Opinion is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise as possible and references should be kept to a minimum. A summary is not required.

The devil lies in details: reply to Stuart Hurlbert

Lauri Oksanen, Dept of Ecology and Environmental Science, Umeå Univ., SE-901 87 Umeå, Sweden (lauri.oksanen@eg.umu.se)

As pointed out in Stuart Hurlbert's recent article, ecologists still at times design their experiments sloppily, creating a situation where various forms of 'non-demonic intrusion' could account for the documented contrasts between treatments and controls. If such contrasts are nevertheless presented to the reader as if they were statistically demonstrated treatment effects, then pseudoreplication is not a pseudoissue and the use of a stigmatizing label of is entirely warranted, as pointed out by Hurlbert. The problems with Hurlbert's concepts start in the context of studies, where the scope of the experiment is to provoke a chain of dramatic and a priori extremely unlikely events, which a given conjecture predicts to happen as a consequence of a given manipulation. As the essence of these experiments is to trigger large dynamical responses in a biological system, they often require much space and/or special conditions, allowing for efficient isolation of the experimental system from potential sources of contamination. These constraints can be incompatible with standard designs (randomization, replication and treatment-control interspersion). In the context of experiments, where it has been necessary to sacrifice randomization, replication or treatment-control interspersion, the logic of inferring treatment effects is the same as used when interpreting causes of spontaneous events or events triggered by manipulations with practical purposes. The observed contrasts can be reasonably interpreted as effects of the treatment if and only if their magnitudes and the timing of their emergence makes alternative explanations utterly implausible (which is up to the reader to judge). If the logic of inference is clearly explained and no claim of statistically demonstrated treatment effect is made, the use of stigmatizing labels like 'pseudoreplication' is unwarranted. However, it might clarify the literature if such imperfectly designed experiments are referred to as experimental events, to be distinguished from perfectly designed experiments, where mechanical interpretation of contrasts between treatments and controls as treatment effects can be regarded as socially acceptable.

In structured discussions, good habits require that the one, who opened the discussion, keeps a low profile in the subsequent debate and enters it only to correct outright misconceptions. Thus, I saw no reason to react to the response of Cottenie and De Meester (2003) to my paper on the trade-offs of experimentalists working on large-scale systems (Oksanen 2001), although our views are by no means identical. Their contribution is sound and does not distort my viewpoints. Unfortunately, however, Hurlbert's (2004) response is of different kind and must be met with a detailed reply. To begin with, I must note that he obviously misunderstood the scope of my paper. It was not supposed to be a review of the literature dealing with pseudoreplication but a critique of those aspects of Hurlbert's (1984) classical paper, which I regarded as counterproductive. I thus saw no reason to cite papers, where the same logic was just applied to different cases. Moreover, I did not attempt to discredit Hurlbert's contribution. In its historical context, Hurlbert's (1984) classical paper was outstanding and definitely deserved its awards. I was critical to other ecologists – not the least to myself – who for years have failed to challenge the weakest and most controversial arguments of Hurlbert. Instead of exposing the paper to sound scientific debate, which washes the gold from the gravel, we have meekly accepted everything that stands there. That kind of worship is never good in science.

In my reply, I will focus on four issues, where the most obvious misconceptions have emerged: (1) the logical equivalence between compound treatments and pseudoreplication, as defined by Hurlbert, (2) the double standards of logical inference that have been established as a consequence of our failure to challenge Hurlbert's (1984) arguments, (3) the meaning of the concept 'experiment' and (4) the connection between the tactical issues of experimental design and the strategic questions concerning the philosophical foundations of sound inference. In the end, I will propose guidelines, which would enable us to get rid of the counterproductive aspects of Hurlbert's propositions without falling back to the real problems, which were rampant in ecological literature before 1984 and which still have not totally disappeared.

Compound treatments and the inevitably 'pseudoreplicated' nature of all experiments

The section on compound treatments represents an especially surprising aspect of Hurlbert's (2004) reply.

To my understanding, the arguments presented in this section are in outright conflict with the bottom line of Hurlbert's (1984) classical paper: that interspersion of treatments and controls is vital, because such design supposedly safeguards experiments against 'non-demonic intrusions' (external impacts logically unrelated to the treatment). In this context, physical space is just a surrogate variable. To be statistically independent *in the context of testing the significance of the nominal treatment effect*, treatments and controls must be interspersed along all potentially relevant environmental axes, so that 'non-demonic intrusions', cannot contribute to the apparent treatment effect. Logically, it makes no difference whatsoever whether lack of interspersion along some axis of environmental variation depends on spatial aggregation, on the connection of experimental units to the same circulation system or on any shared physical attribute other than the nominal treatment.

