Austral ECOLOGY A Journal of ecology in the Southern Hemisphere

Austral Ecology (2013) **38**, 646–663

Pseudofactorialism, response structures and collective responsibility

STUART H. HURLBERT
Department of Biology, San Diego State University, San Diego, CA 92182-4614, USA (Email: shurlbert@sunstroke.sdsu.edu)

Abstract Pseudofactorialism is defined as 'the invalid statistical analysis that results from the misidentification of two or more response variables as representing different levels of an experimental variable or treatment factor. Most often the invalid analysis consists of use of an (n + 1)-way ANOVA in a situation where two or more n-way ANOVAS would be the appropriate approach'. I and my students examined a total of 1362 papers published from the 1960s to 2009 reporting manipulative experiments, primarily in the field of ecology. The error was present in 7% of these, including 9% of 80 experimental papers examined in 2009 issues of *Ecology* and the *Journal of Animal Ecology*. Key features of 60 cases of pseudofactorialism are tabulated as a basis for discussion of the varied ways and circumstances in which the error can occur. As co-authors, colleagues, editors and anonymous referees and editors who approved them for publication, a total of 459 persons other than the senior authors shared responsibility for these 60 papers. Pseudofactorialism may sometimes be motivated by a desire to test whether different response variables respond in the same way to treatment factors. Proper procedures for doing that are briefly reviewed. A major cause of pseudofactorialism is the widespread failure in statistics texts, primary literature and documentation for statistics software packages to distinguish the three major components of experimental design – *treatment structure, design structure, response structure* – and clearly define key terms such as *experimental unit, evaluation unit, split unit, factorial* and *repeated measures*. A quick way to check for the *possible* presence of the pseudofactorialism is to determine whether the number of valid experimental units in a study is smaller than (i) the error degrees of freedom in a multi-way ANOVA; or (ii) the total number of tallies (N) in a multi-way contingency table. Such situations also can indicate the commission of pseudoreplication, however.

Key words: evaluation unit, experimental unit, factorial experiment, pseudoreplication, repeated measure, split-unit design.

> We are usually ignorant which, out of innumerable possible factors, may prove ultimately to be the most important. . . . We usually have no knowledge that any one factor will exert its effects independently of all others that can be varied. . . . If the investigator in these circumstances, confines his attention to any single factor, we may infer either that he is the unfortunate victim of a doctrinaire theory as to how experimentation should proceed, or that the time, material, or equipment at his disposal are too limited to allow him to give attention to more than one narrow aspect of his problem.
>
> – R.A. Fisher (1935)

> Whenever, in experimental, comparative, or social psychology, a systematic investigation of the primary effects and the interacting effects of a number of experimentally controllable factors is being conducted, the principles of efficient factorial design can be invoked with inestimable benefit.
>
> – R.S. Crutchfield (1938)

> The frequency of this error [pseudofactorialism], rare in the older literature, seems clearly a consequence of the ease with which multi-way ANOVAs can be carried out by canned programs at little cost in time or mental effort to the investigator.
>
> – S.H. Hurlbert and M.D. White (1993)

INTRODUCTION

Pseudofactorialism is an error found, most commonly, in multi-way ANOVAs of experimental data (Hurlbert & White 1993). As a context in which to discuss and understand it, some historical perspective on factorial experiments and use of multi-way analyses of variance (ANOVAs) may be helpful.

Concept of the factorial experiment

A (multi-)factorial experiment is defined as a manipulative experiment for assessing the effects of two or more treatment factors (or experimental variables) on the experimental unit (e.g. Fisher 1935; Yates 1937; Cox 1958; Steel & Torrie 1960, 1980; Snedecor & Cochran 1967; Kirk 1982; Mead 1988; Hinkelmann & Kempthorne 2008).

Accepted for publication October 2012.

doi:10.1111/aec.12010

The logic of Fisher's and Crutchfield's imperatives above must have been apparent to many scientists in many disciplines as soon as formal experimental methods came into use, even if many, being 'unfortunate victims' and/or short of resources, stuck to unifactorial experiments. In the field of psychology in particular, the notion that one should study only one treatment factor at a time, 'the law of the single variable' (Peters & VanVoorhis 1940) or 'single factor fetish' (Crutchfield 1938), was a counter-imperative that still had force with some investigators into the early 1940s (Rucci & Tweney 1980). Yet it was an experimental psychologist, Gustav Fechner, who, in his *Elemente der Psychophysik* (1860), had been perhaps the first scientist to explicitly present and advocate factorial treatment structures for experimentation (Stigler 1986). In his book, *Experimental Agriculture*, James Johnston (1849, *fide* Cochran 1976) also hinted at the desirability of looking at different treatment combinations when effects of two types of fertilizer were to be examined.

Not surprisingly, widespread use of factorial experiments occurred only after Fisher's development of the ANOVA. This provided a sophisticated replacement for earlier, more *ad hoc*, more cumbersome ways of analysing data from such experiments. Given that in the 1920s most scientists were still just beginning to become competent in simpler statistical methodologies, one might have expected Fisher to introduce ANOVA with a simple unifactorial example or at most a 2 × 2 treatment structure and a completely randomized design structure. But no such luck! He first demonstrated ANOVA with data from a factorial agronomic experiment concerning fertilizer effects on potato varieties that had a 12 × 3 × 2 treatment structure and a split-split unit design structure (Fisher & Mackenzie 1923). To make matters worse the design of the experiment was flawed and this first analysis incorrect (Yates & Mather 1963; Box 1978; Cochran 1980; Hurlbert 1984).

Fisher and Mackenzie did not employ the term 'factorial' in describing this experiment. In *Statistical Methods for Research Workers* (Fisher 1925) published 2 years later Fisher still is not using the term factorial and discusses such experiments only briefly, with a reanalysis of his 'potatoes and fertilizers' study as the only example, under the heading, *Analysis of variance into more than two portions*. Soon he again strongly advocates their value referring to them only as 'large and complex experiments' (Fisher 1926; Fisher & Wishart 1930). The bud finally burst open with his book *The Design of Experiments* (Fisher 1935) and its 18-page chapter titled *The factorial design in experimentation*, which was immediately followed by Yates (1935) classic, 42-page monograph on the topic.

Other important works followed. Yates (1937) published an extensive and clear technical guide on the analysis of factorial experiments. ANOVA became a *sine qua non* for writers of the next generation of textbooks. Lindquist (1940), 'the first bona fide statistics text devoted to the application of variance techniques in behavioral research' (Rucci & Tweney 1980), had a 14-page section on factorial designs and their ANOVAs. The complexities of multi-way ANOVAs and tediousness of carrying them out with mechanical calculators notwithstanding, researchers across the natural, behavioural and social sciences became enthusiastic users of factorial designs. In a search of the pre-1940 psychological literature, Rucci and Tweney (1980) found six published studies that used factorial designs; and reported that such designs had become quite standard in psychological research by 1957.

Over the next several decades, the conduct of a multi-way ANOVA, once you had decided which one should be done, became exceedingly easy. First, large electronic computers and IBM cards replaced mechanical calculators. Then the latter were replaced by personal computers and statistical software packages. One consequence of this has been that the incentive for making sure the ANOVA you are conducting is an appropriate one has been greatly reduced. If you do the wrong analysis you at least are not wasting much time compared with what you would have wasted, say, 70 years ago.

Major confusion also has been fostered by the jargon and undefined or carelessly defined terms found in the manuals for many statistical packages. These often exhibit little concordance with the classical terminology of experimental design. To a small extent this is understandable in that the manuals must serve analysis of all sorts of observational studies as well as experimental ones.

But even in documents specifically aimed at analysis of manipulative experiments misunderstandings of the most fundamental sort are promoted. SAS (2007), for example, formally defines 'experiment' as 'a process or study that results in the collection of data'. A bit too inclusive perhaps?! The same document defines, 'An experimental or sampling unit [as] the person or objects that will be studied by the researcher. This is the smallest unit of analysis in the experiment from which data will be collected'. With that advice, one wonders why 100% of the users of SAS/SYSTAT are not committing pseudoreplication or pseudofactorialism (*sensu* Hurlbert 1984, 2009, in press; Hurlbert & White 1993); experimental unit and sampling unit (= observational unit = evaluation unit) have been synonymized! In the current edition of the SAS/SYSTAT User's Guide (SAS 2012, The PLAN Procedure, Example 67.2), an observational study with a nested sampling scheme is presented as an example of a 'factorial experiment' and the factors defining the levels of sampling are called 'treatments'. That example has been in the User's Guide at least since 1990, miseducating all who took it seriously.

doi:10.1111/aec.12010

Adding further challenges for the researcher are new statistical methodologies that are quickly incorporated into software packages and that, understood or not, sometimes develop an attraction as the hottest fashion.

In any event, the growth in the use of multi-way ANOVA that has paralleled the increasing use and complexity of factorial experiments has possibly led to an increase in the proportion of the scientific papers containing serious statistical errors. It seems that increasing numbers of experimenters are using methodologies they have not fully understood or digested. And that an increasing fraction of scientists who serve as editors and manuscript referees for journals lack competence to evaluate the increasingly diverse and complex statistical methodologies they find in manuscripts.

Pseudofactorialism defined

Out of the large educational effort demanded by this situation, the present study takes on only one small task: that of analysing a particular class of statistical error termed *pseudofactorialism* that is found in many papers using multi-way ANOVA.

Pseudofactorialism was first discussed by Hurlbert and White (1993). They defined it as 'the invalid statistical analysis that results from the misidentification of two or more response variables as representing different levels of an experimental variable. Often the invalid analysis consists of use of an (n + 1)-way ANOVA in a situation where two or more n-way ANOVAs would be the appropriate approach'.

In their survey of 95 experimental papers on zooplankton ecology, Hurlbert and White (1993) found the error in five papers, or 28% of the 18 papers that used multi-way ANOVA. They concluded that 'The frequency of this error, rare in the older literature, seems clearly a consequence of the ease with which multi-way ANOVAs can be carried out by canned programs at little cost in time or mental effort to the investigator'. The related error of pseudoreplication (*sensu* Hurlbert 1984, 2009) was found in 41% of these 95 papers.

To put the reader on firmer ground, let us consider a simple example of pseudofactorialism. Schmitt (1987) defined as experimental units six large, widely spaced plots on cobble substrate in a subtidal marine environment. Three were maintained as a control treatment, and to each of the other three additional clams were added to see which if any of three naturally occurring clam predators in the area (lobster, octopus, snail) increased in abundance in response to the augmented clam density. Predator abundances were determined over four successive surveys, and the mean abundance of each predator species determined for each plot. Those means were then plugged in to a two-way ANOVA to test for a clam effect, a predator species effect, and the interaction of the two, using an error mean square with 12 (= 2•3(3 – 1)) degrees of freedom (d.f.) in each case. One correct procedure would have been to conduct a separate one-way ANOVA, with only 4 (= 2(3 – 1)) error d.f., for each predator species. Predator species was only a category of response variable, not a treatment factor. A separate one-way ANOVA for each predator species on each date would also have been a valid approach and probably even more suitable to the author's objectives.

