Vol. 191: 301–309, 1999 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published December 30

THEME SECTION

Negative results

Idea and coordination: Howard I. Browman

The uncertain position, status and impact of negative results in marine ecology: philosophical and practical considerations

Howard I. Browman

Institute of Marine Research, Austevoll Aquaculture Research Station, 5392 Storebø, Norway
E-mail: howard.browman@imr.no

MEPS Theme Sections (previously referred to as Comment Sections) represent integrated expert analyses highlighting an important, cutting-edge topic. Theme Sections (TS) are organized by a Contributing Editor (e.g. Browman 1995, 1996). The present TS addresses a topic which has been discussed, condemned or defended by authors and editors for decades: negative results of research efforts.

A rejected manuscript was my motivation for organizing this section. There was nothing wrong with the manuscript itself; the hypothesis was clear and concise and the experiment to test it was appropriate. However, both the editor and the reviewers objected to the fact that the result reported was negative. It was the first time in my career that I had prepared a manuscript about a non-result. The potential value of data which do not support a research hypothesis is often not appreciated. Nevertheless, 'negative' results may be very important for several reasons: they may provide more balance for a subject area thus far supported only (or primarily) by positive results (e.g. the impacts of solar ultraviolet B radiation). They may indicate that a subject area is not as mature or clearly defined as previously suspected (e.g. the first reports of reverse diel vertical migration). They may show that a particular line of research is not worth further efforts (e.g. trophodynamic modelling), or that our current methodologies are inadequate for producing a definitive result (e.g. predicting recruitment in groundfish populations).

The concept of negative results is rather fuzzy. In order to provide broader coverage of its many possible meanings, I sought contributions from long-standing editors of marine science journals, senior scientists/educators, and historians/philosophers of science. Only the latter category responded with enthusiasm. It proved difficult to recruit marine ecologists (fortunately, A. J. Underwood accepted the challenge). Further, it was impossible to get any editor onto my hook (believe me, I tried). Thus, this TS itself produced a largely negative result. Hence, the coverage is not as comprehensive as I would have liked.

Declining my solicitation to contribute, Stephen Jay Gould (Professor of Geology, Museum of Comparative Zoology, The Agassiz Museum, Harvard University) wrote: 'Your suggestion for a forum on publishing negative results in science represents a most important project. In my opinion, this is perhaps the most important effectively undiscussed subject in the entire methodology of science.' In fact, Professor Gould had already written an essay on this topic (Gould 1993)—the main reason for declining my request. In this essay he states: 'The importance of negative results—nature's apparent silence or nonacquiescence to our expectations—is also a major concern in science. Of course, scientists acknowledge the vitality of a negative outcome and often try to generate such a result actively—as in trying to disprove a colleague's favored hypothesis. But the prevalence of negative results does pose an enormous, and largely unaddressed, problem in the reporting of scientific information.' Professor Gould asserts that positive results tell more interesting stories than negative results and are, therefore, easier to write about and more interesting to read (a privileging of the positive). He contends that this may lead to a bias which acts against the propagation of negative results in the scholarly literature. This has been borne out in recent surveys of the medical literature which discuss 'publication bias': studies showing positive results from drugs are published faster and more often than studies showing neutral or negative results, producing a bias that shows drugs in a favourable light (e.g. Johansen & Gotzsch 1999, Rennie 1999). These articles, and Gould's essay, deal with several aspects of the negative results issue not addressed in this TS, and so I encourage readers to look them up.

The issue of negative results remains complex. It reflects our training, our thoughtfulness about what we do as scientists (and how we do it), and our humanity, with all its inherent biases. Hopefully, the essays that follow will provide MEPS readers with a more concrete introduction to the issue.

LITERATURE CITED

Browman HI (1995) Commentaries on current research trends in recruitment studies. Mar Ecol Prog Ser 128:305–310

Browman HI (1996) Predator-prey interactions in the sea: commentaries on the role of turbulence. Mar Ecol Prog Ser 139:301–312

Gould SJ (1993) Cordelia's dilemma. Nat Hist 2/93:10–18

Johansen HK, Gotzsch PC (1999) Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. J Am Med Assoc 282:1752–1759

Rennie D (1999) Fair conduct and fair reporting of clinical trials. J Am Med Assoc 282:1766–1768

When is a negative result anomalous?