Unfortunately, however, treatments inevitably contain elements additional to the nominal treatment, and we never can be sure that the placebo treatments, applied to controls, adequately reproduce these side effects. This issue is not academic hair-splitting. Placebo fences, which have no effect on the movements on either small or large grazers, cannot have the same impact on near-ground wind velocity or snow accumulation patterns as real grazer exclosures. Experimental devices, which eliminate all effects of all predators, inevitably influence the dispersal of the prey and the dynamics of parasites with complex life cycles. Moreover, it is impossible to anticipate all side effects of a given treatment. (My 'hoarfrost accident' provides a good example, Oksanen 2001.) Hence, we should not be naïve. Using Hurlbert's logics, we find that all experiments are inevitably pseudoreplicated, at least to a degree. I totally agree with Hurlbert (1984) that this problem should be minimized and that interspersion in space is normally a good means in this context. We could even agree that interspersion in space is a sufficient precaution against 'non-demonic intrusions' to allow straightforward interpretation of contrasts as treatment effects, but we should realize that this is just a social agreement. Science is an inherently uncertain endeavor, where we can do all things 'right' (in accordance to established social agreements of the scientific community) but nevertheless end up with entirely incorrect interpretations of our experimental results.

On the double standards of inference introduced by Hurlbert (1984)

In his response to my paper, Hurlbert (2004) reiterates his well-known standpoint according to which *'pseudoreplication in any of its various guises is simply an error of statistical analysis and interpretation'*. Simple statistical errors are still committed and need to be corrected (Hurlbert and Meikle 2003). However, the concept 'error of interpretation' does not, to my understanding, have any well-defined meaning whatsoever. For me, the proper interpretation of a demonstrated contrast between two statistical populations hinges on the opinion of scientists concerning the plausibility of different putative causes. Hurlbert (1984, 2004) disagrees and regards it as inherently erroneous to interpret a demonstrated difference between a treatment and a control as a treatment effect, unless the design of the experiment excludes all *imaginable* forms of 'non-demonic intrusion', regardless of the plausibility of the conjecture that such 'intrusions' would account for the demonstrated contrast.

Recall that much of current ecological knowledge derives from case studies, where the observed scenarios are indeed vulnerable to various forms of 'non-demonic intrusion'. A celebrated case of biological control is the decline of *Opuntia* in Australia, which was supposedly an effect of the introduction of *Cactoblastis* (Caughley and Lawton 1981). Yet, this introduction lacks both replication and control. Strictly speaking, we only know that the introduction of *Cactoblastis* was followed by a dramatic decline of *Opuntia*. In other cases, different kinds of barriers provide 'controls'. Examples of this include the impacts sea otter recovery on sea urchins and kelp beds in the Aleutians (Estes and Palmisano 1974, Estes and Duggins 1995), the impacts of the recovery of giant tortoises on the vegetation of Aldabra (Merton et al. 1976), the impact of introduced rabbits on the vegetation of Kerguelen (Werth 1928), and the impact of introduced reindeer on the vegetation of South Georgia (Leader-Williams 1988). In all these cases, the locations of the 'controls' have been determined by the dynamics of the invasion or recovery and by the physical barriers stopping or slowing down the (re)colonization process. Thus, interspersion between 'treatments' and 'controls' has been imperfect at its best, and none of these cases is entirely controlled against 'non-demonic intrusions'. Yet, to my knowledge, nobody has regarded it as erroneous to interpret the contrasts as treatment effects or protested against the use of inferential statistics when documenting the contrasts between 'treatments' and 'controls'. The obvious reason lies in the magnitudes of the contrasts. Common sense and knowledge of the normal variability of these systems tells us that while sources other than the treatment may contribute to these contrasts (or reduce their magnitudes) such 'non-demonic intrusions' just do not provide a plausible alternative explanation to the bulk of the observed contrasts.

Hurlbert (1984:191) objects to the use of corresponding logic in the context of experimental work, because for him *'the essence of experimental work is that the validity of conclusions is not contingent on such assumptions with reality'*. This is one possible way of seeing the

role of experiments in science, but it is hardly the only way. Corresponding introductions could be conducted eve in the explicit purpose of testing ecological conjectures, and even then, external constraints might prevent randomization, replication and/or interspersion of treatments and controls. Would the same logic, which is generally regarded as appropriate in the context of spontaneous events and introductions made for practical purposes, now become inherently erroneous? And would it now suddenly become inappropriate to demonstrate the reliability of one's estimates of population means by means of inferential statistics? My answer to both questions is a definite 'no'. For me, sound inference should follow the logic formalized in my critique (Oksanen 2001, Eq. 1), quite regardless whether we are dealing with spontaneous events, with consequences of various management policies or with real experiments. This is the core of my disagreement with Hurlbert (1984, 2004), who wishes to give referees a blanket warrant to stigmatize such 'misapplication' of inferential statistics', whether explicit or implicit, in the contexts of experiments, and to use it as a reason for recommending rejection.