Since 1993 there appear to have been no other surveys of this error, under pseudofactorialism or any other label. At least two works do mention pseudofactorialism in passing, each defining it incorrectly. Underwood (1997) seems to define it as 'arbitrarily putting unrelated experimental treatments into a more complex experiment'. Tindall *et al.* (2007) define it as using 'a large number of t-tests to analyze . . . [a data set because Hurlbert & White believe] that a correction for multiple tests might be useful to maintain an experiment-wide p-value of 0.05 (e.g. Bonferroni or a modified Bonferroni correction)'. Both mischaracterizations are so puzzlingly incorrect that little more can be said. Hurlbert and White (1993) did not discuss adjustments for multiple comparisons. However in a recent review of the topic, Hurlbert and Lombardi (2012), like many others before them, point out the irrational and arbitrary nature of such adjustments for 'multiplicity' and recommend they never be used.

The specific objectives of this review are to document the frequency and varied guises of pseudofactorialism via two new literature surveys, to analyse the conceptual and terminological problems that foster its commission, and to show how easy is its detection and avoidance.

METHODS

The first literature survey was conducted with the help of students in a graduate course in experimental design that I taught at San Diego State University for many years. As an independent project each student evaluated 20 or 25 recent papers selected by him- or herself and reporting the results of one or more manipulative experiments. The usual criteria used by a student for their selection were that they were in a particular journal or on a particular topic, typically in some area of biology. Each student filled out a form detailing the treatment structure, design structure and response structure (as these three aspects have been distinguished by Finney 1955; Urquhart 1981; Hinkelmann & Kempthorne 1994, 2005, 2008; Valiela 2001; Hurlbert & Lombardi 2004; Hurlbert 2009, in press) of each experiment. They then filled out a form detailing whether any of nine common, specific statistical errors were committed in the statistical analysis of the experiment. Pseudofactorialism was one of those errors. For four of the many years in which this exercise was used (1991,

doi:10.1111/aec.12010

1995, 1996, 1998) I tabulated and kept (and did not lose!) the results for the whole class. I personally checked every instance where a student claimed to have found pseudofactorialism. Most of the cases of pseudofactorialism reported in this article were found by these students or those in earlier or later classes.

For a second survey, I examined every experimental paper published in the *Journal of Animal Ecology* in 2009 (Nos. 1–6) and in *Ecology* during the first half of 2009 (Nos. 1–6). For each paper it was determined whether they reported at least one experiment that was analysed by a multi-way ANOVA, either in classical form or via a generalized linear model or, more rarely, a multi-way contingency table. For each such experiment it was determined whether any of these analyses constituted pseudofactorialism.

For each case of pseudofactorialism found in each survey the following information was recorded and tabulated: (i) the number of treatment factors or classification blocking factors (and levels of each) in the experiment; (ii) the number of pseudofactors (and 'levels' of each) used in the ANOVA; (iii) total number of experimental units in the experiment; (iv) the correct error d.f. available for testing for an effect of the first valid treatment factor listed; (v) the error d.f. actually used in that test; (vi) whether the ANOVA treated the pseudofactor as completely crossed with the treatment factors or as nested under them (i.e. masquerading as a split-unit design); (vii) the number of persons collaborating in the publication of the paper, i.e. the number of co-authors, plus named manuscript reviewers, plus three (one editor and two anonymous referees assumed); and (viii) per cent reduction in the critical value of *F*, for alpha = 0.05, caused by the pseudofactorialism-generated, spurious inflation of error d.f.

The term 'pseudofactor' is used as shorthand for a category of response variable that is erroneously treated as a genuine treatment factor in an ANOVA or comparable procedure. This use of the term should not be confused with its more common use to describe a technical device for facilitating the construction and analysis of experiments, especially those with large numbers of treatments (Yates 1936; Monod & Bailey 1992; Hinkelmann & Kempthorne 2005). When a category of response variable is referenced neutrally, without regard to how it is treated statistically, it may be termed simply a *response variable factor*.

As a check on my interpretations and a courtesy to the authors, after a complete draft of this paper was ready it sent to at least one author for each of the 60 papers cited for pseudofactorialism. They were asked to let me know if they found any error in my analysis and were also invited to comment on the manuscript itself. Reminders were sent to initial non-responders and a few months were allowed for this process. Responses were received on 37 of the papers, and almost all authors agreed their analysis that I had cited constituted pseudofactorialism. Two authors thought they had committed no error and stood firm even after my further explanation. A few said they had no time to look at the manuscript, or said they would get back to me with comments but did not do so. Almost without exception the responses were constructive and good-natured. Many improved the paper. This process was followed to help forestall needless post-publication flurries of complaint or rebuttal based on misunderstandings.

RESULTS

Frequency of occurrence

The frequency of papers containing one or more cases of pseudofactorialism in the two new surveys and in that of Hurlbert and White (1993) is shown in Table 1. When expressed as a fraction of the total number of experimental papers examined, frequency of pseudoreplication ranged from 5% to 9%. Those estimates can

Table 1. Frequency of pseudofactorialism as determined in several surveys

	Number of experimental papers			Frequency of pseudofactorialism (%)	
Survey	Total No. analysed (n_T)	No. using multi-way ANOVA† (n_M)	No. committing pseudofactorialism (n_P)	$100 \bullet n_P/n_T$	$100 \bullet n_P/n_M$
Hurlbert and White (1993)	95	18	4	5	28
SDSU Experimental Design students					
1991	380	?	27	7	?
1995	380	?	35	9	?
1996	260	?	14	5	?
1998	167	?	8	5	?
All years	1187	?	84	7	?
Current ecological literature (2009)					
Ecology, 90(1–6)	47	34	4	9	12
Journal of Animal Ecology, 78(1–6)	33	22	3	9	14
Combined journals	80	56	7	9	12
All surveys	1362	?	95	7	?

†Including general linear model and non-parametric equivalents. Question marks indicate the value is unknown. SDSU, San Diego State University.

be regarded as unbiased for the specific, mostly biological literatures examined as in each case the experimental papers were selected for examination without any evident bias favouring or disfavouring papers likely to exhibit the problem. These estimated frequencies may underestimate the true frequencies, however, if we assume that students missed some cases of pseudofactorialism in these 1187 papers they examined.

When only papers employing multi-way ANOVAs are considered, the observed frequency of pseudofactorialism ranged from 12% to 28%. This frequency could not be calculated for the student surveys, as the students were not asked to record the number of experimental papers in their paper sets that used multi-way ANOVA or multi-way contingency tables.

Sixty variations on a theme

Salient features of 60 cases of pseudofactorialism found in the latter two surveys are documented in the *Supplementary information* (SI). Of these cases, 49 come from the student surveys, seven from my examination of 2009 papers in *Ecology* and *Journal of Animal Ecology*, and four from other sources. Many of the papers found by the students are not tabulated here, for the following reasons.

First, it was decided in the interests of fairness, given the widespread nature of his problem, not to allow any person to appear more than once in this list as a senior author.

Second, some papers seemed to commit pseudofactorialism but gave such scant information on statistical procedures the information desired for the SI was not ascertainable with any confidence. For example, Herman *et al*. (1986) conducted an experiment on the effects of the herbicide atrazine on attached algae in limnocorrals (polyethylene enclosures in a lake), applying atrazine to three limnocorrals and keeping three as controls. Abundance of attached algae was measured at four different depths. The authors stated, 'The ANOVA procedure was used in the analysis . . . with sampling depth considered a separate factor'. We inferred that a two-way ANOVA was used that treated depth as a treatment factor completely crossed with atrazine level, generating 16 (=2•4(3 – 1)) error d.f. That would be a clear case of pseudofactorialism. However, certainty was not possible as no ANOVA table was given or error d.f. mentioned.

A similar case (Hambright 1994) involving temperature measurements made at five depths in experimental ponds with and without fish actually presented a full ANOVA table, with depth treated as an experimental variable fully crossed with 'fish'. So that paper *is* listed in the SI. As usual, no good deed – clarity, in this case – goes unpunished. Just as when measurements on experimental units are repeated over time, there are a variety of analytical options and we briefly review these in Discussion. These might involve using the temperature profile to calculate the mean temperature, the slope of the vertical temperature gradient, or the heat gain per pond. Each such composite response variable is just another way of representing the effect of fish on one thermal property of the ponds.

The main patterns

As is common in many types of biological experiments, especially field experiments, the total number of experimental units in most studies was small: only seven of the 60 experiments analysed had employed more than 40 experimental units (SI, column D). The limited power of such experiments incites a certain concupiscence with respect to additional error d.f., however they might be obtained.

One indicator of possible pseudofactorialism is when the number of error d.f. used in testing for treatment effects exceeds the total number of experimental units in the study. In simple situations (e.g. Lauenroth *et al.* 1978; Schmitt 1987; Hambright 1994: SI), the pseudofactorialism produces spurious error d.f. equal to the true error d.f. times the number of 'levels' of each pseudofactor.

Pseudoreplication of various types, however, also yields error d.f. that exceed the number of experimental units in the experiment. It is an error that can occur by itself, but it can also co-occur with pseudofactorialism. Pseudoreplication occurred in 16 of the 60 analyses with pseudofactorialism (SI, column F, footnote e). That frequency jumps to 27 out of 60 analyses if we include those cases where temporal pseudoreplication resulted from treating time, in a multi-way ANOVA, as a treatment factor completely crossed with treatment factors in experiments with repeated measures response structures. Such cases represent, of course, only one of several ways in which temporal pseudoreplication is committed.

In cases of pseudofactorialism, the response variable factor is most commonly treated analytically as a factor completely crossed with the genuine treatment factors in the experiment. In four of the 60 analyses, however, the response variable factor was treated as a subunit treatment factor in a split-unit design structure (SI, column G). Note that in these cases, the error d.f. for assessing the effect of the whole unit treatment factor(s) are *not* spuriously inflated. In such cases tests for whole unit treatment factors are likely to be little or not at all affected. However, the error d.f. for testing of the putative 'subunit factor' and its 'interaction' with the whole unit treatment factor will be spuriously inflated to a degree that is unknown but dependent on the magnitude of the correlations among the response variables within experimental units. There are no

doi:10.1111/aec.12010

Table 2. Per cent reduction (R) in the critical F values needed for a P value to decrease from 0.10 to 0.05 or 0.01, as a function of numerator and denominator (error) degrees of freedom (d.f.)

