigned experiment where a similar number of measurement units can be considered statistically

Box 3. Non-independence and Statistical Analyses

Issue 1: If Measurement Units Are Non-independent, Should We Replace Them in Our Sample with an Average Value?

This was Hurlbert's recommendation in 1984 [1], but he freely admits in 1990 [10] that for many biologists their comfort with more sophisticated statistical analyses has increased and now multilevel models are sometimes a more attractive option (see [11] for an overview). It often makes sense to use the latter approach, since by taking an average value (of mice within a cage, say) we are discarding information about variation within cages. Hence there is now no real disagreement that multilevel models can provide a powerful approach to dealing with non-independent data points. However, we also note that in many situations an analysis of the mean values is exactly equivalent to the more complex multilevel analysis but has the advantage that the true level of independent replication is explicit and clear [12]. Thus we would caution against immediately reaching for a more complicated but equivalent model just because it is available.

Issue 2: If Some Statistical Techniques Can Cope with Non-independence, Can We Stop Worrying About Non-independence?

This could be one reading of the position advocated by Schank and Koehnle [3]. For example, they say 'The initial reception to Hurlbert's (1984) paper reflected genuine and widespread concern about the design and analysis of experiments. Ultimately, though, increased methodological sophistication, careful thought, and the development of new statistical techniques are solving these problems.' We disagree with any implication that our increasing ability to analyse non-independent data appropriately and effectively should cause us to worry less about non-independent data. The simple fact is that the more strongly dependent data points are, the less efficient our sampling is: the less valuable new information we get for each added data point. Thus, there are practical, financial, logistical, and ethical reasons why we should generally strive to reduce the strength of any non-independence across our sampled data points. Multilevel statistical analysis can help us avoid turning non-independent data into pseudoreplication, but we should generally strive in our experimental design to minimise the strength of the non-independence in the first place.

independent. Sometimes such non-independence can reduce power sufficiently that it seems unwise to go ahead with a planned experiment. It would be good practice to attempt to minimise the extent of non-independence of sampling units in the design of experiments to boost power. However, once an experiment has been completed, concern about non-independence should not immediately imply that the experiment was valueless. It has to be remembered that pseudoreplication is essentially a mismatch between the statistical analysis performed on data and assumptions that we feel comfortable making about the nature of independence of the measurement units, and some statistical techniques can cope with non-independence (Box 3).

Is Non-independence Essentially an Empirical Question That Can Be Decided Only by Analysis of the Data?

This could be one reading of the position advocated by Schank and Koehnle [3]. For example, they say 'pooling across units of analysis is not necessarily a statistical error. It is a decision made after an appropriate statistical analysis reveals that there are no detectable dependences across what we thought might have been the unit of analysis and is quite commonly done in multilevel analyses.' We think there is a need for several notes of caution in adopting this philosophy. Imagine that we are again comparing the masses of male and female mice from a laboratory population and include the identity of the cage that each mouse is kept in as a random factor in our analysis to control for possible non-independence of cage mates. If we perform such a statistical analysis and find that 'cage identity' does not come out as being significant, this implies that we have no evidence that the specific cage a mouse is kept in influences its weight. However, no evidence of an effect is not the same as evidence of no effect; we could be making a type II error and failing to detect an effect that really exists. Such an error is likely when our experiment was probably not designed specifically to give us strong power to test that hypothesis. The quote above from Schank and Koehnle [3] suggests that if they found that cage identity was not a significant factor in their more complex model, they would drop this factor out, essentially treat all mice as independent, and perform a *t*-test (for example) to compare the two sexes. In general we take a different philosophical view.

Where our measures are subsamples of our experimental units – for example, in a nested experimental design where multiple mice in a cage are measured but the treatment is applied to the whole cage – cage identity should always be included in the statistical analysis (see [13] for an overview). However, what if different treatments can be applied at random to individuals within the same cage? Here, whether you should include cage identity in the analysis is less straightforward.

