Anim. Behav., 1996, **52,** 419–422

COMMENTARIES

Sunfish cognition and pseudoreplication

CELIA M. LOMBARDI* & STUART H. HURLBERT†
**Museo Argentino de Ciencias Naturales*
†Department of Biology, San Diego State University

(Received 11 October 1994; initial acceptance 26 January 1995; final acceptance 4 February 1996; MS. number: AS-1145)

In a recent study Dugatkin & Wilson (1992) tested for cognitive abilities in bluegill sunfish. They found that when a focal fish was allowed to forage with different companions, it was able to remember with which ones it had had greatest success and to use this information in future interactions. Moreover, fish seemed to prefer to associate with familiar conspecifics over unfamiliar ones. A number of subsidiary analyses were also reported.

Lamprecht & Hofer (1994) claimed that results presented by Dugatkin and Wilson did not provide unambiguous support for their conclusions. They criticized the authors' way of quantifying preference, their experimental protocol and their statistical analyses. They noted that some of these analyses constituted pseudoreplication (sensu Hurlbert 1984) and, for one data set, suggested an alternative type of analysis.

Dugatkin & Wilson (1994) responded to these criticisms by, among other things, commenting on the nature of pseudoreplication and statistical independence, re-analysing one data set in the manner suggested by Lamprecht & Hofer (1994), and implying that the latter had suggested only 'refinements' in their analyses. They concluded that 'the statistical significance of our results . . . remain[s] unchanged' (page 1461).

We are offering comment on this exchange because it seems likely to leave readers confused as to the nature of pseudoreplication and the seriousness and extent of statistical error in Dugatkin & Wilson (1992). The original critique by Lamprecht & Hofer (1994) was perhaps too gentle. Its effectiveness was also diminished by the editor's liberality in allowing Dugatkin & Wilson (1994) to claim in response that correct re-analysis of their data would not affect the statistical significance of their many results. Our comments will be restricted to strictly statistical matters. Because Dugatkin & Wilson (1992) used a variety of designs and protocols in their study, these cannot be summarized here even briefly; thus the reader should refer to the original article to follow our commentary.

Dugatkin & Wilson (1992) carried out 11 different types of statistical analyses, eight of which were carried out separately for each of the two tanks of fish used. We give the results of these analyses and of a twelfth analysis undertaken at the behest of Lamprecht & Hofer (1994) (Table I) Ten of these analyses represent clear cases of sacrificial pseudoreplication (sensu Hurlbert 1984; Hurlbert & White 1993; see below), and one (analysis 1) appears to represent some simpler miscalculation of error degrees of freedom. As is usually the case, pseudoreplication is most clearly evidenced by an error degrees of freedom (or specified sample size, *N*, in the case of a chi-squared test) that exceeds the number of independent observations made. In the right-hand column of Table I, we indicate the tests and error degrees of freedom that we regard as appropriate to the hypotheses Dugatkin & Wilson (1992) wished to test.

Sacrificial pseudoreplication may be characterized as follows. An experiment is conducted with a satisfactory design involving multiple experimental units per treatment and multiple samples or measurements per experimental unit. The data

Correspondence: C. M. Lombardi, Museo Argentino de Ciencias Naturales, Avenida Angel Gallardo 470, 1405 Buenos Aires, Argentina. S. H. Hurlbert is at Department of Biology, San Diego State University, San Diego, CA92182, U.S.A.