	Numerator d.f., $a = 0.01$			Numerator d.f., $a = 0.05$		
Denominator d.f.	1	2	4	1	2	4
1	99	*	*	75	*	*
2	92	91	*	51	53	*
4	79	76	74	41	38	33
8	69	64	60	35	30	27
16	64	57	51	32	26	23
∞	59	50	42	29	23	19

*This combination of treatment and error d.f. not possible for a comparison of two or more treatments. R is calculated as $R = (1 - F_{0.10}/F_a)$ where a is 0.05 or 0.01 and F is the ratio of the mean square for the treatment factor to the mean square for error that will yield a P value of 0.10, 0.05 or 0.01 respectively.

grounds for regarding the response variable factor as a treatment factor of any sort. The several sources of confusion here are addressed in Discussion. These concern the definitions of experimental unit, split-unit design structures, and repeated measures response structures. Confusion over these lies at the bottom of all sorts of misery in the literature in addition to the problem of pseudofactorialism itself. The disdain of many professional statisticians for use of clear, standardized terminologies often is to blame (Hurlbert, in press).

When the response variable factor is treated as a treatment factor fully crossed with the genuine treatment factors, analysis of the latter is compromised by the spurious inflation of the error d.f. employed in the analysis. This will tend to bias P values downward. How much downward will be a function of how much the incorrect analysis has altered sums of squares, a matter impossible to discern in most papers.

Nevertheless, as an index of bias due to incorrect error d.f. alone, we can determine how much the critical F value is reduced for any specified alpha. The per cent reduction is quite variable from one case of pseudofactorialism to another but, for alpha = 0.05, that reduction ranged from 1% to 97% in the cases documented (SI, column I). Such reductions may or may not affect decisions. For example, for neither a P value of 0.80 nor one of 0.01 is a 50% reduction in P likely to change conclusions regarding the existence of a treatment effect.

When P values are closer to alpha values used by paleoFisherians and Neyman-Pearsonians (see Hurlbert & Lombardi 2009) to dichotomize results into 'significant' and 'not significant', one can calculate, for any given treatment (numerator) and error (denominator) d.f., what reduction in the F value for the treatment factor is required to lower a P value of, say, 0.10 to, say, 0.05 or 0.01 (Table 2). As shown, a lowering to 0.05 generally requires a 20–50% reduction in the F value, while a lowering to 0.01 requires a 40–90% reduction.

The total number of collaborators involved in publication of these 60 papers is 459 (SI, column H). Add on the 60 senior authors and we have 519 persons who found pseudofactorialism to be acceptable procedure – or who perhaps reviewed statistical analyses in a manuscript only superficially. Very few of these 519 persons are professional statisticians. Good applied statisticians, busy people all, often are reluctant manuscript referees for disciplinary journals outside their own field. Presumably they feel that tutorials on basic statistics simply should not be their responsibility. A large but uncounted number of the 519 collaborators are or have been editors or editorial board members of scientific journals; they and many others of the 519 also have served as referees for other scientific journals. Is the fox is in the henhouse, so to speak, and the farmer on vacation?

How do we get out of this positive feedback loop? Referring to the high frequency of various types of statistical malpractice in experimental papers on zooplankton ecology, we noted 'the tremendous burden it places on conscientious editors, reviewers, thesis advisors, and statisticians. The morass of incorrect statistical analyses in the literature creates a Sisyphean task for them. It provides an abundance of negative models that continually are undoing their instructional efforts' (Hurlbert & White 1993). The drum will not be beaten further on this point.

A caveat

It should be emphasized that the citing of a case of pseudofactorialism in a paper is in no way an indictment of the overall quality of a paper. Some responding authors wanted me to make this clear with respect to their paper and perhaps even mention that their conclusions would have been unaffected by a corrected statistical analysis. But that is simply not feasible. And evaluation of the overall quality of even a single paper would require a difficult and often

doi:10.1111/aec.12010

subjective evaluation of a dozen factors, including other statistical analyses presented in the paper, that are not germane to the topic of this review.

DISCUSSION

Pseudofactorialism is just another simple error that scientists are deceived into committing (and recommending!) by the long-standing terminological chaos in experimental design and statistics generally, as I have commented elsewhere (Hurlbert 2009, in press). That chaos permeates most statistics textbooks, the manuals for statistical software, the primary literature of statistics itself and that of all the natural and social sciences.

In the following discussion, the tabulated cases of pseudofactorialism (SI) are used to illustrate more concretely some key terminological and technical issues. Many statisticians see no need for making their terminologies as precise as their symbolic notation and as consistent across disciplines as possible (Hurlbert 2009, in press). Such statisticians thus may find this discussion tedious, irrelevant and with insufficient symbolic notation and mathematics to merit their attention. No apology is offered. They are wished 'Good luck!' in their continued attempts to communicate with non-statisticians.

Single-unit designs

Some confusion derives from the fact that we have never had an umbrella term for all those design structures that are not split-unit. The term *single-unit designs* may serve for that purpose. This would refer to all experiments where the experimental unit is defined at only one scale.

Thus there are two primary dimensions to the design structure of any experiment. The first is whether the design is single-unit or split-unit. The second is whether completely randomized, randomized block or systematic assignment of treatments to experimental units is used. This is specified separately for each level in split-unit, split-split unit etc. design structures. If we adhere to the classical definition of 'split unit', then only factorial experiments can have split-unit design structures.

'Factor' and 'factorial'

The word *factor* has many usages and connotations, and these are hardly less diverse in statistics than in the English language generally. Many statistics texts use it, with a qualifier, for a very limited number of formal terms, such as treatment factor, blocking factor and so on. A few define many more such formal terms. For example, Lee (1975) recognizes: *subject factor, error factor, treatment factor, blocking factor, group factor, trial factor, unit factor, fixed factor, random factor, qualitative factor, quantitative factor, phantom factor, control factor, nest factor, terminal factor* and *constant factor*.

It is evident that the term *factorial* derives from *factor*. So a neophyte might be excused from thinking that every experiment is 'factorial' in some sense. However, most statisticians have long paid at least lip service to the definition of factorial experiment given at the beginning of this article, adopting *factorial* as shorthand for *multi-factorial*. The simplest factorial treatment structure involves two treatment factors.

Yet many textbooks, confusing the distinctions among *treatment, design* and *response structures*, misuse the term *factorial*. Raktoe *et al*. (1981) aim 'to present a systematic and unified approach to the subject of factorial designs'; but they omit all consideration of factorial experiments with split-unit design structures. They are willing, however, to label as 'factorial' an experiment with a single treatment factor and a randomized block design structure, just as did Federer (1975).

Winer *et al*. (1991) describe as a '2 × 2 × 3 factorial experiment' what is in fact a 2 × 3 factorial experiment with a completely randomized split-unit design structure. Winer *et al*. characterize it as a three-factor design by treating 'hospital' (= the (whole) experimental unit) as a treatment factor. They suggest (p. 361) that in a 'bona fide factorial experiment' the design must be a single-unit one. A simple 2 × 2 factorial experiment with a split-unit design structure is not 'bona fide' in their unconventional language.

Sokal and Rohlf (1995) reflect the same confusion, routinely mixing ANOVA terminology and design terminology. They refer to 'factorial analysis of variance' (p. 369) and claim that, 'Split plot designs are quite often incorrectly analyzed as factorial ANOVAS' (p. 386). The latter point is a useful and valid one only if 'factorial ANOVAS' is replaced with 'single-unit designs'.

Underwood (1997) describes a simple unifactorial experiment with a standard randomized block design and labels it a 'factorial experiment'. But a blocking factor is not a treatment factor, and mere use of a two-way ANOVA is not sufficient to justify use of the label 'factorial'.

Quinn and Keough (2002) repeat the error of Raktoe *et al*. (1981) and Winer *et al*. (1991) and designate as 'factorial' only experiments with single-unit design structures.

As they were originally defined, 'Factorial experiments are experiments which include all combinations of several different sets of treatments or "factors" ' regardless of the design structure – completely randomized or randomized block, split-unit or single-unit – that is employed (Yates 1937; see also Cochran & Cox

doi:10.1111/aec.12010

1950, 1957; Federer 1955; Finney 1955; Cox 1958; Steel & Torrie 1960; Mead 1988; Hinkelmann & Kempthorne 1994, 2008; Steel *et al.* 1997). Our usage should continue to conform to that long-established definition. We can acknowledge the validity of the term *fractional factorial designs* or, better, *fractional factorial treatment structures* for situations where one or more of the possible combinations of employed treatment levels is absent from an experiment.

Metric-interaction mismatch

The great advantage of factorial experiments is that they permit assessment of treatment factor interactions. For a multi-way ANOVA to accomplish this meaningfully, it is critical that the logically appropriate metric for effect size be used. This is a matter that has been ignored in the great majority of the 60 papers cited for pseudofactorialism and is a problem in addition to that of the pseudofactorialism itself. It merits comment here because for many authors, the desire to test for interaction between a genuine treatment factor and a response variable factor was evidently one incentive to commit pseudofactorialism.

As discussed most thoroughly by Mead (1988), for many response variables, perhaps most biological ones, 'the natural side on which most measurements should be made is a log scale . . . it should be assumed that data for continuous variables should be transformed to a log scales unless there is good reason to believe that this is not necessary'. Equivalently, one can say that multiplicative rather than additive models are usually appropriate for such response variables.

Hurlbert and White (1993) also noted that usually ' "magnitude of effect" is most appropriately and meaningfully measured as *per cent* change rather than as *absolute* change . . . [and that] assessments of factor interaction in multi-way ANOVAs are meaningful only when the data are log-transformed prior to analysis'. Failure to so transform, they termed 'metric-interaction mismatch'. Consider a case where 'species' is a true treatment factor in an experiment on effects of nutrient additions on population densities of different species of algae, with each species in a separate set of aquaria (experimental units). If we are interested in treatment interactions, we would want to test with a two-way ANOVA whether each species showed the same per cent change in response to nutrient additions, not whether the absolute increment or decrement was the same for all species. Log-transformation or a multiplicative model does this.

If we simply have a unifactorial experiment on effects of nutrient additions on aquaria each containing many algal species we *might* still have an interest in assessing the null hypothesis that all species responded the same way. However, this would have to be done via an analysis that did not treat species as a treatment factor but still did reflect definition of effect size as per cent change.

Experimental units *versus* evaluation units

In any experiment it is evident that, if ANOVA is used, the error d.f. available for testing for a treatment effect is a function of the number of replicate experimental units used per treatment (or treatment combination) and the design structure of the experiment. The spurious additional d.f. resulting from treating the pseudofactor as crossed with the real treatment factor(s) can be regarded as reflecting failure to properly specify the experimental unit in an experiment and its analysis.

Thus in the experiment of Schmitt (1987) mentioned earlier, the three predator populations in each plot were each treated as an experimental unit rather than as the evaluation units that they were. In Hambright (1994), the water masses defined at five depths in each pond were likewise mistakenly treated as experimental units rather than as evaluation units. An evaluation unit is defined as 'that element of an experimental unit on which an individual measurement is made' (Hurlbert & White 1993, after Urquhart 1981).