Michael Ruse

University of Guelph, Department of Philosophy, 367 MacKinnon Building, Guelph, Ontario N1G 2W1, Canada

E-mail: mruse@uoguelph.ca

Some years ago, I served on a committee of the (American) National Academy of Sciences, putting together a booklet on proper conduct in science, paying particular attention to the problem of fraud. I was there, as a philosopher and historian of science, to add a different dimension to the learning of the various (very distinguished) scientists, and as is my wont I let everyone have my opinion, at length, on every possible occasion. I was certain that fraud was a bad thing and a troublesome thing (not necessarily identical qualities), because the concocted results lead us all astray and we waste masses of time putting things right, if we ever do.

One of the scientists on the committee was the molecular biologist from MIT, Philip Sharp, then heavily favoured for a Nobel Prize, which honour came his way just a few years later. I found that he clearly did not share my concerns — sceptical would be a polite way of putting his extreme disbelief. His was a strong conviction that the issue of fraud was much overblown and not the threat that I (like the media and the U.S. Congress) then took it to be. Our disagreements were not personal, and finally in one of the intervals between discussion, I challenged him on this. I suggested that Sharp's position was almost as if he thought that fraud was no big deal because scientific work — experimentation particularly — was really no voyage of discovery. It was no thrust into the unknown where a phony result could destroy the work and happiness of many, but a stylized dance where experiment was really more a frill than an adventure of discovery. A concocted result was no big deal because if it were true, it did no harm and if it were false no one would believe it anyway.

I should say by way of background that our committee was meeting at the time of the brouhaha over cold fusion. The claim was widely disbelieved, and subsequently proven to be false. The physicists on the committee were scathing and quite dismissed any appeal for supportive evidence. 'Like trying to drive a nail into a beam of oak using a pound of butter,' as one of them put it to me. Sharp, responding to the question of whether scientists generally know the outcome of experiments beforehand, sided with the physicists and replied at once without hesitation: 'Of course that is so. It is silly nonsense to say that you don't know what will come up. I would never dream of doing an experiment where I don't know the results before I set out. It would be a criminal waste of expensive equipment and chemicals, and an unwarranted waste of the time of my grad students and post-docs.' Then he paused, and I really do not think it was for effect. 'Of course,' he said thinking back, ' there have been a couple of times when the experiments didn't work out. That's when it gets really exciting.'

I have often thought about Sharp's comments, and have tried them out on many people. I think now I have a bit of a grasp of what he meant. I think also that his comment goes beyond the question of fraud and tells you something about negative results as well as positive results — results, that is, where the work has been done honestly and fairly. If you take a position on science influenced by the philosopher Karl Popper (1959) (which I was doing rather at the time of the committee), then Sharp does not make much sense. Experimentation is a voyage of discovery: trying out bold conjectures and hypotheses, and seeing what will come up. All results count, equally in a sense, and this applies to negative results as much as positive ones. There is no such thing as bad information, and indeed a case can be made for saying that negative results are more important than positive ones. Famously, for Popper the name of the game is falsifiability — the aim of the scientist must be to show false the most cherished of hypotheses — and a negative finding is the best possible grist for the mill.

But take now the alternative philosophy of science of Thomas Kuhn. Kuhn (1970) argues that successful science occurs within paradigms, and that the scientist as such never challenges the paradigm. At least, the scientist in everyday life ('normal science') never challenges the paradigm. This means that negative results are not exciting and not significant, because they do not and cannot challenge the paradigm. Positive results at least have the virtue of burnishing the paradigm, like a hymn of praise in a church service, and perhaps even extend the paradigm's scope. But negative results are worthless because at most they reflect on the scientist's inadequacies, just as failure to finish a crossword tells us nothing about the crossword, but much about your lack of word power.