0003–3472/96/020419+04 $18.00/0

Table I. Summary of statistical analyses presented by Dugatkin and Wilson (1992: analyses 1–11; 1994: analysis 12)

Analysis of comparison	Tank	Result†	Proper procedure
1. Correlation: size versus number eaten	1	r^2=0.35, **df=5**, *P*>0.1	Correlation, *df*=4
	2	not given	Correlation, *df*=4
2. Foraging success: alone versus paired	1	*t*=0.479, **df=42?**, *P*>0.6	*t*-test (paired), *df*=5
	2	*t*=0.63, **df=42?**, *P*>0.5	*t*-test (paired), *df*=5
3. Capture time: alone versus paired	1	*t*=2.42, **df=42**, *P*<0.05	*t*-test (paired), *df*=5
	2	*t*=2.45, **df=42?**, *P*<0.05	*t*-test (paired), *df*=5
4. Number items eaten: variation among partners (only for 'red')	1	*F*=3.22, *df*=4, ?, *P*>0.05	*F*-test, *df*=4, 24
	2	*F*=1.29, *df*=4, ?, *P*>0.05	*F*-test, *df*=4, 24
5. Regression: feeding success versus aggression	1	r^2=0.09, **df=209**, *P*<0.001	Regression, *df*=4
	2	r^2=0.049, **df=209**, *P*<0.005	Regression, *df*=4
6. Preference for same individual in both experiments	1	χ^2=11.26, *P*<0.001, **N=60?**	*t*-test (one-sample), *df*=5
	2	χ^2=9.6, *P*<0.001, **N=60?**	*t*-test (one-sample), *df*=5
7. Preference for 'companion in success'	1	χ^2=9.6, *P*<0.005, **N=120?**	*t*-test (one-sample), *df*=5
	2	χ^2=8.13, *P*<0.005, **N=120?**	*t*-test (one-sample), *df*=5
8. Number of items eaten: with chosen versus with not chosen companion	1	*t*=2.76, **df=59**, *P*<0.01	*t*-test (paired), *df*=5
	2	*t*=2.13, **df=59**, *P*<0.05	*t*-test (paired), *df*=5
9. Preference determinants (various)	1	χ^2, *P*>0.25, **N=120?**	?
	2	χ^2, *P*>0.25, **N=120?**	?
10. Number of eaten	1 versus 2	*t*=7.13, **df=208**, *P*<0.0001	*t*-test, *df*=10
11. Preference: familiar versus unfamiliar	1+2	χ^2=32.1, *P*<0.0001, **N=36**	*t*-test (one-sample), *df*=1
12. Alternative analysis for comparison #7 above (suggested by Lamprecht & Hofer 1994)	1+2	Wilcoxon matched-pairs, **1-tailed**, **N=12**, *P*<0.025	*t*-test (one-sample), 2-tailed, *df*=1

†Every value for degrees of freedom (*df*) or *N* given in **bold** signals an incorrect statistical analysis, usually sacrificial pseudoreplication.

are analysed incorrectly, however, by ignoring or sacrificing the information on the nested structure of the data set and treating each sample or measurement as if it represented a separate experimental unit.

Avoiding pseudoreplication is not simply a matter of statistical 'refinement', as implied by Dugatkin & Wilson (1994). The usual consequence of pseudoreplication is underestimation of the *P* value, often by several orders of magnitude (Hurlbert 1984; Machlis et al. 1985; Hurlbert & White 1993; S. H. Hurlbert & C. M. Lombardi, unpublished data). This is again demonstrated by the one case of pseudoreplication (analysis 11) where the authors give sufficient information to allow a correct analysis. That correct analysis entails a one-sample *t*-test of the null hypothesis that the mean (0.972) of the values for the two tanks (17/18=0.944; 18/18=1.000) differs significantly from the value (0.500) dictated by the null hypothesis. This test yields a *P* of 0.04 which, although still significant in the conventional sense, is very different from the '*P*<0.001' that Dugatkin & Wilson obtained by pooling data for the replicate tanks (Table I). For the other pseudoreplicated analyses we still have no sense of what the correct *P* values are.

In responding to their critics, Dugatkin & Wilson (1994) misinterpreted the concept of pseudoreplication. That phenomenon is not a consequence of constraints on design, as implied by their statement that there are situations where one 'must accept a risk of pseudo-replication to do the study at all' (page 1459). Pseudoreplication is simply a type of incorrect statistical analysis. For no study design, however weak, is pseudoreplication an inevitable result.