Such confusion is the same as that which underlies the commission of pseudoreplication (Hurlbert 1984, 2009). The distinction between the pseudoreplication and pseudofactorialism resides in that in the former the *same* response variable has been measured in each experimental unit on multiple evaluation units of the *same* type (e.g. multiple samples of the same species in the experimental unit) whereas in the latter, *different* response variables have been measured on, usually, *different* types of evaluation units (e.g. samples, multiple or not, of different species in the experimental unit).

Great inflation of error d.f. over the correct error d.f. can occur when both pseudofactorialism and pseudoreplication occur in the same analysis, especially if there is more than one pseudofactor involved. For example, that inflation was 315-fold in de la Cruz *et al.* (1989) and 508-fold in Diffendorfer *et al.* (1995) (SI, columns E and F). In two unifactorial experiments (Gardner *et al.* 1995; Howe & Marshall 2002) the inflation was essentially infinite as there were zero error d.f. for a correct analysis because all treatments were unreplicated.

In rare situations, pseudofactorialism causes no inflation of error d.f. This was the case in Shakarad *et al.* (2001), where there were only two levels of the pseudofactor (sex) and where effect of the treatment factor ('selection') in a unifactorial randomized complete block design was tested for using the block × treatment × pseudofactor interaction mean square

doi:10.1111/aec.12010

rather just the block × treatment mean square. The experimental unit in this study was a fly population; the authors effectively regarded the males in a population as constituting one experimental unit and the females as constituting another.

Split-unit designs: what they are not

In four of the 18 studies where there were multiple species present in each experimental unit and the response variable factor 'species' was treated as if it were a subunit treatment factor in a split-unit design rather than, as in the other 14 cases, a factor crossed with the valid treatment factors in a single-unit design (SI, column G). In two of these four (Mamalos *et al.* 1995; Finzi *et al.* 2001) species was explicitly labelled a subplot (= subunit) factor, in Preen (1995) it was labelled a repeated measures factor, and in Arnone & Körner 1995) it went unlabelled.

These authors recognized that, at least in some sense, species was a factor 'nested' under the valid treatment factor(s). It was, but only as a label for a set of response variables or evaluation units of qualitatively different types, for example individuals or populations of different species.

The authors also may have been told that to carry out a separate statistical analysis on each species was somehow improper or invalid and that they should include all species in an overall ANOVA of some sort. That sort of bad advice is indeed common.

In a valid split-unit design, the subunits must receive the experimental treatments independently and have the same physical independence from each other as do the whole units – as must the experimental units in any kind of experiment. What transpires on one subunit cannot be allowed to influence what transpires on another (Cox 1958; Federer 1975; Mead 1988; Wiley 2003; Kozlov & Hurlbert 2006; Hurlbert 2009, in press). Otherwise biased estimates of treatment effects and standard errors are highly probable.

Federer (1986) once made a statement frightening in its implications for the quality of much past and ongoing experimental work in many fields:

> The assumption of [physical] independence among experimental and/or sampling units is an untenable one for many situations. Since statistical theory is considerably easier for independent observations, we statisticians hide in our i.i.d. world. In many experiments in agriculture, biology, medicine, ecology and other areas, the experimental units are not independent. There may be competition, memory, carry-over, etc. between adjacent units. It is felt that the phenomenon of competition between adjacent units is present in most agricultural, biological and ecological experiments. Statistical analyses for independent observations is universal but incorrect for these data. Thus, incorrect or inappropriate analyses are being conducted on thousands of experiments each year.

Indeed, large numbers of inappropriate statistical analyses are being published, but there is no evidence that lack of physical independence of experimental units has been a major contributor to the problem. *Mis-identification* of the experimental unit, on the other hand, has been.

Federer's statement can be understood only in the light of his unusual concept of the experimental unit, as evidenced by many examples in his later book (Federer & King 2007). There, if a response variable is measured on the same experimental unit on two successive occasions or if two different response variables are measured on the same experimental unit, the authors consider that they are dealing with two different experimental units rather than, as is the case, just two different evaluation units. Federer and King (2007), for example, describe a randomized complete block design experiment testing effects of five treatments on strawberry production. At harvest time the strawberries are sorted into four different quality categories and weighed. The authors treat these categories as 'split plot treatments', implying that each of the four sets of strawberries obtained constitutes or represents a separate experimental unit. In another example (p. 79), 10 different mixed species hay crops are each grown on six plots; on harvest, the crop for a plot is sorted into weeds, legumes, and grass 'to form the split plot treatments'. The potential for statistical correlations among the four strawberry response variables or among the three plant type response variables is obvious; but it has no bearing on whether or not the actual and only experimental units in these studies, the plots, were physically independent.

Physical independence is a necessary but not sufficient condition for independence of errors (statistical independence). The latter is achieved by also employing randomization in the assignment of treatments to experimental units, sub- or whole-. In an ecological field experiment, all the plots could be physically independent of each other. But if all those in the northern end of the field were assigned to treatment A and all those in the southern end to treatment B, independence of errors would not likely obtain.

For the four studies in SI that are cited above and that treated species as a subunit factor, the multiple species populations in each experimental unit are unlikely to meet this standard criterion for independent subunits. Arnone and Körner (1995) planted 77 individual plants representing seven species (including three tree species) spatially intermingled with each other on 6.7-m^2 plots and determined growth of each species after 530 days. The potential for the seven

doi:10.1111/aec.12010

species influencing each others' growth in various ways would seem high. At the end of the experiment, the leaf area index (m^2 leaf surface per square metre of ground) in the chambers was approximately 4.

Preen (1995) studied in a unifactorial experiment the effects of simulated dugong grazing on plots in natural seagrass beds containing two spatially intermixed species of sea grass. Finzi *et al.* (2001) studied in a unifactorial experiment effects of elevated CO_2 levels on five spatially intermixed plant species occurring in their large outdoor chambers. Mamalos *et al.* (1995) studied in a factorial experiment the effects of nitrogen and phosphorus additions to field plots on root activity of five spatially intermixed plant species occurring on their field plots. In all these cases, too, there can be *no* grounds for considering 'species' to be a subunit treatment factor. As in all those studies where it was considered a treatment factor crossed with the valid treatment factors in a single-unit design, 'species' is just a category of response variable, a response variable factor.

An experiment by Gulmon (1992) gives insight into when 'species' might or might not be validly treated as a subunit treatment factor. This experiment was designed to examine the effect of date of first watering (four dates) on germination rate for seven plant species. Four 25 cm × 50 cm flats of soil were set up for each watering date, and seeds for the seven species were sown in seven parallel rows within each flat. The total number of seeds sown per flat was 2095. Because Gulmon's statistical analysis treated species as completely crossed with watering date in a single unit design, I considered this an example of pseudofactorialism (SI, column G).

But would this judgement have been justified if Gulmon had analysed this as a split-unit design? I would say no. After all, the species were essentially in spatially distinct subplots (= monospecific rows) just as in a standard agricultural split-unit experiment. Although the rows were separated from each other by only a few cm and the overall seed density quite high, the potential for events on one row affecting what transpired on other rows was low *given the nature of the experiment*. Specifically, every time a seed germinated the event was tallied and the seedling was removed from the flat. Now if seedlings had not been plucked and the plants had been allowed to grow to maturity so as to determine effects of watering date on biomass at plant maturity, then each flat would have turned into a tangle of vegetation – and the necessary physical independence of the supposed subunits would have been lost.

So, as it was, Gulmon analytically treated a valid split-unit design as a single-unit one. Potvin (1993) has called that 'the most frequent and . . . damaging error occurring in the statistical analysis of greenhouse or growth chamber experiments'. It spuriously inflates error d.f and deflates estimates of error. Potvin presents and analyses a hypothetical data set for an experiment with a split-unit design with two levels for the whole unit treatment factor (A), six levels for the subunit treatment factor (B), and two whole (experimental) units per level of A. Properly done, the analysis yields a $P = 0.025$ for the test for an effect of A. When I analyse these data as if they came from a single-unit design, the test for an effect of A yields a $P = 0.00000021$. This is a consequence of the error d.f. having jumped from 2 in the first case to 12 in the second. P values for tests for factor B and the A × B interaction were much less affected as the error d.f. jumped only from 10 to 12 in going to the incorrect analysis.

Statisticians themselves have long been and continue to be a major source of confusion on these matters (Hurlbert, in press). Federer (1977) labels as a randomized complete block split-unit design one where the 'whole unit' is a pen of piglets in which each piglet gets a different nutritional supplement, the putative subunit factor. In his text on experimental design, Gill (1978) presents an experiment to compare the growth rates of steers of s different size classes on f different feed types, with s steers (one from each size class) maintained in each of p pens per feed type. Gill refers to this as a randomized complete blocks design, with the pen as block, and then analyses it as a split-unit design. Neither example conforms to a split-unit design, as there is high potential for the individual piglets or steers in each pen – each supposedly representing an independent subunit – to influence each other's growth. They are actually experiments with completely randomized single-unit design structures and multiple response variables.

Worse yet, as discussed in Hurlbert (in press), two recent statistics texts (Federer & King 2007, as discussed above, and Casella 2008) present many more examples where evaluation units within the same experimental unit are taken to represent physically independent subunits in a split-unit design. Casella (2008) gives one example where plants are exposed to different levels of ozone and defines the upper and lower portions (where measurements are made) of each plant as subunits in a split-unit design. When successive measurements are made on an experimental unit over time, Casella (p. 208) refers to time as a treatment factor and to the experimental unit on any given monitoring date as a subunit ('split plot').

Intersection with temporal pseudoreplication

The reader will note that in 20 of the 60 studies, time or some correlate of it is a or the pseudofactor (SI). The label used for this factor varied among studies and included *date, year, time, instar, days, season, cohort* and *period*. In all cases, these correspond to measurements made on the same experimental units under the same

doi:10.1111/aec.12010

treatments at successive points in time. They represent *repeated measures*, as classically defined.

In some cases, repeated measurements serve simply to document how effect sizes change with distance in time from the beginning of the experiment. In other cases, the interest is in temporal change in response variables that is associated with particular environmental conditions that are changing with time, for example night *versus* day, spring *versus* summer *versus* autumn, high tide *versus* low tide.

Temporal pseudoreplication is defined as an analysis that treats measurements made successively over time on one or more experimental units as if each measurement represented or came from an independent experimental unit (Hurlbert 1984, 2009; Hurlbert & White 1993; Hurlbert & Lombardi 2004). This error can be committed with many different statistical procedures, for example *t*-tests, ANOVAs, chi-squared tests, non-parametric tests, randomization tests etc. In the special case where it is committed via a multi-way ANOVA we have chosen to classify the error as both pseudofactorialism *and* temporal pseudoreplication. It is then a unique type of pseudofactorialism that results from repeated measurements made *on the same type* of evaluation unit, or example the same species population, the same soil constituent, the same physiological variable etc. In that regard it differs from other types of pseudofactorialism.