A negative result comes about because you have not done the experiment properly, or because you have chosen the wrong organism to which your model cannot apply, or because your equipment is dirty or your statistics inappropriate, or some such thing. And who cares about that — except you the next time the granting agency meets?! Negative results are certainly not worth publishing, and editors who know their business keep them out. There is quite enough without explicit record of human failure.

But Sharp (and Kuhn) show that there is something more than this. Sometimes negative results do count. In Kuhn's language these are *anomalies*, showing that there is something wrong with the paradigm. And obviously these should be seized on and used and published. The problem is how we are to distinguish anomalous results from just plain run-of-the-mill negative results. In a way, the only answer is hindsight and history. A negative result only becomes an anomaly when someone shows that it will fit into a new paradigm and leave the old one behind. But how is one to tell beforehand if a result is an anomaly and thus worth publishing or simply negative and to be ignored and go unreported?

At one level, it takes genius, something most editors do not have — or pretend to have. The trouble is that there are so many geniuses who turn out not to be. Swans who turn into ugly ducklings. Any editor knows only too well how many papers come their way full of strange results and wonderful new theories to account for them. (I should say that the results are rarely found by these authors themselves, but gleaned from eclectic reading. The Creationist literature is a paradigm.) But how does one distinguish interesting negativity? Obviously track record is important — someone who has found interesting things in the past is worth listening to in the present. And combined with this is experience. Doing science is like auto mechanics — it is a skill as much a book learning. A first-rate mechanic just knows when something strange is up, even if he cannot articulate his feelings. The sound is just not right. Similarly, a first-rate experimentalist just knows when an experiment's failure is interesting. He knows how reliable his test organisms or his equipment or whatever are. He knows when negativity might be more than that.

In a perfect world, the experimentalist might then be able to go on to explain what is up, and offer a new explanation — show why the paradigm must be rejigged in this case or perhaps even rejected and replaced. But science is not done in a perfect world, and sometimes (often?) experimentalists have to turn to others — theoreticians — for help in interpreting their negative results. The best they can offer is the negative conclusion, hoping that others will show why it is significant.

And all of this goes to show, I trust, why editors are rightly unimpressed by negative results and loathe to publish them. And why it is a jolly good thing that editors do not always get their own way, and negative results are somehow circulated through the scientific community.

LITERATURE CITED

Kuhn TS (1970) The structure of scientific revolutions. University of Chicago Press, Chicago

Popper KR (1959) The logic of scientific discovery. Harper Books, New York

Negative results as positive knowledge, and zeroing in on significant problems

Douglas Allchin

Minnesota Center for the Philosophy of Science, University of Minnesota, Minneapolis, Minnesota 55455, USA

E-mail: allchin@umn.edu

According to philosopher of science Karl Popper, no researcher could be happier than to find negative results. In his view, one cannot prove theories conclusively, beyond all doubt — even with positive results. Positive evidence may accumulate, but because no one ever has all the facts, one never knows whether an alternative theory (perhaps not yet articulated) may be correct. At best, he claimed, one can falsify a hypothesis using results that clearly contradict it. And Popper was partly right: sometimes, 'negative' evidence can productively guide research away from conclusions that might otherwise have seemed reasonable.

But Popper was no research scientist. His claims betrayed an idealization of science as governed by relatively simple formal logic and expressing all its conclusions in the form of universal laws. Under such conditions a sole exception, or anomaly, is a deathblow. In practice, the art of falsification is more subtle. Researchers must consider methodological assumptions, statistical analyses, details of experimental design and test conditions that Popper never fully addressed. Even fellow philosophers sympathetic to Popper now consider his views deeply flawed (e.g. Lakatos 1978, Kuhn 1970, Mayo 1996). Assembling negative results and interpreting them effectively is no simple task.