I freely admit that the issue is problematic. It is possible to maintain that the term 'experiment' should be reserved to such empirical studies, where every possible precaution has been taken to exclude all forms of 'non-demonic intrusions', but we then should find some other, value-neutral name for those manipulative studies, which are then left in a no-man's land between experiments and descriptions. Before deciding on this issue, it is useful to define the concept 'experiment' and to consider the role of experiments in the scientific process.

What is an experiment?

According to Hurlbert's (2004) definition, *a manipulative experiment is an exercise designed to determine the effects of one or more experimenter-manipulated variables (= experimental variables or treatment factors) on one or more characteristics (= response variables) of some particular type of system (= the experimental unit). Its primary defining features are: that (1) the experimenter can assign treatments or levels of each experimental variable at random to the available experimental units; and (2) that there are two or more levels established for each experimental variable used*. While being a good description of a certain category of experiments, this definition is lengthy and narrow. For the sake of interdisciplinary communication, it also seems useful to avoid homemade definitions, if it the use of generally agreed ones is possible. I thus looked how the word 'experiment' is defined in my Finnish (Facta 2001) and Swedish (Nationalencyclopedin 1990) encyclopedias. Although the two encyclopedias emphasize a bit different aspects of an experiment, the definitions are consistent and, to my understanding, suitable for ecological experiments, too. A short version can be stated as follows:

An experiment is a deliberate and active manipulation of an empirical system, conducted in order to test the validity of a conjecture or the utility of a procedure.

In these definitions, active manipulation is emphasized, making the concept 'manipulative experiment' to a tautology. Conversely, there is no mention of either randomization or treatment levels. The essence of an experiment, as defined above, is that different imaginable outcomes speak clearly for or against the validity of a given conjecture (or the utility of a given procedure). As a model for a maximally powerful experiment, both encyclopedias presents Galilei's classical study, where two metal balls of different size were simultaneously dropped from the tower of Pisa and were found to land practically simultaneously, contrary to the predictions of Aristoteles. Indeed, the above definition does not embrace astronomical events, spontaneous invasions of plants or animals, or practically motivated introductions. My reason for discussing such situations was to point out that even events with very imperfect design, if seen as experiments, could constitute strong tests of conjectures.

On the relationship between epistemology and practical methodology

The reason why even very imperfect design can be compatible with very strong message cannot be made fully understandable without discussing the epistemological dimension of science. Unfortunately, we ecologists tend to be poorly trained in this area. As a rule, we profess to use the hypothetico-deductive approach, but in reality, we tend to apply reasoning, which is very close to the logic referred to as inductionism or classical empiricism. So did I until getting involved in a political debate, where Popper's (1963) ideas were emphasized. Let us thus clarify the basic concepts. Contrary to Hurlbert's (2004) view, inductionism is not 'excessive reliance on inductive reasoning' (whatever that might mean). Inductionism is a philosophical position, outlined by Sir Francis Bacon (1620) and further developed by several other philosophers, especially by John Stuart Mill (1843). According to this philosophical school, it is possible to firmly establish causal laws by means of induction, if certain procedure rules are strictly followed. The characteristic standpoint of an inductionist is that he/she emphasizes the need to empirically prove putative causal relationships and to proceed from the specific to the general first when such a connection has been empirically established. Hurlbert's (1984) basic idea-

that the p-values reported by experimentalists could be firmly connected to the hypotheses tested in the experiments, provided that the rules of sound experimental design (randomization, replication, interspersion) are followed, represents this kind of logic. The same logic is even succinctly illustrated in the statement of Cottenie and De Meester (2003) that *one sample is enough for confirmation on the level of individual system but replication is necessary for population inference.* Ecological systems may indeed have properties, which make the inductionistic epistemology appropriate. For a meaningful discussion, however, it is helpful to recognize one's position and to realize the existence of other positions.

The basic point of departure of the hypothetico-deductive approach is that, though indispensable, induction is an inherently uncertain procedure (Popper 1963). Emphasis is thus shifted from solid inference from specific cases to general laws towards construction of testable conjectures. To be testable, conjectures must be all-statements, referring to entire categories. The basis for population inference is thus built in the conjecture itself. It is supposed to apply to all elements of a category specified by the theoretician. If a single element behaves in a way, which is in conflict with the predictions of the conjecture, the conjecture must be regarded as falsified (at least in its initial form and with respect to its initial scope, Lakatos 1972). In this context, the essence of a good experiment is not its design *per se* but the existence of clear predictions, which refer to the experimental system and which are very unlikely to be corroborated for reasons independent of the conjecture to be tested. If this is the case, then even spontaneous events and very imperfectly designed experiments can make strong tests. If this is not the case, replication and randomization are to no avail. The results will be uninstructive anyway.