If, on the basis of the results of previous, similar experiments or an understanding of the underlying mechanisms, you feel there is a potential source of non-independence that you can account for (like cage identity in our example), we agree that it is a good idea to include this factor in the statistical analysis. However, we believe that, you having made that decision, that factor should then stay in the statistical analysis, because that factor not reaching significance in your analysis is not good evidence that it is of trivial importance. To us, removing it is unjustified on statistical grounds, and if we perform only the simpler *t*-test sometimes (on the basis of pretesting the data), we cannot be confident that the type I error rate of the *t*-tests that we do perform will remain at the nominal (generally 5%) value. We would use such model-simplification approaches only where the philosophy behind the original study is one of exploration; that is, in a situation where we feel we know little about the system and are performing a study to suggest factors that might be worth further exploration in the future. If, however, the study is not exploratory but has been motivated by a wish to test specific hypotheses (like the sex difference

in the mass of mice in our example), we do not feel that such model simplification is wise, for the reasons given immediately above (see [14] for a fuller discussion).

Our philosophical approach also implies that there is a cost to including factors that might possibly be a source of non-independence. For example, imagine that we conduct an experimental study where two different fertilisers are given to individual plants growing in separate pots, and the pots are laid out in a square formation on the shelf of a greenhouse at random with respect to treatment. In principle, pots on one side of the shelf might experience conditions that are more similar to one another than those experienced by pots on opposite sides, so we might consider fitting shelf side as a factor to deal with this. However, conditions might also vary from front to back, so to be safe perhaps we fit 'front or back of shelf' as another factor. The cost of doing so is that each factor included in our statistical model takes away degrees of freedom from the testing of the main hypotheses of interest. Hence, we should include in our model only those factors that (on the basis of previous empirical evidence or understanding of mechanisms) we expect to have an important effect. So the best approach is to try to reduce causes of non-independence as much as possible when designing our study and then to record only those factors that we still fear might have an influence.

Can Rodents in the Same Cage Ever Be Viewed as Independent Measures of a Treatment?

Suppose your research involves rodents that are gregarious and thus, for welfare reasons, are kept in groups of several individuals to a cage. If you apply the same treatment to all individuals in each cage, individual rodents are not independent measures of the effect of the treatment. For example, we might randomise cages to have either an antibiotic or a placebo added to the food in the bowl shared by all rodents in that cage. Appropriate analysis should involve the use of either the mean value for each cage or a more complex, multilevel model including cage identity as a factor. This situation is uncontroversial. However, what if the treatments can be applied directly to individual rodents so that individuals in the same cage can be randomly allocated to different treatments? For example, imagine that individuals within each cage are each individually randomised to receive either a single-dose antibiotic or a sham injection at the start of the experiment and then have their subsequent activity levels measured. Here a fundamental disagreement emerges. Hurlbert's view is clear: rodents in the same cage can never be independent measures of treatments applied to them as they are not sufficiently isolated or physically independent of each other (e.g., [15]). In forming this view, he apparently equates physical separation with statistical independence. The same philosophy was adopted by another recent review on pseudoreplication within neuroscience [16]. However, other influential authors (e.g., [17,18]) take a quite different view: that such a study can, in principle, be treated as a randomised block design, with cage identity as a blocking factor.

We believe that there is an effective way to resolve this disagreement. Specifically, we believe that whether rodents in the same cage can be regarded as independent measures of any general effects of treatments applied to them is not simply a question of physical isolation, but instead depends on the biology of the specific case. For individual rodents in the same cage to provide independent measures of the effect of the treatment, we need to be confident that any effects of the influence of one cage mate on another do not act in such a way that individuals given the same treatment become more (or less) similar to one another than they are to individuals given different treatments. That is, we need to be confident that any effects of social behaviours or other influences of one cage mate on another do not confound treatment effects; or, in statistical terms, there is no treatment-by-cage interaction. Imagine that one active rat in a cage stimulates the activity of all members of the cage in the same way. The increased activity

Box 4. Should Individuals in Different Treatment Groups Be Housed Together?