Still, Popper's arguments are a valuable reminder of the qualified role of even 'positive' results. Confirmation or agreement from a single experiment or study (or even several) is not always reliable. For example, positive instances themselves do not necessarily exclude

the possibility of significant counter-instances. Nor do they address possible alternative explanations for the same results: theories can, and often do, overlap in their predictions. Results that are merely consistent with a proposed theory or explanation should not wholly convince us. Our brains tend to seek confirmation of our beliefs and discount perception of instances that challenge them. But no one regards this cognitive bias as good science. Even positive results deserve skeptical analysis.

To rely on positive results means also to rule out experimental error and alternative explanations at the same time. Theories must not just pass tests; they must pass severe tests. They must survive a likely opportunity to fail. In a sense, the researcher must invite or aim for potential negative results (to expose them should they exist). While Popper hinted at a notion of such tests, Mayo (1996) articulates the nature of severity more fully. First and foremost is the concept, familiar to every research scientist, of controls. Some controls are central to the experimental design. They may be at the core of assessing the relative warrant of 2 theoretical maps — or, they may help assess the empirical domain or scope of a particular concept. Other controls are accessory: they are checks to ensure that the intended experimental conditions are actually realized. Controls help rule out error. Second is the concept of error statistics. Since much reasoning about experimental data is statistical, one needs to measure its precision (and hence, reliability). We turn to p-values, error bars and the like as important labels of the rigor of a test and/or its conclusion. Tests must also have statistical power, or discriminatory sensitivity. Such methods strongly shape our confidence in experimental conclusions. They assess the uncertainty and the chance of error. The meaning of positive results depends on the various methods we use for assessing their reliability.

Just so for negative results. Measures of reliability apply to positive and negative findings symmetrically. 'Negative' results can thus represent 'positive' knowledge when we are confident of the conclusions. A measure of departure from theoretical expectations, for example, or rejection of a null hypothesis, can be statistically significant and, hence, noteworthy. What matters are the controls and the statistical analysis of the data, not the 'negative' dimension of the conclusions. (Here, I hedge the question of what should constitute statistical significance: such standards will vary at least among fields of study and their historical development.) There is a tendency among scientists (in their casual rhetoric) to undervalue so-called negative results. Instead, a researcher should embrace negative results—when they are reliable. 'Wrong' outcomes may be personally disheartening, but they can nonetheless be meaningful. By contrast, one should disvalue inconclusive results that leave only uncertainty. Ultimately, what matters is the severity of the tests, not whether results agree or disagree with theoretical expectations.

Under this perspective, researchers should guard against uncertain results, not negative results. That is, one should fear experimental outcomes being ambiguous or inconclusive, rather than being 'wrong' according to an established hypothesis. Certainly this is why one typically invests so much statistical effort in advance to ascertain the minimal data collection. Hence, if a study is worth doing at all, it is worth doing well. Prospectively, it ought to deliver significant results—worth publishing regardless of the specific outcome. If not, then perhaps the experiment needs to be redesigned. The question itself should be posed or framed experimentally to deliver an answer that will matter. The experimenter must risk failure. Else why investigate? Thinking must shift from an exclusively right/wrong distinction to include a certain/uncertain distinction. 'Positive' knowledge is defined by being certain, not by being either right or wrong. The fundamental aim is reliability. The lesson is key—and one worth instilling in students.

Of course, not all scientific investigations are experimental (in the sense of testing a clearly formulated hypothesis). Some are exploratory. For example, one may search for a new, more effective methodology. Or one may try to isolate an unfamiliar phenomenon and tease it into relief, with no clues yet about what causes it. Here, there are no well-formed theories, no clear null hypotheses. Such explorations are generally riskier, since no one knows even how to proceed. The possibilities may be immeasurably many, or ill-defined. That is, one cannot ensure that any search is exhaustive. The researcher hopes to get lucky (on a hunch, perhaps) or to capitalize opportunistically on a chance observation. Indeed, a successful result can *demonstrate* the way to go (Allchin 1992). A 'negative' result, however (as Kuhn 1970 noted), indicates merely an incomplete recipe, nothing about the impossibility of 'positive' results. Failures of this type are less informative. Very little is ruled out. No finding can significantly guide others. Uncertainty remains. Information may pass along informal channels among colleagues, but it is not the stuff of publication. But then, the scientist knows at the outset that such research is risky. In contrast to the results characterized above, a 'negative' result here is really a *non*-result.