The connection between epistemology and experimental design can be illustrated by the different ways in which I saw at the classical intertidal experiments (Paine 1966, 1974, 1980, Dayton 1971, Menge 1972, Lubchenko 1980) before and after having read Popper's (1963) main work. It was obvious for me all along that the treatment effects reported in these papers did not correspond to the definitions of treatment effect that I had learned in the obligatory statistics course, though I regarded this as a minor concern only. For me, the magnitudes of the changes and the timings of their emergence implied that they were primarily effects of the treatment. As an implicit inductionist, I was nevertheless concerned with the fact that the designs did not enable me to obtain accurate estimates for the magnitudes of true treatment effects, which I felt as an obstacle for the induction of generalities from these specific cases. The longer I got with Popper's text the less I cared about these limitations. After all, I had already been exposed to two contrasting, general conjectures on the distribution and abundance of organisms along environmental gradients, and both conjectures – one formulated by Gleason (1926), and Whittaker (1975), the other by Cajander (1906, 1909), Gause (1934), and MacArthur (1972) – created clear predictions concerning even intertidal systems. To my surprise, I found the results of the intertidal experiments as being at variance with the predictions of the supposedly more modern Gleason-Whittaker conjecture, according to which species are distributed individually along environmental gradients, while the Cajander-Gause-MacArthur conjecture, emphasizing biotic feedbacks, was corroborated. The change of epistemological view thus placed the intertidal studies to a new context and influenced my willingness to see studies of this kind published in major journals.

My own experiments have been designed and conducted in the same spirit. Their main scope has been to put ecological systems to states, where my conjectures predict dramatic changes in or contrasts between different systems. The primary role of statistics in these experiments has been to document that the patterns seen in the samples faithfully represent trends in sampled statistical populations. Indeed, I have nevertheless tried to follow the conventional design (replication, randomization, interspersion) *whenever feasible*. As pointed out by Hurlbert (1984), such design has immense strengths, which should not be sacrificed without a very good reason. Randomization, replication and interspersion strongly reduce the likelihood for spurious corroboration and allow for convincing documentation of even small treatment effects. Moreover, as pointed out in the final parts of Hurlbert's reply (2004), experiments tend to be expensive to set up and to maintain (not least in the conditions of arctic and alpine ecosystems), while it is often relatively cheap to monitor even such parameters, for which no a priori predictions could be derived. Interesting information has indeed been obtained from these aspects of my studies (Oksanen and Moen 1994). However, my priorities have been in the deductive moments of the experiments. Sometimes, other concerns (appropriate scales, safeguards against contamination) have thus lead to less conventional experimental designs.

When replication and/or interspersion has not been feasible or when the critical predictions have forced me to combine local experiments with essentially macro-ecological comparisons (Moen and Oksanen 1998), I have focused on such treatments, where the predicted changes and/or contrasts are very large, as compared to the normal variability of the system(s). Contrary to the statement of Hurlbert (2004), however, this approach does not apply just '*to a very narrow class of situations, where we know beforehand that the treatment effect will be so much greater than ''background variation'' that treatment replication can be dispensed with'*. I cannot even understand the meaning of this statement. If we knew the magnitude of the treatment effect beforehand, there

would be no point with conducting the experiment! My preferred approach in situations, where replication is infeasible, has been to first monitor the behavior of measurables of interest in my study systems and, out of such data, to induce the plausible magnitudes of changes due to 'non-demonic intrusions'. Then I have compared the magnitudes of these 'spontaneous' changes to the magnitudes of changes predicted by my conjecture. If the smallest responses, which could still be regarded as consistent with my conjecture, are considerably larger than the *largest* contrasts that spontaneous variability of the measurables could conceivably account for, I have regarded it as reasonable to conduct an experiment even if perfect control against 'non-demonic intrusions' was not possible. Contrary to the suggestion of Hurlbert (2004), this has nothing to do with the use of 'sledgehammer treatments'. Indeed, I prefer such treatments, which have minuscule immediate effects but are predicted to trigger very large responses, preferentially large enough to change the system beyond recognition. Given all this, Hurlbert's (2004) claim that my '*box checking protocol would seem to only require whether the difference between the two sample means is positive or negative'* is positively amazing.

Let us illustrate the problems and trade-offs with an example from the real world. An issue hotly debated among ecologists is whether the potential of food-limited herbivores to impose dramatic impacts on the benthic vegetation of the Aleutians and on the terrestrial vegetation on various oceanic islands (see above) is a specific property of impoverished island communities (Strong 1992, Polis and Strong 1996) or a generic property of terrestrial and benthic ecosystems. An obvious way to test these two conjectures against each other is to create correspondingly predation-free experimental systems, consisting of continental vegetation and of some herbivores normally using it. In this context, the predictions of contesting conjectures are clear indeed. Conjectures emphasizing the ability of plants to defend themselves and de-emphasizing the importance of predation (Murdoch 1966, Haukioja and Hakala 1975, White 1978, Rhoades 1985, Seldal et al. 1994) predict that what will be observed in such systems is just business as usual. Conjectures emphasizing the existence of community-level trophic cascades, in turn, predict that the herbivores erupt and the vegetation will change beyond recognition, at least in relatively productive ecosystems, initially dominated by woody plants or tall forbs (Hairston et al. 1960, Fretwell 1977, Oksanen et al. 1981, Oksanen and Oksanen 2000).