Along the same line, Dugatkin & Wilson (1994) claimed still not to 'see how [they] could have designed the study differently to reduce further the problem of pseudo-replication' (page 1460). Yet for every design they used there was a correct statistical analysis that would have allowed a valid test of the hypothesis being examined (Table I, right-hand column). Correct analyses would have yielded generally much higher *P* values than those they reported and put many of their conclusions in doubt. We cannot specify in detail here how each re-analysis we suggest (Table I) should be carried out. In general, it would be a matter of condensing the observations to the point where one would have a single mean value (or pair of values) for each fish (analyses 1–10) or for each tank (analyses 11–12). For analyses 6, 7 and 11 this would entail a shift from categorical variables (0, 1) and chi-squared tests to continuous variables (%) and *t*-tests; such a shift can be recommended for numerous papers in *Animal Behaviour* and other journals that have used chi-squared tests (and *G*-tests) in similarly inappropriate manners (e.g. Hurlbert 1984, page 206, Table 6; Hurlbert & White 1993, Figure 4). It is not clear how Dugatkin & Wilson (1992) assessed preference determinants by chi-squared tests (analysis 9), but evidently 120 observations were treated as independent, thus yielding pseudoreplication for each of these analyses. Other approaches are needed here and several possibilities exist.

The Wilcoxon matched-pairs test (analysis 12) of Lamprecht & Hofer (1994) was suggested by them to be applicable to either each group of six fish taken by itself 'or for both groups combined' (page 1458). Its separate application to each group would be valid and an acceptable alternative to the one-sample *t*-test we have proposed as a substitute for analysis 7 (Table I). However, its application to the pooled data for the two groups, which is the application that Dugatkin & Wilson (1994) opted for, represents sacrificial pseudoreplication. The two groups or tanks were initially set up to provide two replicate independent observations, and there is no reason not to treat them as such, as can be done via a one-sample *t*-test with one degree of freedom.

We argue that in using a *one*-tailed Wilcoxon matched-pairs test (Table I, analysis 12) Dugatkin & Wilson (1994) committed an additional error. The 'tailedness' of their other tests (Dugatkin & Wilson 1992) was not specified, so one assumes they were two-tailed. No reason for this difference was given. One-tailed tests are common in the pages of *Animal Behaviour*, perhaps because many statisticians (e.g. Siegel 1956; Sokal & Rohlf 1981) have indicated them to be appropriate if the direction (positive or negative) of a result is predicted. Their use in basic research and in most types of applied research is, however, almost always inappropriate because it makes no allowance for the interest that both the individual investigator and the scientific community usually will have in results that run counter to prediction (Kimmel 1957; Eysenck 1960; Fleiss 1981; Armitage & Berry 1987; Pillemer 1991; Lombardi & Hurlbert, in press).

Generally we do not find it productive to provide a critique of statistical errors in individual papers, especially for the field of animal behaviour where the evidence to date indicates that papers with serious statistical errors outnumber those without such errors (Machlis et al. 1985; Kroodsma 1989; S. H. Hurlbert & C. M. Lombardi, unpublished data). Thus our usual approach to such problems has been to survey and discuss large numbers of cases of statistical error simultaneously (Hurlbert 1984; Hurlbert & White 1993; Lombardi & Hurlbert, in press; S. H. Hurlbert & C. M. Lombardi, unpublished data). In this case, however, a rather large number of persons (authors, colleagues, reviewers, editors and other critics) have participated in an exchange that has left a large number of instances of pseudoreplication not only published but seemingly well defended. This is more an indictment of the system than of the paper under discussion. Weaknesses in the system include the poor training in statistics that most animal behaviourists receive, the high frequency with which gross statistical errors escape filtration or remediation by the review process, and the ease with which the same review process allows authors, when caught with statistical errors, to obscure the constructive message of their critics.