Why might genuine negative results have developed such an unfavorable image, especially in publishing? Experiments have yet another important dimension: relevance. Publishing merit depends not only on the soundness or reliability of the conclusions, but also on the fruitfulness of the information to others (Hull 1988).

Will the information be valuable to future applications or research? Are the conclusions novel? Namely, does it warrant communicating to peers and/or registering in an archive? (One asks the very same questions in advance, in evaluating grant applications.) Research that does not address significant problems leads nowhere. (Who cares if the conclusions are reliable?) Hence, an additional screen or filter characterizes acceptability or value for publication: is the problem posed by the research significant? *Science* and *Nature* earn their prestige as premier journals based not so much on the reliability of the published results as on the widespread relevance of the studies. Does 'negative results' sometimes refer to results that matter to hardly anyone—even if positive? 'Negative', here, may be a code word for irrelevant or insignificant. When negative results are important (along the terms suggested above), they are indeed published. Moreover, their publication profile generally reflects both the experimental rigor and their signficance for current discourse in the field. Negative results do get press—sometimes very good press. That is the stuff of scientific revolutions, great and small.

Ultimately, then, negative results can be positive knowledge. It depends on an experimental design that supports clear (certain) conclusions. But it is equally, if not more important, to zero in on significant problems.

Acknowledgements. The author appreciates support from NEH #FS-23146.

LITERATURE CITED

Allchin D (1992) How do you falsify a question?: crucial tests v. crucial demonstrations. PSA 1992 1:74–88

Hull D (1988) Science as a process. University of Chicago Press, Chicago

Kuhn TS (1970) The structure of scientific revolutions, 2nd edn. University of Chicago Press, Chicago

Lakatos I (1978) The methodology of scientific research programmes. Cambridge University Press, Cambridge

Mayo D (1996) Error and the growth of experimental knowledge. University of Chicago Press, Chicago

The role of negative evidence in science

David L. Hull

Department of Philosophy, Northwestern University, Evanston, Illinois, 60208, USA

E-mail: d-hull@nwu.edu

Do manuscripts presenting negative evidence have a more difficult time getting published than other sorts of manuscripts? Common wisdom has it that they do, but too often what everybody knows turns out to be false. The phrase 'negative evidence' implies that this evidence has implications for some hypothesis or other. A datum cannot count as 'evidence' unless it is evidence for or against some hypothesis or other. Hence, pure data papers, to the extent that they exist, do not count one way or the other with respect to this problem. The issue then becomes the character of these hypotheses. Are they of major importance, concerning the fundamentals of some area of science, or more limited in their scope? A second issue is whether any evidence can be purely negative; that is, counting against one hypothesis but not counting for any alternative hypotheses. Finally, does it matter whether or not the hypothesis in question has already appeared in the literature and become fairly well known or whether the author of the manuscript thought it up himself and then shot it down?

'Negative evidence' can refer to evidence that shows that a central tenet of some particular area of science is false. At one time, evolutionary biologists thought that natural selection was so powerful that deleterious mutations would be rapidly eliminated from populations. As a result, natural populations should be (genetically) quite homogeneous. The results of gel electrophoresis showing that natural populations are surprisingly heterogeneous count as negative evidence. Yet these papers were published quite rapidly (Hubby & Lewontin 1966, Lewontin & Hubby 1966). This evidence counted against the view that natural selection is so powerful and sure-sighted. It also counted for hypotheses about the mechanisms responsible for genetic heterogeneity, e.g. heterozygote superiority.