Unfortunately, to create such experimental systems is easier said than done. For relatively small grazers (e.g. voles), they can be constructed indoors (Moen et al. 1993), but the limited space then inevitably creates a sledgehammer design, which is undesirable for many reasons. It is possible to construct outdoor exclosures, but there the real choices found so far have been semi-agricultural systems, grass dominated already to begin with (Norrdahl et al. 2002), and field systems, where predator exclosures have worked for short periods only (Ekerholm et al., unpubl.). The best bet has been to work on islands in a big lake, providing efficient barriers against the invasion of mammalian predators (Hambäck et al. 2004). Such design is, however, far from perfect, because it is not possible to randomize pieces of terrain to be either outlying islands or mainland controls. Moreover, it only takes a single site visit on a windy day to realize that the results could indeed be influenced by 'non-demonic intrusions' (wave action, spray, ice piling during the breakups...), influencing plants, voles and/or researchers. The existence of some differences between islands and mainland references is thus trivial. However, the prediction of the cascade conjecture is more specific than so. The islands are predicted to display the Aldabra-Kerguelen syndrome, with vole densities way above mainland levels and with intense winter grazing leading to total destruction of woody vegetation and to expansion of herbaceous plants. The results have corroborated these rather specific predictions. Nevertheless, they have been very difficult to communicate to a scientific community, preoccupied with Hurlbert's (1984) concerns of pseudoreplication and perfect experimental design.

I hope that the above example suffices to illustrate that my 'box checking' is much more than just a sign check of differences between two population means and that Hurlbert's (2004) claim that *the probability of 'confirming' the substantive hypothesis or prediction will approach 50 percent as the number of measurements made in each experimental unit becomes large* is patently incorrect. This statement only applies to an imaginative world, where refutation of a statistical null hypothesis automatically amounts to corroboration of a scientific conjecture. Predictive science just does not work like that. It works by deriving predictions with strong, quantitative aspects, or by deriving several statistically independent predictions, whose simultaneous corroboration has low a priori probability.

Given the level of epistemological knowledge illustrated in Hurlbert's (2004) reply, it seems likely that the inductionistic view implicitly embedded in his recommendations does not emerge from a careful consideration of the viewpoints of Sir Francis Bacon (1620) and John Stuart Mill (1843) against the arguments of Popper (1963) and Lakatos (1972), but is of more implicit nature. This may not be so damaging. In the end, philosophers of science have seldom done more than explained past successes of philosophically naïve scientists. Historically, imaginativeness and common sense have been much more important for scientific successes than grasp of prevailing epistemology. The problem is that Hurlbert (1984, 2004), in effect, forbids the com-

bined use of statistics and common sense in the interpretation of results of such experiments, where replication and/or interspersion of treatments and controls have not been feasible. The challenge is to get rid of this arbitrary and counterproductive limitation without losing the positive consequences of Hurlbert's (1984) classical paper – vast improvement in our collective awareness of problems of statistics and experimental design.

Exorcising the devil of details

To detect the problematic aspects of Hurlbert's position, it is useful to compare Hurlbert's (1984) original arguments to his recent ones (Hurlbert 2004). In the 1984 paper, Hurlbert clearly regards all use of inferential statistics, based on sub-samples of a single treatment and a single control, as pseudoreplication, and objects to all explicit and implicit use of inferential statistics in the context of unreplicated experiments. Indeed, Hurlbert (1984, p. 208) treats 'implicit pseudoreplication' – i.e. the use of error bars to show the reliability of population estimates in graphical presentations – as if it were an especially dangerous form of 'pseudoreplication': *''Disallow implicit pseudoreplication, which, as it often appears in the guise of very ''convincing'' graphs, is especially misleading!''* In his response to my critique, Hurlbert (2004) sounds softer as he states that he only recommended editors to *''disallow the use of inferential statistics, when they are misapplied. This can hardly be considered controversial advice''*. So it seems – but the devil lies in details. For a reader, who has actually read Hurlbert's (1984) classic, it should be immediately obvious that 'misapplied' and 'applied' are synonyms in this context. Like Mephistopheles in Goethe's Faust, the devil of details enters the scene in the form of a fuzzy poodle, which gets its hard core first when connected to Hurlbert's previous papers.