We can imagine experimental designs where levels of a treatment are applied to individuals and then those individuals are randomly allocated to a smaller number of enclosures. This could be analysed as a randomised block design with each enclosure being a separate block. However, if you are intending to use measures of individuals within the enclosures as independent measures of your treatments, you need to be sure that enclosure effects (including interactions among individuals in the enclosure) will not induce an enclosure-by-treatment interaction. If such an interaction is suspected, each enclosure should provide one data point to look at general effects of the treatments, whereas the use of individuals as independent data points would amount to pseudoreplication. Care needs to be taken if you plan to study the effect of a given treatment and subsets of individuals from different treatment groups share common environments (e.g., fish tanks, rodent cages, other enclosures). In this situation, if you want to gain the power benefits associated with analysis where individuals (rather than enclosures) are treated as independent measures of the treatment, you need to convince others that there is no strong treatment-by-enclosure interaction (i.e., that cage identity and treatment are not confounded). How such convincing might be achieved is discussed in the main text.

Probably without realising it, we make an assumption that such confounding is not occurring in a way that will damage the validity of our study almost every time we conduct an experiment. Whenever we perform an experiment in a single location – for example, within a single laboratory – there is the potential for local effects to bias any measure of a treatment that we make. This is true regardless of whether our experimental individuals are physically isolated from one another within the laboratory. When we analyse such experiments using individual measures as independent data points, we are implicitly assuming that the results from this laboratory would generalise to other similar laboratories (i.e., that there is no treatment-by-laboratory interaction).

of all individuals in that particular cage would not be confounded with treatment; only if the increased activity was more pronounced in individuals of one treatment group would confounding occur. The same would be true of any effect of one individual on another (e.g., reaction to the noise of others) that caused a treatment-by-cage interaction, even if rodents in the same cage were physically separated by partitions. For this reason we do not feel that the presence or absence of physical interaction is a good proxy for deciding whether individuals can be treated as independent data points. Of course, this logic does not just apply to rodents in cages (Box 4).

This section highlights a general theme of our understanding of statistical independence and pseudoreplication: it is difficult to offer hard and fast rules; rather, whether two experimental units can provide independent information on a particular treatment cannot be practically resolved with certainty and comes down to a matter of opinion (based on weight of evidence). While there is little doubt that physical isolation will usually increase our confidence that individuals are independent, and so is often a very worthy aim in experimental design, lack of such isolation due to the constraints of an experimental system does not necessarily imply reduced replication.

Should We Conduct Experiments Only Where There Is No Danger of Non-independence?

Schank and Kehnle [3] suggest that readers of Hurlbert [1] could be led to believe that he advocated conducting experiments only when complete independence could be guaranteed, although Hurlbert later [10] states that this was not his position. Our view is that it is normally impossible to entirely eliminate all potential for non-independence at the experimental design stage, nor is it normally possible to demonstrate independence conclusively at the analysis stage. What you can do is strive to minimise the potential for non-independence of data points when designing the study and record potential factors that you fear might still cause non-independence. You can then explain in your report of the study how this approach has led you to adopt the particular statistical analysis that you present. Based on this explanation, and the results that you present, the reader should be able to form an opinion about whether they feel that: (i) there was likely to be non-independence of data points; and (ii) such potential non-independence has been handled appropriately. However, we emphasise that this will be a

matter of opinion. In most cases, most scientists in a particular field will reach the same opinion on reading the same report, but this will not always be the case.