The same story can be told for less fundamental hypotheses. For a long time, biologists thought that the relation between monarch and viceroy butterflies was one of Batesian mimicry. Monarch butterflies are protected from predation because they taste bad to the birds that feed off them. So the story goes, the viceroy gains some protection from predators by looking like the nasty tasting monarch. However, a very simple experiment raised significant doubts about this belief (Ritland & Brower 1991). Viceroys also taste bad. The relation is more likely to be one of Mullerian mimicry (both species get mutual benefit from tasting bad and looking like each other) or a mixture of Batesian and Mullerian mimicry. Once again, evidence against one hypothesis turns out to be evidence for alternative hypotheses.

Finding examples in which negative evidence counts against one view and for no others is extremely difficult. The closest that I can come to such an example is the precession of the perihelion of Mercury. For a long time Mercury did not obey Newton's laws of celestial mechanics very well. The evidence that showed that Mercury was frequently not where it should be counted as negative evidence with respect to Newton's laws. However, it did not count for any

alternative theory. Advocates of theories alternative to Newton's theory could explain lots of data but not the precession of the perihelion of Mercury. At the time, no one could. As it turned out, no modification of Newton's theory was able to accommodate these recalcitrant data. That took Einstein's theory.

Evidence that counts for or against fundamental tenets of science is very likely to see print. Let anyone come up with evidence that seems to favor Lamarckian over Darwinian inheritance, and it is sure to get a wide airing. But very rarely is negative evidence so momentous. How about evidence for or against work-a-day hypotheses? The relation between monarch and viceroy butterflies is very limited and particularized. Hundreds of cases of such mimicry have been studied. Evidence against the viceroy example appeared rapidly in *Science* because this is the paradigm example of Batesian mimicry. Just about every text on evolutionary theory includes color plates of these two butterflies. Anyone challenging the less well-known examples of mimicry would have a harder time making it into the pages of widely circulated journals. Instead authors of such papers would have to submit their findings to more specialized journals.

Another factor affecting the fate of negative evidence is the inherent plausibility of the hypotheses that it bears on. For example, the notion of overall similarity has seemed extremely plausible to biological systematists for a long time. One possible goal for systematists is to produce classifications of plants and animals that reflect increasing degrees of overall similarity. One group of systematists (numerical taxonomists) attempted to refine the notion of overall similarity and make it more explicit. The results were the opposite of what they had anticipated. As it turns out, too many alternative measures of overall similarity were formulated, and no reasons could be found for preferring one over all the others. Overall similarity turned out to be an illusion. No such thing exists or, put differently, too many equally plausible alternatives exist. Initially, papers attempting to make the notion of overall similarity more explicit were published because they looked as if progress was being made, but gradually the conclusion that overall similarity is a delusion seemed increasingly inevitable. As negative as these papers might be, they were published to warn off later biologists.

I think that papers that fulfill this function of negative evidence might have a hard time getting published. One author or series of authors publishes papers presenting the results of experiments that they have run. Other authors find what they take to be errors in these papers. Can they get these corrections published? Another possibility is that a novel hypothesis occurs to a scientist. This scientist proceeds to test his bright idea only to discover that it is mistaken. Might he or she publish these findings to ward off other scientists who might be tempted to pursue this dead end? Journals do have arrangements for publishing corrections, but the widespread belief is that publishing corrections is very difficult. One alternative is to write a paper that presents positive evidence for your views and include these corrections as well. Since this solution seems so obvious, I suspect that lots of scientists employ it already without any great fanfare.

All of the preceding is an example of armchair 'philosophizing'. Given what I know in general about science, these are the most obvious distinctions and most likely results, but I do not know how real editors and referees behave with respect to these issues. I have done the necessary research for other questions about publication in scientific journals (Hull 1983, 1988), but I have not conducted any of the studies relevant to the problem of negative evidence. If I did, could I get these results published, especially if they were negative? I think so. Refuting what everyone believes is important enough for journals to publish these refutations. But I don't know.