To my understanding, Hurlbert (1984, 2004) has constructed a world of his own, where demonstration of the magnitudes of contrasts has no value whatsoever in evaluation of the possible contribution of 'non-demonic intrusion' to observed contrasts between treatment(s) and control(s), where all implicit and explicit use of statistics in the combination of imperfectly designed experiments is regarded as 'misapplication' and where it is regarded as 'error of interpretation' to use common sense when interpreting such statistically demonstrated contrasts, which cannot mechanically be regarded as treatment effects. It is this devil of details that must, in my opinion, be exorcised.

To identify the devil, let us assume that Hurlbert had been there when the pivotal experiments of Galilei were conducted. We know from first principles that two balls never hit the ground exactly simultaneously. Moreover, heavy balls, with higher weight to surface ratio, fall a bit more rapidly in earth's atmosphere than light ones. Had Galilei used 100 replicates of light and heavy balls, he had doubtlessly found that the heavier balls have a statistically significant tendency to hit the ground first. In a 'Hurlbert world', the correct interpretation had been that Aristoteles was right. Fortunately, Galilei lived in the real world, where even quantitative aspects could be weighed in and common sense could be used. We ecologists, too, should get back to the real world. Twenty years in a cold place with unreasonably harsh discipline was long enough a sentence even in Stalin's Soviet Union. But on our way back to reality, we should remember why we were sentenced to the 'Hurlbert world' in first place. Before 1984 even many top ecologists did have poor command of statistics and could put into their p-values aspects, which just were not there. Hurlbert's (1984) contribution had probably never got so uncritical response, unless the vast majority of ecologists criticized by him had in fact been ignorant of the basics of experimental design and had indeed misinterpreted the meanings of their p-values.

On our way back to the real world, let us first recognize the fact that experiments are conducted for vastly different purposes. One class of experiments consists of controlled manipulations of various physical, chemical and biological variables, where the treatment levels are fixed by the experimentalist, and where the experimentalist is primarily interested in short-term immediate responses of the system. In these experiments, the experimentalist normally tries to avoid such treatment levels, which would dramatically change the characteristics of the system, since this would make it difficult to identify direct responses of the system to different treatments and to study their interactions. Experiments of this kind are common in agriculture, forestry and medicine and have spread to basic ecology from that direction. Textbooks of biostatistics are primarily concerned with experiments of this kind. Both Hurlbert's (1984) classical paper and his recent definition of manipulative experiment (Hurlbert 2004) clearly refer to this category of experiments.

However, ecologists even conduct experiments, where the importance of a putative population or community ecological mechanism is studied by removing or by introducing an actor, predicted to play pivotal role for a given community. Here the concept of 'treatment level' has limited relevance, as the only levels of interest are undisturbed presence and total absence. Moreover, the focus is on such dynamical responses, where the system changes beyond recognition – as that is what should happen if a strong interactor had been removed or introduced. Since the experiments focus on dynamic responses of ecological systems, including components, which occur in vastly different densities, and since the outcome can be very sensitive to contamination, the

relevant manipulations must often be conducted in large spatial scales and efficient isolation of the treatment(s) from potential sources of contamination is important. (A single predator entering a predator removal area could quickly reduce herbivore densities, thus strongly ameliorating the cascading impacts of predator removal on plants.) This constrains the design options of the experimentalist. On the other hand, the focus on very strong responses reduces the potential importance of 'non-demonic intrusion'.

One way to solve the dilemma is to create a more nuanced vocabulary. We can define a perfectly designed experiment as an experiment following Hurlbert's (1984) instructions of randomization, replication and interspersion. In this context, it is reasonable to require that the treatments and controls be interspersed in all obvious and potentially relevant dimensions, not just in a two-dimensional physical space. At least in mountainous terrain, interspersion with respect to altitude is even more important than interspersion in horizontal dimensions, since the intensity of many plausible sources of 'non-demonic intrusion' (e.g. drought, waterlogging, freezing, gales, summer blizzards) correlates with altitude. Moreover, we should require the experimentalist to reproduce all obvious side effects of the treatment with appropriate placebos. If any of these conditions is clearly violated in an experiment, claimed to be perfectly designed, then the use of a stigmatizing label like 'pseudoreplication' is warranted indeed, because the *p*-values reported by the experimentalist are presented under the false pretense of maximal rigor, which the design does not actually match.

Perfectly designed experiments can be contrasted with experimental events, where the scope of the study is to check, whether a given manipulation initiates a dramatic scenario, predicted by a given conjecture. Here the strength of the experiment depends on the probability of the predicted scenario to emerge due to reasons other than the treatment. Rejection of the statistical null hypothesis of no treatment effect is neither a sufficient nor a necessary condition for such an experiment to be instructive. Even an impeccably demonstrated treatment effect to the predicted direction can amount to an ambiguous result, if the observed effect is smaller than the predicted one. Conversely, a dramatic chain of events, resulting to a huge contrast between an unreplicated treatment and an unreplicated control, can amount to strong corroboration, if such dramatic events do not belong to the normal behavior of the system. Experiments of this kind are to be evaluated by the same criteria as used in the interpretation of spontaneous events or consequences of practically motivated manipulations. Contrasts between treatments and controls could be interpreted as effects of the treatment, if the readers (including the referees) can agree that their magnitude and the timing of their emergence make alternative explanations implausible.