Most authors in the SI who treated time as a pseudofactor gave no evidence of sensing any potential problem with doing so. McGrath *et al.* (2009) were an interesting exception. They monitored on nine dates the number of birds visiting 17 pairs (blocks) of mesquite trees, one member of each pair having had most of its flowers removed by hand. As the birds were migrants that typically remained in the study area for less than 2 days, the authors opined that they 'were able to resample experimental pairs [of trees] every three days without fear of pseudoreplication'. In addition to mistakenly implying that that pseudoreplication is an error of design rather than one of analysis and interpretation, the authors were also wrongly implying that so long as different *evaluation* units (birds) were monitored on each date, it did not matter that they were in the same *experimental* unit (tree). They stated that their generalized linear model analysis treated 'date as a covariate', but they had 128 error d.f. in the test for an effect of flower reduction on bird numbers. That suggests that 'date' was additionally treated as crossed with pair (block) and treatment.

A similar rationale was invoked by Hairston (1986) in the analysis of his classic field experiment on competition and predation among salamander species. This involved removing particular salamander species from particular sets of plots and following population change in remaining species. At night he and his students visually assessed numbers of salamanders active on the surface of his large experimental plots on 32 monitoring dates over 4 years. Dividing the whole duration of the experiment into five periods, he applied two types of ANOVAs to the data for each period (four to nine monitoring dates per period). One was a repeated measures ANOVA (details unspecified). The other was a one-way ANOVA that treated repeated censuses on a given plot as if they represented independent experimental units and thus qualified as temporal pseudoreplication. The latter incorrect analysis not surprisingly always yielded lower *P* values for treatment effects, at least as far as can be inferred from Hairston's tables.

Hairston acknowledged that his one-way ANOVAs 'technically violate the assumption of independence'. He claimed, however, that 'the relatively small proportion of all specimens present seen during one visit reduces the importance of the violation'. He thought, like McGrath *et al.* (2009), that as long as he rarely or never saw the same individual salamander more than once, he could treat the successive counts as if they came from different independent experimental units. The criterion is irrelevant. Whether a response variable is measured on the same or different evaluation units (individuals, materials, locations, components etc.) within a given experimental unit, the successive measurements for that experimental unit cannot be assumed to be independent of or uncorrelated with each other.

Split units in time?

None of the cited studies having a repeated measures response structure was described by its authors as having a split-unit design or analysed with a standard split-unit ANOVA with time treated as the 'subunit factor'. It is common, however, that simple repeated measures response structures are described as having 'split units in time', simply because a standard split-unit ANOVA is mathematically identical to a standard repeated measures ANOVA, as long as the latter does not employ Huynh and Feldt (1976), Greenhouse and Geisser (1959) or other corrections to the error d.f. (The unfortunate use of 'repeated measures' as a label for cross-over designs or analyses thereof is common in some disciplines, e.g. psychology, but conflicts with classical terminology and is to be discouraged.)

Cochran (1939) noted the 'analogy' between the analysis for a split-unit design structure and that for a repeated measures response structure but also warned that 'the analogy should not . . . be carried too far . . .' Federer (1955) threw caution to the wind and suggested that 'since the design and analytical features [of a repeated measures response structure] are of the same nature as split plot designs, there is little reason to set up a separate category for these designs'. Steel and Torrie (1960) were one of the earliest texts to have

doi:10.1111/aec.12010

a chapter subsection titled 'Split plots in time'. Over the last half century many texts have made similar use of the label, frequently referring to time as a 'treatment factor' and the successive measurements made on one experimental unit as representing different experimental units. No good has come of this blurring of the distinction between design structure and response structure. The concept of 'splits in time' was introduced as an early attempt to deal with experiments involving repeated harvests from the same crops in a given season (e.g. alfalfa) or over a series of years (e.g. perennial species such as citrus) when computing facilities were primitive. Since that time multivariate-based approaches have become available to deal with such situations when the researcher is not content to simply conduct date-by-date analyses, and the conceptual distinction between design structure and response structure has been understood more widely. Application of the label 'split unit' or 'split plot' to anything other than true split-unit design structures should now be regarded as an anachronism to be avoided.

Pseudofactorialism in contingency tables

Contingency tables and corresponding procedures such as chi-squared or *G* tests are not well suited to analysis of experiments except where there is a single response measured on each experimental unit and this is treated as a categorical variable.

As discussed, when ANOVA or other parametric procedures are used to analyse an experiment, the commission of pseudofactorialism and/or pseudoreplication is typically indicated by the error d.f. exceeding the number of true experimental units used in the experiment.

With contingency tables the presence of either of these errors is typically signalled by *N*, the total number of tallies in the table, exceeding the number of experimental units in the experiment. The number of d.f. in a chi-squared or *G* test reflects only the number of levels defined for each categorical response variable, not the number of experimental units in the study. The high frequency of pseudoreplication in contingency table analyses, even in statistics texts, has long been the subject of critical review (Lewis & Burke 1949; Wolins 1982; Hurlbert 1984, 2009; Hurlbert & White 1993; Wickens 1993; Lombardi & Hurlbert 1996; Hurlbert & Meikle 2003).

In the present set of 60 papers, only two presented examples of pseudofactorialism in the context of a contingency table analysis. One of the experiments conducted by Metaxas and Young (1998) looked at the effects of diet (four types) and algal density (four levels) on vertical distribution of sea urchin larvae in cylindrical, salinity-stratified water columns (four depth strata defined), with each of the 16 treatment combinations replicated four times (SI). The experiment thus involved a total of 64 experimental units. For each replication, '100–200 larvae' were introduced into the cylinder, and after 30 min the number of larvae present in each depth stratum was estimated.

Their statistical analysis consisted initially of applying a *G* test to a four-way contingency table with dimensions corresponding to diet, algal density, 'replicate', and stratum; the response variable was implied to be number of larvae. The d.f. were specified to be 144. *N* was not specified but presumably was more than 25 600 (= 4 × 4 × 4 × 4 × '100–200'), well above the number of experimental units in the experiment. In treating 'stratum' as a treatment factor this analysis commits pseudofactorialism. In treating the individual larvae (apparently) as the experimental units, the *G* test commits pseudoreplication. In treating 'replicate' as a treatment factor because it tested for and found 'heterogeneity' among replicates, it committed a sin yet to be named!

Subsequent treatment of this data set entailed breaking it down into and testing three-way and two-way contingency tables, with somewhat complex rationales given for doing so. Most of these analyses also contained pseudofactorialism and pseudoreplication. Metaxas and Young cite Sokal and Rohlf (1981) as one of the guides they relied on for their procedures. On page 750 of that book (page 746 of its 1995 edition), Sokal and Rohlf present a data set on fruit fly larvae very similar in structure to that of Metaxas and Young. They demonstrate with it how to apply log-linear models to a three-way contingency table. Sokal and Rohlf are vague as to the design of the experiment that generated their fruit fly data, but their analysis also seems to represent both pseudofactorialism and pseudoreplication.

Distrust authority. Say it loud and say it clear.

A second, simpler case, not referenced in the SI, is provided by another analysis in Schmitt's (1987) study of the effects of increasing clam densities in subtidal plots on abundances of predaceous invertebrates there. The number of live and dead (empty shell) individuals of each of two snail species were determined for each of three control plots and each of three 'clams added' plots. The counts were then pooled across replicates within clam treatments, thus setting the stage for sacrificial pseudoreplication. Then log-linear frequency analysis was applied to a three-way contingency table (2 treatments × 2 species × 2 conditions (live, dead)) with $N = 1197$. That analysis treated species and condition as if they were treatment factors and thereby committed pseudofactorialism as well as pseudoreplication.

Better would have been to simply use a *t*-test, for each snail species, to test whether the response variable 'per cent alive' differed between the two clam treatments.

doi:10.1111/aec.12010

Dealing with multivariate response structures

Pseudofactorialism is one of many classes of problems that originate, in part, from modern design and analysis of experiments having evolved initially in such close contact with agronomic research. Strong conceptual frameworks and clear terminology developed more slowly for the topic of response structure than they did for the topics of treatment structure and design structure simply because agronomic experiments typically had simple response structures consistent with their very focused objectives. Often the only response variable of strong interest was yield at end of growing season – a single response variable measured at a single point in time. A major exception has been the long-standing interest mentioned above in how to deal with experimental data from successive harvests over time for those crops that yield them.

Although in most other disciplines, such as ecology, psychology and medicine, complex response structures have long been common, textbooks lagged in providing coherent discussions of response structure. Books that cover that topic as well as chapters 15 and 16 in Mead *et al.* (2003) are vanishingly rare. As Urquhart (1981) intimated, the majority of books do not even distinguish response structure as a subject matter distinct from design structure (e.g. Box *et al.* 1978; Winer *et al.* 1991; Sokal & Rohlf 1995; Steel *et al.* 1997; Quinn & Keough 2002; Federer & King 2007; Casella 2008; Milliken & Johnson 2009). This is a large topic on which we offer only a few suggestions here.

The principal dimensions of response structure are (i) the number and nature of response variables measured in each experimental unit; (ii) the number and location of evaluation units on which each response variable is measured in an experimental unit; and (iii) the monitoring schedule over time for each response variable. The variety of potential response structures is great, as is their potential complexity – and as is the number of technically *valid* options available for the statistical analysis of any given one.

When measurement of a given response variable is repeated over time, there are a variety of analytical options (e.g. Mead 1988; Everitt 1995; Keselman *et al.* 2001; Mead *et al.* 2003; Casella 2008; Hinkelmann & Kempthorne 2008; Milliken & Johnson 2009). For many, perhaps most, experimental data sets perfectly reasonable interpretations are achieved by conducting a separate statistical analysis of that response variable for each monitoring date. That approach is perfectly valid and does not require any corrections for multiple comparisons (e.g. Gill 1978; Mead & Curnow 1983; Mead 1988; Finney 1990; Soto & Hurlbert 1991; Underwood 1997; Hurlbert & Lombardi 2012), many claims to the contrary notwithstanding.

Repeated measures ANOVA is sometimes recommended as an alternative approach. It certainly is preferable to treating time as a treatment factor crossed with the actual treatment factor(s) and committing pseudofactorialism. However use of repeated measures ANOVA implies there is value in testing two trivial null hypotheses: (i) that the response variable is constant over time; and (ii) that the treatment effect is constant over time. Moreover, such ANOVAs must be considered invalid in the absence of some sort of correction (e.g. Greenhouse & Geisser 1959; Huynh & Feldt 1976) for the departure of the within-experimental unit correlation structure from the sphericity assumption.