For example, a common belief among biologists is that molecular biology is rapidly driving traditional whole-organism biology out of existence. Departments with names that include such terms as 'evolution', 'ecology' and 'environment' are being closed down all across the US to be replaced by departments of molecular biology. A student of mine, Nicole Ducharme, ran a small study to test this hypothesis, and to her dismay her results were negative. She could find no such trends. We will see how much difficulty she finds in getting these negative results published. I suspect that she will succeed because she is refuting a widely-held belief of prime importance to a lot of biologists. Negative evidence with respect to some minor finding that is of interest to almost no one is quite another matter.

One of the problems with all of the preceding examples is that they concern papers that did get published. If they had been rejected, I would have no way of knowing of their existence. What we really need is a study that includes manuscripts that did *not* get published as well as those that *did*. Were manuscripts presenting largely negative evidence rejected more frequently than other sorts of manuscripts?

LITERATURE CITED

Hubby JL, Lewontin RC (1966) A molecular approach to the study of genic heterogeneity in natural populations. I. Genetics 54:577–594

Hull DL (1983) Thirty-one years of systematic zoology. Syst Zool 32:315–342

Hull DL (1988) Science as a process. University of Chicago Press, Chicago

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterogeneity in natural populations. II. Genetics 54:595–609

Ritland DB, Brower LP (1991) The viceroy butterfly is not a Batesian mimic. Nature 350:497–498

Publication of so-called 'negative' results in marine ecology

Antony J. Underwood

Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories, A11, University of Sydney, New South Wales 2006, Australia

E-mail: aju@bio.usyd.edu.au

A logical framework for falsification. I assume in these comments that the only results worth publishing at all are those based on some consistent and coherent framework which defines the purpose of the study, the rationale for the particular measurements to be made (whether observational, 'mensurative' [Hurlbert 1984] or manipulative), the domain to which the results are supposed to apply, and the context in which the results can be interpreted. These notions themselves dictate the rationale for choice of particular methods, the appropriateness of sampling and experimental designs and, of course, wherever appropriate and necessary, the form and interpretation of any statistical analysis to be used.

In other words, there must be a defined and defensible logical framework. Among others available, the framework I advocate is that described in full elsewhere (Medawar 1969, Heath 1970, Underwood 1990, 1991, 1997a). A study or phase of a research programme begins with observations (problems, patterns, puzzles) and it proceeds in an orderly attempt to explain those observations. One (or usually more) explanatory models (theories, conceptual frameworks) is proposed which can account for the observations. To contrast among such different models, from each is deduced a testable hypothesis (or a series of hypotheses). Ideally, these are 'bold' in Popper's (1969) sense of being imaginative or visionary. More importantly, they must predict different things from the basis of the different models. Hypotheses, therefore, must consist of predictive statements (by definition, the results predicted cannot be known to the predictor) and should provide the maximal possible contrasts for the various models. The study then creates the circumstances (a manipulative experiment) or visits to sites where the circumstances exist, to provide an experimental test(s) of the hypothesis(es).

Because of the well-known and repeatedly demonstrated fact that inductive reasoning is irrational (Hume 1779), it is usually customary (and often statistically imperative; Underwood 1990, Winer et al. 1991) to turn the hypothesis into its logical complement — the null hypothesis. Thus, a prediction that 'Action X (e.g. removal of predators) will cause an increase in some variable (e.g. density of prey)' is converted to the null hypothesis that 'Action X will cause no change or a decrease in the specified variable.' The experiment then sets out to attempt to disprove the null hypothesis.

If the null hypothesis is demonstrated to be wrong (or is statistically improbable), the hypothesis and model are supported (e.g. Simberloff 1983, Connor & Simberloff 1986, Underwood 1990, 1991). The logical basis for such a conclusion by falsification is well known (e.g. Trusted 1979). Alternatively, if the null hypothesis is retained (i.e. the experiment failed to disprove it), the hypothesis and model have been falsified.

In neither case is the single experiment the end of the study. Where models have not been supported, new models must be proposed — which must now include explanation of the observations gained during the experiment which falsified the previous model(s). Where a model is supported, it needs to be probed and tested more stringently, by proposing more general or more specific hypotheses (conjectures that are more bold; Popper 1969). These, in turn, must be tested.