Indeed, the experimentalist should make it clear that the contrasts do not represent statistically demonstrated treatment effects. Nevertheless, the results could be reasonably interpreted as effects of the treatment by the same criteria by which the bending of Mercury's light was interpreted as an effect of sun's gravity, the decimation of *Opuntia* in Australia was interpreted as an effect of the introduction of *Cactoblastis*, and the dramatic changes in the benthic and terrestrial vegetation of the Aleutians, Aldabra, South Georgia and Kerguelen were interpreted as effects of introductions or recoveries of grazer or predator populations. To my understanding, the classical intertidal experiments belong to this category. They were not perfectly designed, but as experimental events, their messages were nevertheless very powerful. I am not sufficiently familiar with the systems to judge whether the lack of perfection in the designs of these experiments was a consequence of constraints or of statistical ignorance, nor is this issue especially important. As the likelihood of such large contrasts emerging exactly when the treatment was being applied could be regarded as negligibly small ($p_I \approx 0$, Oksanen 2001, Eq. 1), the changes could indeed be regarded as effects of the treatments.

Clear distinction between perfectly designed experiments and experimental events and the associated distinction between statistically demonstrated treatment effects and what can be reasonably interpreted as effects of a given treatment would be one way to re-enter the real world – to reintroduce the possibility of using imagination and common sense even in the context of experimental studies – without losing the conceptual rigor introduced by Hurlbert (1984). With this distinction, an experimentalist, who for one reason or another must deviate from the perfect design, also becomes at once aware of the limitations of such approach. Only such contrasts, which cannot reasonably emerge as consequences of 'non-demonic intrusions', will then count as interesting results, and the experimentalist must realize that the final word in this issue will be said by other colleagues. If this is clear, then pseudoreplication will indeed be a pseudoissue and error bars in graphs will not constitute 'implicit pseudoreplication' but provide valuable information, helping the reader to judge whether the divergence between the treatment(s) and the control(s) was dramatic enough to make it proven beyond reasonable doubt that the emerging contrast was an effect of the treatment.

Acknowledgements – Sincerest thanks to Göran Englund and Tarja Oksanen for constructive criticism of a previous draft and to Stuart Hurlbert for frank and instructive correspondence. The work has been supported by a grant from Vetenskapsrådet (Swedish Council for Natural Sciences).

References

Bacon, Sir F. 1620. Novum organum scientiarum. – Johannes Billius, London, UK. (cit: L. Jardine and M. Silverthorne, eds 2000. The new organon. – Cambridge Univ. Press, Cambridge, UK.)

Cajander, A. K. 1906. The struggle between plants in nature. – Luonnon Ystävä 9: 296–300 (in Finnish, a shortened English translation in Trends Ecol. Evol. 6: 294–296).

Cajander, A. K. 1909. Über die Waldtypen. – Acta For. Fenn. 1: 1–175.

Caughley, G. and Lawton, J. 1981. Plant-herbivore systems. – In: May, R. M. (ed.), Theoretical ecology, 2nd ed. Saunders, Philadelphia, pp. 132–166.

Cottenie, K. and De Meester, L. 2003. Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). – Oikos 100: 394–396.

Dayton, P. K. 1971. Competition, disturbance, and community organization: the provision and subsequent utilization of space in a rocky intertidal community. – Ecol. Monogr. 41: 351–389.

Estes, J. A. and Palmisano, J. F. 1974. Sea otters, their role in structuring near shore communities. – Science 185: 1058–1060.

Estes, J. A. and Duggins, D. O. 1995. Sea otters and kelp forests in Alaska: generality and variation in community ecological paradigm. – Ecol. Monogr. 65: 75–100.

Facta 2001. 1981. Werner Söderström Oy, Porvoo, Finland

Fretwell, S. D. 1977. The regulation of plant communities by food chains exploiting them. – Persp. Biol. Med. 20: 169–185.

Gause, G. F. 1934. The struggle for existence. – Williams & Wilkins, Baltimore, Maryland.

Gleason, H. A. 1926. The individualistic concept of plant association. – Torrey Bot. Club Bull. 53: 7–26.

Hairston, N. G., Smith, F. E. and Slobodkin, L. B. 1960. Community structure, population control, and competition. – Am. Nat. 94: 421–425.

Hambäck, P. A., Oksanen, L., Ekerholm, P., Lindgren, Å., Oksanen, T. and Schneider, M. 2004. Predators indirectly protect tundra plants by reducing herbivore abundance. – Oikos, in press.