When the repeated measures correspond to a particular aspect of time, for example night *versus* day, there may be specific interest in assessing how effect sizes produced by the treatment factors change with the time factor. For example, if Y_d represents the value of the response variable during the day and Y_n represents its value on the same experimental unit at night, one could do analysis that tested whether there is a treatment effect on Y_d/Y_n or $Y_d/(Y_d + Y_n)$ or $(Y_d - Y_n)$. Which of these composite response variables was most appropriate one would need to be decided with care. The error d.f. involved would be the same as when simple response variables were tested. The works cited above give many suggestions as to the types of composite variables that can be useful.

When there are multiple response variables a parallel line of reasoning argues for the validity of carrying out a separate statistical analysis for each. Most researchers acknowledge, on first principles, that different response variables are always affected differentially by treatment factors. They may want to document *in what way* the response variables behave differently but testing the trivial null hypothesis that they all behave the same way is usually of no interest. And of course for response variables measured in different units, only a separate analysis for each variable makes sense. Consider, for example, an experiment on the effects of nitrogen fertilization on field plots where the three response variables are soil pH, soil water content (per cent, by weight), and plant biomass (g per square metre).

When there are multiple response variables of a similar sort and measured in the same units (e.g. biomass of each of several species in a plot), however, some researchers may still wish to test whether the different variables respond in the same way to the treatment factors, that is, whether there is an interaction between the treatment factor(s) and the response variable factor. Some may even 'feel' that it is imperative to do so.

As most texts offer no clear guidance for meeting that objective, researchers have been easily confused by the terminological chaos in statistics and led into the commission of pseudofactorialism. One simple intuitive approach is to create composite variables, like the ratio of two response variables (e.g. abundances of two species) and test the effect of the treatment factor on that ratio. This parallels the suggestion above for

doi:10.1111/aec.12010

testing for an interaction between a treatment factor and time, when experimental units are measured once at night and once during the day. Mead *et al*. (2003), under the heading 'Joint (Bivariate) Analysis', give a clear introductory exposition on other approaches useful for assessing correlations between two response variables that may aid interpretation of results.

More complex ANOVAs are possible that would test for a treatment effect, a response variable effect, and their interaction. However, these ANOVAs would require error d.f. adjustments similar to those needed when a repeated measures ANOVA is applied to data for a single response variable (e.g. Greenhouse & Geisser 1959; Huynh & Feldt 1976). These correct, or attempt to correct, for the anticipated correlation structure of the response variables within an experimental unit. When response variables are treated as different levels of a subunit factor, as in the earlier mentioned examples in Federer and King (2007) and Casella (2008), and there is no adjustment of error d.f., the analysis represents pseudofactorialism.

Many statistics texts do have a chapter or chapter section on multivariate ANOVA (MANOVA). This procedure can handle repeated measures in time and multiple response variables simultaneously. But it probably is the least useful of all approaches to analysis of multivariate response structures. Nevertheless some authors state MANOVA to be almost obligatory for response structures of any complexity that meet requirements of the methodology. Scheiner (1993) states, 'When more than one response variable has been measured the most appropriate method of analysis is usually multivariate analysis of variance (MANOVA) in which all dependent variables are included in a single analysis. Unfortunately ecologists often do not use MANOVA when they should'. Gotelli and Ellison (2004) state, 'If . . . we have a vector of correlated dependent variables [e.g. multiple response variables and/or repeated measures in time] we rely on a multivariate analysis of variance (MANOVA) or other multivariate methods . . .'

Our 60 sets of authors who committed pseudofactorialism ignored such advice, which could have offered them at least technical salvation. Probably they avoided MANOVA for the same reason as do most other experimenters: it is an unnecessary and cumbersome methodology for extracting information from experimental data sets with multiple response variables. Reporting of MANOVA results generates a high ratio of 'statistical baggage' to 'scientific information'. MANOVA tests the uninteresting null hypothesis that for no response variable for no date is there any difference among treatments.

Beware the referees, too

One final case of pseudofactorialism listed might be mentioned for a particular lesson it has for young scientists.

As a graduate student at my university, Jim Mills (1986) set up four different herbivory treatments using 63 plots in an area of recently burned chaparral vegetation in San Diego County. All plots contained, inter alia, shoots of two naturally occurring different shrub species, chamise and ceanothus. Response of those two species to the herbivory treatments was the focus of the study. When the data were in, Jim did an appropriate one-way ANOVA for each of the species, wrote up the manuscript and submitted it to *Ecology*.

He was pleased when it was judged acceptable if several changes were made. One reviewer, for example, had gently questioned why he 'used separate tests rather than a two-way ANOVA' (Anon., pers. comm. to J. Mills, 1986). Motivated to assure that his article would be published in such a prestigious journal as *Ecology*, Jim responded deferentially to what he perceived as a strong suggestion. A revised manuscript with a new, 'pseudofactorial' two-way ANOVA was submitted and accepted. Unfortunately, it still contained, unrevised in the Acknowledgements section, Jim's thanks to myself and another San Diego State University (SDSU) colleague, Boyd Collier, for our 'statistical advice'. But we were not consulted on the revision. Glad that is cleared up at last! Jim is now editor of the *Journal of Wildlife Diseases* and keeping an eagle eye out for *pseudofactorialistas* among both his authors *and* his referees.

The moral of the story is that bad statistical advice is often given out by paper referees. Do not accept any advice without fully understanding and believing in it. Continue to distrust authority.

Two psychological drivers

Although poor terminology and poor understanding of statistics are the proximate causes of pseudofactorialism, there are psychological factors that encourage the error. Two big and interrelated ones we may call *anovamegaphilia* and *alpha paranoia*.

By anovamegaphilia is meant the preference of some statisticians and researchers to tackle analysis of an experiment by building the most inclusive models possible and carrying out the biggest, most complex ANOVAs. The mere size of an analysis, and complexity and sophistication of the mathematics involved, can be psychologically attractive to some, especially statisticians of a more theoretical bent.

What might be considered Exhibit A for this phenomenon is a 22-page chapter in Federer and King (2007) with the title, 'World's Record for the Largest Analysis of Variance Table (259 Lines and for the Most Error Terms (62) in One Analysis of Variance'. Seriously! The table was for a complex, long-term agronomic experiment initiated in the 1950s on pineapple in Hawaii. It only shows the partitioning of the 768 d.f.

doi:10.1111/aec.12010

available, contains as usual no indication of effect sizes, and seems never to have been published in a scientific journal.

The second psychological factor, alpha paranoia, refers to the 'multiple comparison problem', the irrational fear of the supposed high risk of making one or more type I errors when many statistical tests are conducted. It is that fear that still drives some scientists to recommend MANOVA and/or to disallow or discourage separate analysis of individual response variables, separate date-by-date statistical tests, and any other approach that creates 'multiplicities'. That fear likely was a motivating factor behind some of the cases of pseudofactorialism reported here. It has driven thousands more to unnecessary and arbitrary use of procedures that involve fixing set-wise or experiment-wise type I error rates. Justification for terming this fear 'irrational' and fixed set-wise error rate procedures 'arbitrary' and documentation of the large number of statisticians and scientists who for decades have been making this same argument is provided in another review (Hurlbert & Lombardi 2012) and numerous books and papers cited therein.

CONCLUSION

The causes of pseudofactorialism would seem to be, in part, the same as those responsible for many other types of statistical errors in the scientific literature. These would include, most prominently: a long-persisting lack of a single, standardized and widely accepted terminology for experimental design; a high frequency of unreliable advice (books, articles, manuals for software packages, courses, advisors, referees, editors), including that bad advice implicit in the resultant faulty statistical analyses in the literature; and insufficient effort on the part of authors to learn the basic principles of experimental design and analysis and to overcome the prevalence of bad advice. Improvement of this state of affairs is clearly a long-term project.

Detecting and/or avoiding pseudofactorialism is, however, a manageable short-term project. Having a specific label for the error may help. The graduate students in my experimental design course had little difficulty in spotting the problem in papers they examined independently from me, after having had a 20-min lecture on the topic. And they had had only one or two courses in statistics prior to my course. To check for pseudofactorialism, one first determines whether a multi-way ANOVA (or equivalent) was used. Then one determines what the experimental unit was in the study and how many were employed. If the error d.f. for tests of treatment effects exceed the number of experimental units used, one checks to see if time or a response variable factor (e.g. species) was treated as an experimental treatment factor in the multi-way ANOVA. If it was you can put a *pseudofactorialista* notch in the handle of your pistol.

ACKNOWLEDGEMENTS

This paper is dedicated to the SDSU biology graduate students in my experimental design course who, over almost three decades, earned their black belts in pseudofactorialism-psleuthing and dug up most of the examples presented here and many more. Jim Mills (1986) and the wayward *Ecology* referee are thanked for first opening my own eyes to the problem. Brett C. Harvey independently discovered and precisely defined the phenomenon, writing me in January 1990 with an offering of Schmitt (1987) and Schlosser and Ebel (1989) as examples. Michael White discovered most of the cases of pseudofactorialism reported in our first publication on the phenomenon (Hurlbert & White 1993). David Hambright, Paul Murtaugh and N. Scott Urquhart are thanked for review of the manuscript and their suggestions and challenges. Finally, thanks to all the cited authors who responded graciously and constructively to my gentle provocations.

REFERENCES

Adler L. S., Schmitt J. & Bowers M. D. (1995) Genetic variation in defensive plant chemistry in *Plantago lanceolata* (Plantaginacea) and its effect on the specialist herbivore *Junonia coenia* (Nymphalidae). *Oecologia* **101,** 75–85.

Arnone J. A. III & Körner C. (1995) Soil biomass and carbon pools in model communities of tropical plants under elevated CO_2. *Oecologia* **104,** 61–71.

Bell J. D. & Westoby M. (1986) Importance of local changes in leaf height and density to fish and decapods associated with sea grasses. *J. Exp. Mar. Biol. Ecol.* **104,** 249–74.

Boag P. T. (1987) Effects of nestling diet on growth and adult size of zebra finches (*Poephila guttata*). *Auk* **104,** 155–66.

Bowman W. D., Theodose T. A. & Fisk M. C. (1995) Physiological and production responses of plant growth forms to increases in limiting resources in alpine tundra: implications for differential community response to environmental damage. *Oecologi* **101,** 217–27.

Box G. E., Hunter W. G. & Hunter J. S. Jr (1978) *Statistics for Experimenters*. Wiley, New York.

Box J. F. (1978) *R.A. Fisher: The Life of a Scientist*. Wiley, New York.

Brett M. T. (1992) *Chaoborus* and fish-mediated influences on *Daphnia longispina* population structure, dynamics and life-history strategies. *Oecologia* **89,** 69–77.