APPENDIX

We applaud the fruitful manner in which Drs Wilson and Dugatkin (1996) have responded, in their reply, to our suggestions concerning analysis of their data on sunfish cognition (Dugatkin & Wilson 1992). Their re-analyses are fine, as far as we can judge, and should serve as helpful models for other researchers. We disagree, however, with several implications and statements on important statistical matters in the latter part of their reply. An amicable correspondence via email has not resolved these differences. Therefore we offer the following caveats or counterclaims so that readers will be informed of the specific nature of the unresolved issues. (1) We do not imply and do not believe that repeated measurements on an individual can 'never . . . be treated as statistically independent' (page 424). There are many situations where they can be so treated. (2) Two flips of a coin cannot be regarded as statistically independent simply because 'the outcome of one flip does not influence the outcome of the next flip' (page 424). The statistical independence of two or more measurements cannot be assessed in the absence of a precise specification of the hypothesis to be tested with those measurements. (3) One avoids pseudoreplication not to be 'conservative [or] to underestimate the number of independent events' (page 424), but rather to make sure that the validity of an analysis is not compromised by unreasonable assumptions, e.g. that two fish or two tanks behave identically. (4) We cannot 'trade' the likelihoods of type I and type II errors (page 424), although we can change their relative magnitudes. Type II errors, of course, are a possibility only for those researchers who interpret high *P* values as indicating 'no effect'. (5) The fixing of alpha serves no purpose in research, because interpretation of *P* values should not be a 'yes/no' matter, and because there are no reasons why the alpha value selected by an author should influence the evaluation of the research by an editor, reviewer or other reader. (6) Our assumption concerning the experiment involving multiple fish in each of two tanks is that fish behaviour patterns in the two tanks were not identical; i.e. that with enough measurements a real 'tank effect' would be demonstrated. That generic assumption is more accurately labelled 'realistic' than 'conservative', both on logical grounds and on the empirical evidence of 'uniformity trials' in many fields of research.

REFERENCES

Armitage, P. & Berry, G. 1987. *Statistical Methods in Medical Research*, 2nd edn. London: Blackwell.

Dugatkin, L. A. & Wilson, D. S. 1992. The prerequisites for strategic behaviour in bluegill sunfish, *Lepomis macrochirus*. *Anim. Behav.*, **44,** 223–230.

Dugatkin, L. A. & Wilson, D. S. 1994. Choice experiments and cognition: a reply to Lamprecht & Hofer. *Anim. Behav.*, **47,** 1459–1461.

Eysenck, H. J. 1960. The concept of statistical significance and the controversy about one-tailed tests. *Psychol. Rev.*, **67,** 269–271.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, 2nd edn. New York: Wiley.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.*, **54,** 187–211.

Hurlbert, S. H. & White, M. D. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. *Bull. mar. Sci.*, **53,** 128–153.

Kimmel, H. D. 1957. Three criteria for the use of one-tailed tests. *Psychol. Bull.*, **54,** 351–353.

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

Lamprecht, J. & Hofer, H. 1994. Cooperation among sunfish: do they have the cognitive abilities? *Anim. Behav.*, **47,** 1457–1458.

Lombardi, C. M. & Hurlbert, S. H. In press. Misprescription and misuse of one-tailed tests. *Ecol. Monogr.*

Machlis, L., Dodd, P. W. D. & Fentress, J. C. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. *Z. Tierpsychol.*, **68,** 201–214.

Pillemer, D. 1991. One- versus two-tailed hypothesis tests in contemporary educational research. *Educ. Res.*, **20,** 13–17.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.

Sokal, R. R. & Rohlf, F. J. 1981. *Biometry*. 2nd edn. San Francisco: Freeman.

Wilson, D. S. & Dugatkin, L. E. 1996. A reply to Lombardi & Hurlbert. *Anim. Behav.*, 1996, **52,** 423–426.