Haukioja, E. and Hakala, T. 1975. Herbivore cycles and periodic outbreaks: formulation of a general hypothesis. – Rep. Kevo Subarctic Res. Stn 12: 1–9.

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

Hurlbert, S. F. 2004. On misinterpretations of pseudoreplication and related issues: a reply to Oksanen. – Oikos 104: 591–597.

Hurlbert, S. F. and Meikle, W. G. 2003. Pseudoreplication, fungi and locusts. – J. Econ. Entomol. 96: 533–535.

Lakatos, I. 1972. Falsification and the methodology of scientific research programmes. – In: Lakatos, I. and Musgrave, A. (eds), Criticism and the growth of knowledge. Cambridge Univ. Press, pp. 91–196.

Leader-Williams, N. 1988. Reindeer on South Georgia: the ecology of an introduced population. – Cambridge Univ. Press.

Lubchenko, J. 1980. Algal zonation in the New England rocky intertidal community: an experimental analysis. – Ecology 61: 333–344.

MacArthur, R. H. 1972. Geographical ecology. – Harper and Row, New York.

Merton, L. F., Bourn, D. M. and Hnatiuk, R. J. 1976. Giant tortoise and vegetation interaction on Aldabra atoll. I. Inland. – Biol. Conserv. 9: 293–304.

Menge, B. A. 1972. Competition for food between intertidal starfish species and its effect on body size and feeding. – Ecology 53: 635–644.

Mill, J. S. 1843. A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence and methods of scientific investigation. – Parker and Son, London, UK. (cit: shortened version, E. Nagel. ed. 1974. Hafner, New York.)

Moen, J. and Oksanen, L. 1998. Long-term exclusion of folivorous mammals in two arctic-alpine plant communities: a test of the hypothesis of exploitation ecosystems. – Oikos 82: 333–346.

Moen, J., Gardfjell, H., Oksanen, L. et al. 1993. Grazing by food-limited microtine rodents on a productive experimental plant community: does the "green desert" exist? – Oikos 68: 401–413.

Murdoch, W. W. 1966. Community structure, population control, and competition-a critique. – Am. Nat. 100: 219–226.

Nationalencyclopedin. 1990. Bra Böcker AB, Höganäs, Sweden.

Norrdahl, K., Klemola, T., Korpimäki, E. et al. 2002. Strong seasonality may attenuate trophic cascades: vertebrate predator exclusion in boreal grassland. – Oikos 99: 419–430.

Oksanen, L. 2001. Logic of experiments in ecology: is 'pseudoreplication' a pseudoissue? – Oikos 94: 27–28.

Oksanen, L. and Moen, J. 1994. Predictability of plant responses to the exclusion of grazers in three Fennoscandian tundra habitats. – Ecoscience 1: 31–39.

Oksanen, L. and Oksanen, T. 2000. The logic and realism of the hypothesis of exploitation ecosystems. – Am. Nat. 155: 703–723.

Oksanen, L., Fretwell, S. D., Arruda, J. et al. 1981. Exploitation ecosystems in gradients of primary productivity. – Am. Nat. 118: 240–261.

Paine, R. T. 1966. Food web complexity and species diversity. – Am. Nat. 100: 65–75.

Paine, R. T. 1974. Intertidal community structure: experimental studies on the relationship between a dominant competitor and its principal predator. – Oecologia 15: 93–120.

Paine, R. T. 1980. Food webs, linkage, interaction strength and community infrastructure. – J. Anim. Ecol. 49: 667–685.

Polis, G. A. and Strong, D. R. 1996. Food web complexity and community dynamics. – Am. Nat. 147: 813–846.

Popper, K. 1963. Conjectures and refutations. – Harper & Row, New York.

Rhoades, D. F. 1985. Offensive-defensive interactions between herbivores and plants: their relevance in herbivore population dynamics and ecological theory. – Am. Nat. 125: 205–238.

Seldal, T., Andersen, K.-J. and Högstedt, G. 1994. Grazing-induced proteinase inhibitors: a possible cause for lemming population cycles. – Oikos 70: 3–11.

Strong, D. R. 1992. Are trophic cascades all wet? Differentiation and donor-control in speciose ecosystems. – Ecology 73: 747–754.

Werth, E. 1928. Überblick über die Vegetationsgliederung von Kerguelen sowie von Possession-Eiland (Crozet-Gruppe) und Heard-Eiland. – In: von Drygalski, E. (ed.), Deutsche Südpolar-Expedition 1901–1903. Teil 8. de Gruyter Verlag, Berlin, pp. 300–326.

White, T.C.R. 1978. The importance of a relative shortage of food in animal ecology. – Oecologia 3: 71–86.

Whittaker, R. H. 1975. Communities and ecosystems. – MacMillan.