Bushek D. & Allen S. K. Jr (1996) Host–parasite interactions among broadly distributed populations of the eastern oyster *Crassostrea virginica* and the protozoan *Perkinsus marinus*. *Mar. Ecol. Prog. Ser.* **139,** 127–41.

Casella G. (2008) *Statistical Design*. Springer, New York.

Cipollini M. L., Drake B. G. & Whigham D. (1993) Effects of elevated CO_2 on growth and carbon/nutrient balance in the deciduous woody shrub *Lindera benzoin* (Lauraceae). *Oecologia* **96,** 339–46.

doi:10.1111/aec.12010

Cochran W. G. (1939) Long-term agricultural experiments. *Suppl. J. R. Stat. Soc.* **6,** 104–48.

Cochran W. G. (1976) Early development of techniques in comparative experimentation. In: *On the History of Statistics and Probability* (ed. D. B. Owen) pp. 3–25. Dekker, New York.

Cochran W. G. (1980) Fisher and the analysis of variance. In: *R.A. Fisher: An Appreciation (Lecture Notes in Statistics)*, Vol. 1. (eds E. Fienberg & D.V. Hinkley) pp. 17–34. Springer, New York.

Cochran W. G. & Cox G. M. (1950) *Experimental Designs*. Wiley, New York.

Cochran W. G. & Cox G. M. (1957) *Experimental Designs*, 2nd edn. Wiley, New York.

Cox D. R. (1958) *Planning of Experiments*. Wiley, New York.

Crutchfield R. S. (1938) Efficient factorial design and analysis of variance illustrated in psychological experimentation. *J. Psychol.* **5,** 339–46.

de la Cruz A. A., Hackney C. T. & Bhardwaj N. (1989) Temporal and spatial patterns of redox potential (Eh) in three tidal marsh communities. *Wetlands* **9,** 181–90.

DeStaso J. III & Rahel F. J. (1994) Influence of water temperature on interactions between juvenile Colorado River cutthroat trout and brook trout in a laboratory stream. *Trans. Amer. Fish. Soc.* **123,** 289–97.

Diffendorfer J. E., Gaines M. S. & Holt R. D. (1995) Habitat fragmentation and movements of three small mammals (*Sigmodon, Microtus*, and *Peromyscus*). *Ecology* **76,** 827–39.

Edwards M. S. (1998) Effects of long-term kelp canopy exclusion on the abundance of the annual alga *Desmarestia ligulata* (Light F). *J. Exp. Mar. Biol. Ecol.* **228,** 309–26.

Everitt B. S. (1995) The analysis of repeated measures: a practical review with examples. *Statistician* **44,** 113–35.

Fechner G. (1860) *Elemente der Psychophysik, 2 Vol*. Breitkopf & Hartig, Leipzig.

Federer W. T. (1955) *Experimental Design: Theory and Application*. Macmillan, New York.

Federer W. T. (1975) The misunderstood split plot. In: *Applied Statistics* (ed. R. P. Gupta) pp. 9–39. North Holland Publishing, Amsterdam.

Federer W. T. (1977) Sampling, blocking and modeling considerations for split plot and split block designs. *Biom. J.* **19,** 181–200.

Federer W. T. (1986) Whither statistics? *Statistical Design: Theory and Practice: Proceedings of a Conference in Honor of Walter T. Federer* (eds C. E. McCulloch, S. J. Schwager, G. Casella & S. R. Searle) pp. 211–31. Cornell University, Ithaca.

Federer W. T. & King F. (2007) *Variations on Split Plot and Split Block Experiment Designs*. Wiley, Hoboken, New Jersey.

Finney D. J. (1955) *Experimental Design and Its Statistical Basis*. University Chicago Press, Chicago.

Finney D. J. (1990) Repeated measurements; what is measured and what repeats? *Stat. Med.* **9,** 639–44.

Finzi A. C., Allen A. S., DeLucia E. H., Ellsworth D. S. & Schlesinger W. H. (2001) Forest litter production, chemistry, and decomposition following two years of free-air CO_2 enrichment. *Ecology* **82,** 470–84.

Fisher R. A. (1925) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher R. A. (1926) The arrangement of field experiments. *J. Minist. Agric. (G. B.)* **33,** 503–13.

Fisher R. A. (1935) *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Fisher R. A. & Mackenzie W. A. (1923) Studies in crop variation. II. The manurial response of different potato varieties. *J. Agric. Sci.* **13,** 311–20.

Fisher R. A. & Wishart J. (1930) The arrangement of field experiments and the statistical reduction of the results. Technical Communication No. 10, Imperial Bureau of Soil Science. His Majesties Stationery Office, London, 24 pp.

Fraser L. H. & Grime J. P. (1998) Top-down control and its effect on the biomass and composition of three grasses at high and low soil fertility in outdoor microcosms. *Oecologia* **113,** 239–46.

Gardner S. D. L., Taylor G. & Bosac C. (1995) Leaf growth of hybrid poplar following exposure to elevated CO_2. *New Phytol.* **131,** 81–90.

Gill J. L. (1978) *Design and Analysis of Experiments in the Animal and Medical Sciences*, Vol. 2. Iowa State University Press, Ames.

Gotelli N. J. (1988) Determinants of recruitment, juvenile growth, and spatial distribution of a shallow-water gorgonian. *Ecology* **69,** 157–66.

Gotelli N. J. & Ellison A. M. (2004) *A Primer of Ecological Statistics*. Sinauer Associates, Sunderland.

Graham J. H., Fletcher D., Tigue J. & McDonald M. (2000) Growth and developmental stability of *Drosophila melanogaster* in low frequency magnetic fields. *Bioelectromagnetics* **21,** 465–72.

Grayson K. L. & Wilbur H. M. (2009) Sex- and context-dependent migration in a pond-breeding amphibian. *Ecology* **90,** 306–12.

Greenhouse S. W. & Geisser S. (1959) On methods in the analysis of profile data. *Psychometrika* **24,** 95–112.

Gulmon S. L. (1992) Patterns of seed germination in California serprentine grassland species. *Oecologia* **89,** 27–31.

Hairston N. G. Sr (1986) Species packing in *Desmognathus* salamanders: experimental demonstration of predation and competition. *Am. Nat.* **127,** 266–91.

Hambright K. D. (1994) Can zooplanktivorous fish really affect lake thermal dynamics? *Arch. Hydrobiol.* **130,** 429–38.

Hart S. P. & Marshall D. J. (2009) Spatial arrangement affects population dynamics and competition independent of community composition. *Ecology* **90,** 1485–91.

Herman D., Kaushik N. K. & Solomon K. R. (1986) Impact of atrazine on periphyton in freshwater enclosures and some ecological consequences. *Can. J. Fish. Aquat. Sci.* **43,** 1917–25.

Hinkelmann K. & Kempthorne O. (1994) *Design and Analysis of Experiments, Vol. 1. Introduction to Experimental Design*, 2nd edn. Wiley-Interscience, New York.

Hinkelmann K. & Kempthorne O. (2005) *Design and Analysis of Experiments, Vol. 2: Introduction to Experimental Design*, 2nd edn. Wiley-Interscience, New York.

Hinkelmann K. & Kempthorne O. (2008) *Design and Analysis of Experiments, Vol. I: Introduction to Experimental Design*, 3rd edn. Wiley-Interscience, New York.

Hollertz K. (2002) Feeding biology and carbon budget of the sediment-burrowing heart urchin *Brissopsis lyrifera* (Echinoidea: Spatangoida). *Mar. Biol.* **140,** 959–69.

Hovel K. A. & Morgan S. G. (1997) Planktivory as a selective force for reproductive synchrony and larval migration. *Mar. Ecol. Prog. Ser.* **157,** 79–95.

Howe S. A. & Marshall A. T. (2002) Temperature effects on calcification rate and skeletal deposition in the temperate coral, *Plesiastrea versipora* (Lamarck). *J. Exp. Mar. Biol. Ecol.* **275,** 63–81.

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

doi:10.1111/aec.12010

Hurlbert S. H. (2009) The ancient black art and transdisciplinary extent of pseudoreplication. *J. Comp. Psychol.* **123,** 434–43.

Hurlbert S. H. (in press) Affirmation of the classical terminology for experimental design via a critique of Casella (2008). *Agron. J.*

Hurlbert S. H. & Lombardi C. M. (2004) Research methodology: experimental design, sampling design, statistical analysis. In: *Encyclopedia of Animal Behavior* (ed. M. M. Bekoff) pp. 755–62. Greenwood Press, London.

Hurlbert S. H. & Lombardi C. M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann. Zool. Fenn.* **46,** 311–49.

Hurlbert S. H. & Lombardi C. M. (2012) Lopsided reasoning on lopsided tests and multiple comparisons. *Aust. N. Z. J. Stat.* **54,** 23–42.

Hurlbert S. H. & Meikle W. G. (2003) Pseudoreplication, fungi and locusts. *J. Econ. Entomol.* **96,** 533–5.

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

Huynh H. & Feldt L. S. (1976) Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J. Educ. Stat.* **1,** 69–82.

Jaenike J. & Anderson T. J. C. (1992) Dynamics of host-parasite interactions: the *Drosophila-Howardula* system. *Oikos* **64,** 533–40.

Johnston J. F.W. (1849) *Experimental Agriculture, Being the Results of Past, and Suggestions for Future Experiments in Scientific and Practical Agriculture*. Blackwood, Edinburgh.

Kenyon R. A., Loneragan N. R. & Hughes J. M. (1995) Habitat type and light affecting sheltering behaviour of juveniles tiger prawns (*Penaeus esculentus* Haswell) and success rates of their fish predators. *J. Exp. Mar. Biol. Ecol.* **192,** 87–105.

Keselman H. J., Algina J. & Kowalchuk R. K. (2001) The analysis of repeated measures designs: a review. *Br. J. Math. Stat. Psychol.* **54,** 1–20.

Kirk R. E. (1982) *Experimental Design: Procedures for the Behavioral Sciences*, 2nd edn. Brooks/Cole, Pacific Grove.

Kozlov M. & Hurlbert S. H. (2006) Pseudoreplication, chatter, and the international nature of science: a response to D.V. Tatarnikov. *J. Fundam. Biol. (Moscow)* **67,** 145–52. [In Russian; English translation available from S. Hurlbert or via Google Scholar].

Lauenroth W. K., Dodd J. L. & Sims P. L. (1978) The effects of water- and nitrogen-induced stresses on plant community structure in a semiarid grassland. *Oecologia* **36,** 211–22.

Lederhouse R. C., Ayres M. P. & Scriber J. M. (1990) Adult nutrition affects male virility in *Papilio glaucus* L. *Funct. Ecol.* **4,** 743–51.

Lee W. (1975) *Experimental Design and Analysis*. W.H. Freeman, San Francisco.

Lewis D. & Burke C. J. (1949) The use and misuse of the chi-square test. *Psychol. Bull.* **46,** 433–89.