To appreciate the problem, we return to the example study of a sex difference in the mass of mice. We already discussed sharing a cage as a potential source of non-independence. However, it is easy to imagine others. If (as seems likely from our experience of other systems – including humans) there is a significant genetic component to body mass, genetically -related individuals might be non-independent. It could be that the issue is spatial but not as simple as being cage based. For example the mouse cages might be spread across a number of rooms in the facility and (e.g., if rooms differ considerably in temperature or level of disturbance) individuals from the same room (but not necessarily the same cage) might be more similar to each other than to random individuals from across the population. The issue could be about previous history. For example, the mice might have been used in previous experiments that involved dietary manipulation and those that were involved in the same experiment (or the same treatment group within an experiment) might be more similar to each other than random individuals. Non-independence might even be not an intrinsic aspect of the experimental subjects but of the measurement regime; it might be that human experimenters differ in their measurement technique and so two mice measured by the same individual are more similar on average than two mice measured by different members of the research team. In truth, for any study there is a never-ending list of plausible mechanisms that might cause non-independence. This does not mean that it is impossible to conduct good experiments; it does mean that your aim in an experiment should be to reduce the potential effects of factors that *a priori* you expect to cause significant non-independence and to look to measure those factors that you feel might still be of concern after you have attempted to reduce non-independence. It is unreasonable to expect an experiment to be unequivocally free of any non-independence. It is reasonable to expect obvious factors that are likely to have a significant effect to be either controlled by careful design or measured and controlled for statistically. As discussed above, not only is it normally impossible to unequivocally demonstrate complete independence *a priori*, it is also normally practically impossible to statistically demonstrate that the collected data are entirely and unequivocally independent. Therefore it is up to authors to demonstrate that they have taken reasonable steps to reduce or statistically control for non-independence and for reviewers, editors, and readers to be able to form an opinion about how successful the authors have been in this. No party should realistically demand unequivocal demonstration of complete independence.

Concluding Remarks

Having non-independence in data points and pseudoreplicating are not the same thing. Researchers should be able to demonstrate that in a given experiment, through careful design and appropriate analysis, they have minimised and controlled the risk of non-independence weakening their study. If they do that to the satisfaction of others, they have avoided pseudoreplication.

Acknowledgment

The authors thank four reviewers for stimulating suggestions that improved this work substantially.

References

1. Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54, 187–211
2. Oksannen, L. (2001) Logic of experiments in ecology: is pseudoreplication a pseudoissue? *Oikos* 94, 27–38
3. Schank, J.C. and Koehnle, T.J. (2009) Pseudoreplication is a pseudoproblem. *J. Comp. Physiol.* 123, 421–433
4. Davies, G.M. and Gray, A. (2015) Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecol. Evol.* 5, 5295–5304
5. Forstmeier, W. *et al.* (2016) Detecting and avoiding likely false-positive findings – a practical guide. *Biol. Rev.* 92, 1941–1968

6. Cottenie, K. and De Meester, L. (2003) Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). *Oikos* 100, 394–396
7. Hurlbert, S.H. (2004) On the misinterpretation of pseudoreplication and related matters: a reply to Oksanen. *Oikos* 104, 591–597
8. Oksannen, L. (2004) The devil lies in details: reply to Stuart Hurlbert. *Oikos* 104, 598–605
9. Hargrove, W.W. and Pickering, J. (1992) Pseudoreplication: a *sine qua non* for regional ecology. *Landsc. Ecol.* 6, 251–258
10. Hurlbert, S.H. (1990) The ancient black art and transdisciplinary extent of pseudoreplication. *J. Comp. Psychol.* 124, 434–443
11. Aarts, E. *et al.* (2014) A solution to dependency: using multilevel analysis to accommodate nested data. *Nat. Neurosci.* 17, 491–496
12. Murtaugh, P.M. (2007) Simplicity and complexity in ecological data analysis. *Ecology* 88, 56–62
13. Krzywinski, M. *et al.* (2014) Points of significance: nested designs. *Nat. Methods* 11, 977–978
14. Colegrave, N. and Ruxton, G.D. (2017) Statistical model specification and power: recommendations on the use of test-qualified pooling in analysis of experimental data. *Proc. Biol. Sci.* 284, 20161850
15. Hurlbert, S.H. (2013) Review of *Biometry*, 4th edn, by R.R. Sokal & F.J. Rohlf. *Limnol. Oceanogr. Bull.* 22, 62–65
16. Lazic, S.E. (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* 11, 5
17. Festing, M.F.W. *et al.* (2002) The Design of Animal Experiments: Reducing the, Use of Animals in Research through Better Experimental Design. In *Laboratory Animal Handbooks* (Vol. 1, 1st edn), Royal Society of Medicine Press
18. Sokal, R.R. and Rolf, F.J. (2012) *Biometry* (4th edn), W.H. Freeman