Lindquist E. F. (1940) *Statistical Analysis in Educational Research*. Houghton-Mifflin, Boston.

Lombardi C. M. & Hurlbert S. H. (1996) Sunfish cognition and pseudoreplication. *Anim. Behav.* **53,** 419–22.

McCluney K. E. & Sabo J. L. (2009) Water availability directly determines per capita consumption at two trophic levels. *Ecology* **90,** 1463–9.

McGrath L. J., van Riper I. I. I. C. & Fontaine J. J. (2009) Flower power: tree flowering phenology as a settlement cue for migrating birds. *J. Anim. Ecol.* **78,** 22–30.

McIntosh W. J. (1986) The effect of imagery generation on science rule learning. *J. Res. Sci. Teach.* **23,** 1–9.

Main K. L. (1987) Predator avoidance in seagrass meadowns: prey behavior, microhabitti selection, and cryptic coloration. *Ecology* **68,** 170–80.

Mamalos A. P., Elisseou G. K. & Veresoglou D. S. (1995) Depth of root activity of coexisting grassland species in relation to N and P additions, measured using nonradioactive tracers. *J. Ecol.* **83,** 643–52.

Mattila J. & Bonsdorff E. (1998) Predation by juvenile flounder (*Platichthys flesus* L.): a test of prey vulnerability, predator preference, switching behaviour and functional response. *J. Exp. Mar. Biol. Ecol.* **227,** 221–36.

Mead R. (1988) *The Design of Experiments*. Cambridge University Press, Cambridge.

Mead R. & Curnow R. N. (1983) *Statistical Methods in Agriculture and Experimental Biology*. Chapman and Hall, London.

Mead R., Curnow R. N. & Hasted A. M. (2003) *Statistical Methods in Agriculture and Experimental Biology*, 3rd edn. Chapman & Hall, New York.

Metaxas A. & Young C. M. (1998) Responses of echinoid larvae to food patches of different algal densities. *Mar. Biol.* **130,** 433–45.

Milliken G. A. & Johnson D. E. (2009) *Analysis of Messy Data, Volume I: Designed Experiments*, 2nd edn. CRC Press, New York.

Mills J. N. (1986) Herbivores and early postfire succession in Southern California. *Ecology* **67,** 1637–49.

Moltschaniwskyj N. A. & Martínez P. (1998) Effect of temperature and food levels on the growth and condition of juvenile *Sepia elliptica* (Hoyle 1885): an experimental approach. *J. Exp. Mar. Biol. Ecol.* **229,** 289–302.

Monod H. & Bailey R. H. (1992) Pseudofactors: normal use to improve design and facilitate analysis. *Appl. Stat.* **41,** 317–36.

Montague J. R., Aguinaga J. A., Ambrisco K. L., Vassil D. L. & Collazo W. (1991) Laboratory measurement of ingestion rate for the sea urchin *Lytechinus variegatus* (Lamarck) (Echinodemrata: Echinoidea). *Fla Sci.* **54,** 129–34.

Morin P. J., Lawler S. P. & Jackson E. A. (1990) Ecology and breeding phenology of larval *Hyla andersonii*: the disadvantage of breeding late. *Ecology* **71,** 1590–8.

Parsons W. F. J., Ehrenfeld J. G. & Handel S. N. (1998) Vertical growth and mycorrhizal infection of woody plant roots a potential limits to the restoration of woodlands on landfills. *Restor. Ecol.* **6,** 280–9.

Patzkowsky M. E. (1988) Differential response of settling larvae to resident colony density in two species of *Bugula* (Bryozoa: Cheilostomata). *J. Exp. Mar. Biol. Ecol.* **124,** 57–63.

Peters C. C. & VanVoorhis W. R. (1940) *Statistical Procedures and Their Mathematical Bases*. McGraw-Hill, New York.

Polak M., Opoka R. & Cartwright I. L. (2002) Response of fluctuating asymmetry to arsenic toxicity: support for the developmental selection hypothesis. *Environ. Pollut.* **118,** 19–28.

Polis G. A. & McCormick S. J. (1987) Intraguild predation and competition among desert scorpions. *Ecology* **68,** 332–43.

Potvin C. (1993) ANOVA: experiments in controlled environments. In: *Design and Analysis of Ecological Experiments* (eds S. M. Scheiner & J. Gurevitch) pp. 46–68. Chapman & Hall, New York.

Power E. A. (1988) Enclosure experiments to test effects of log storage on a large salmonid producing lake; water quality and zooplankton density. *Verh. Int. Verein. Theoret. Angew. Limnol.* **23,** 1578–85.

doi:10.1111/aec.12010

Preen A. (1995) Impacts of dugong foraging on seagrass habitats: observational and experimental evidence for cultivation grazing. *Mar. Ecol. Prog. Ser.* **124,** 201–13.

Quinn G. P. & Keough M. J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Raktoe B. L., Hedayat A. & Federer W. T. (1981) *Factorial Designs*. Wiley, New York.

Ritchie M. E., Tilman D. & Knops J. M. H. (1998) Herbivore effects on plant and nitrogen dynamics in oak savanna. *Ecology* **79,** 165–77.

Rucci A. J. & Tweney R. D. (1980) Analysis of variance and the 'second discipline' of scientific psychology: a historical account. *Psychol. Bull.* **87,** 166–84.

Sanders J. J. & Gordon D. M. (2003) Resource-dependent interactions and the organization of desert ant communities. *Ecology* **84,** 1024–31.

SAS (2007) *Concepts of Experimental Design: A SAS White Paper*. Design Institute for Six Sigma & SAS Institute, Cary, 34 pp.

SAS (2012) *SAS/SYSTAT(R) 9.3 User's Guide*. SAS Institute, Cary.

Scheiner S. M. (1993) MANOVA: multiple response variables and multispecies interactions. In: *Design and Analysis of Ecological Experiments* (eds S. M. Scheiner & J. Gurevitch) pp. 94–112. Chapman & Hall, New York.

Schelske C. L., Rothman E. D., Stoermer E. F. & Santiago M. A. (1974) Responses of phosphorus limited Lake Michigan phytoplankton to factorial enrichments with nitrogen and phosphorus. *Limnol. Oceanogr.* **19,** 409–19.

Schlosser I. J. & Ebel K. K. (1989) Effects of flow regime and cyprinid predation on a headwater stream. *Ecol. Monogr.* **59,** 41–57.

Schmitt R. J. (1987) Indirect interactions between prey: apparent competition, predator aggregation, and habitat segregation. *Ecology* **68,** 1887–97.

Schröder A., Nilsson K. A., Persson L., van Kooten T. & Reichstein B. (2009) Invasion success depends on invader body size in a size-structured mixed predation-competition community. *J. Anim. Ecol.* **78,** 1152–62.

Shakarad M., Prasad N. G., Rajamani M. & Joshi A. (2001) Evolution of faster development does not lead to greater fluctuating asymmetry of sternopleual bristle number in *Drosophila*. *J. Genet.* **80,** 1–7.

Snedecor G. W. & Cochran W. G. (1967) *Statistical Methods*, 6th edn. Iowa State University Press, Ames.

Sokal R. R. & Rohlf F. J. (1981) *Biometry*, 2nd edn. W.H. Freeman, New York.

Sokal R. R. & Rohlf F. J. (1995) *Biometry*, 3rd edn. W.H. Freeman, New York.

Soto D. & Hurlbert S. H. (1991) Long-term experiments on calanoid–cyclopoid interactions. *Ecol. Monogr.* **61,** 246–65.

Sousa W. P., Schroeter S. C. & Gaines S. P. (1981) Latitudinal variation in intertidal algal community structure: the influence of grazing and vegetative propagation. *Oecologia* **48,** 297–307.

Steel R. G. D. & Torrie J. H. (1960) *Principles and Procedures of Statistics*. McGraw-Hill, New York.

Steel R. G. D. & Torrie J. H. (1980) *Principles and Procedures of Statistics: A Biometrical Approach*, 2nd edn. McGraw-Hill, New York.

Steel R. G. D., Torrie J. H. & Dickey D. A. (1997) *Principles and Procedures of Statistics*, 3rd edn. McGraw-Hill, New York.

Stigler S. M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge.

Stoll P., Oggier P. & Baur B. (2009) Population dynamics of six land snail species in experimentally fragmented grassland. *J. Anim. Ecol.* **78,** 236–46.

Strauss S. Y., Stanton M. L., Emery N. C. *et al*. (2009) Cryptic seedling herbivory by nocturnal introduced generalists: impacts, survival, performance of native and exotic plants. *Ecology* **90,** 419–29.

Streams F. A. (1987) Within-habitat spatial separation of two *Notonecta* species: interactive vs. noninteractive resource partitioning. *Ecology* **68,** 935–45.

Tindall S. D., Ralph C. J. & Clout M. N. (2007) Changes in bird abundance following Common Myna control on a New Zealand island. *Pac. Conserv. Biol.* **13,** 202–12.

Underwood A. J. (1997) *Experiments in Ecology*. Blackwell, London.

Urquhart N. S. (1981) The anatomy of a study. *HortScience* **16,** 100–16.

Valiela I. (2001) *Doing Science: Design, Analysis, and Communication of Scientific Research*. Oxford University Press, Oxford.

Welker J. M., Wookey P. A., Parsons A. N., Press M. C., Callaghan T. V. & Lee J. A. (1993) Leaf carbon isotope discrimination and vegetative responses of Dryas octopetala to temperature and water manipulations in a high Arctic polar semi-desert, Svalbard. *Oecologia* **95,** 463–9.

Wickens T. D. (1993) Analysis of contingency tables with between-subjects variability. *Psychol. Bull.* **113,** 191–204.

Wiley R. H. (2003) Is there an ideal behavioural experiment? *Anim. Behav.* **66,** 585–8.

Winer B. J., Brown D. R. & Michels K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edn. McGraw-Hill, New York.

Wolins L. (1982) *Research Mistakes in the Social and Behavioral Sciences*. Iowa State University Press, Ames.

Yates F. (1935) Complex experiments. *Suppl. J. R. Stat. Soc.* **2,** 181–247.

Yates F. (1936) A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.* **26,** 424–55.

Yates F. (1937) *The Design and Analysis of Factorial Experiments*. Technical Communication No. 35, Imperial Bureau of Soil Science, Harpenden, England.

Yates F. & Mather K. (1963) Ronald Aylmer Fisher, 1890–1962. *Biogr. Mem. Fellows R. Soc.* **9,** 90–129.

Young T. P. & Okello B. D. (1998) Relaxation of an induced defense after exclusion of herbivores: spines on *Acacia drepanolobium*. *Oecologia* **115,** 508–13.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Characteristics of 60 examples of pseudofactorialism from the literature, 1974–2009.

doi:10.1111/aec.12010