In such a procedure, results set in a logically consistent and clearly defined framework for the experimental processes can *never* be considered negative. In either possible outcome of an experiment, the results falsify the null hypothesis, thereby providing support (so far) for a model, OR they falsify a model or series of models requiring it (so far) to be inadequate to explain the previous knowledge (i.e. making it useless).

Bias and Type I error: obsessions with α. Biologists in general, and ecologists in particular, are obsessed with not making Type I errors. They (we) have great reluctance to err by disproving a null hypothesis when it should really be retained (this is a Type I error). To keep the probability of such errors small (and conventionally choosing a probability of $p = 0.05$), experimenters run serious risks of increasing the probability of a Type II error. Type II errors are failures to reject a null hypothesis (failure to support a model) when it is actually correct (e.g. Cohen 1977, Winer et al. 1991, Underwood 1997a). Nowhere has this obsession become more odd than in analyses of environmental sampling and experiments where Type II errors are failures to find impacts because sampling is inadequate (e.g. Fairweather 1991, McDonald & Erickson 1994, Mapstone 1995, Gray 1996, Underwood 1997a, b).

Obsessions about α — the probability of Type I errors — have led to publication of numerous erroneous studies (1 in 20 at $p = 0.05$!). At the same time,

pre-occupation with α must have led to a failure to publish many potentially useful studies (of unknown, but possibly large β, the probability of Type II error). This leads to philosophical conundrums. For example, what is the size of scientists 'allowable' life-time pool of mistakes? If you have already done 60 experiments leading to published, statistically significant results, using $p = 0.05$ to set α, can you be allowed to go on publishing papers? By now, 3 of the 60 are, on average, wrong. How many more wrong studies may you publish? As an alternative problem, suppose you analyse a set of 4 variables out of 5 collected in some study and publish the results using Bonferoni corrections to the statistical tests so that α is maintained as $p = 0.05$ over all the variables. Now someone else (or you) decides later to analyse the 5th variable. All tests must be re-adjusted to retain α at $p = 0.05$! The previous publication should be corrected—leading to the irrational conclusion that its significant results may have always been a Type I error—due to a test on a variable that you previously chose not to do!!

Such games are unhelpful, but underline the point that obsessions with Type I error are not necessarily healthy obsessions.

Post hoc panicking about the power of experiments. The alternative mode, that is increasingly becoming widespread (particularly in some areas of ecology), is to deal with so-called negative results (failures to reject a null hypothesis) as though they must be Type II errors. Consider a study in which spatial differences in the survival of juvenile fish on a coral reef are attributed to (explained by the model of) predation by larger fish. This leads to the hypothesis that removal of predatory fish should lead to an increase in survival of juvenile fish compared with that in control areas where predators forage naturally. The experiment is done and results in a small (say 10%) difference in survival that is not statistically significant. Often, at this point, sense is abandoned and panic creates retrospective or post-hoc power analyses. These calculate the probability of rejecting the null hypothesis (no change or decrease in survival) in favour of the observed alternative of a 10% increase in survival. Commonly, the outcome is a small power—say $p = 0.40$ (or β, the probability of Type II error, is large, at $p = 0.60$). This leads to irrational consequences—'if the experiment were done with many more replicates, it would have been powerful ($p = 0.95$; $\beta = 0.05$) and the null hypothesis would have been rejected' or, much worse, 'predation really does explain the original observations; the experiment was not good enough.' Alternatively, for the existing size of experiments, power can be calculated for different probabilities of Type I error until probabilities of Type I and Type II errors are the same. In this case, suppose that $\alpha = 0.40$ and $\beta = 0.40$ for the observed 10% difference in survival. If you now choose to reject the null hypothesis at $p = 0.40$, it would have been rejected. Predation can thus be demonstrated to be the correct model because α can be altered until the desired result is achieved!

Such bizarre games are statistically unsound, logically indefensible and scientifically reckless. They stem from a failure to have specified in advance *how much* increase in survival should occur when predators are removed. This is the 'effect size' for the statistical test (Cohen 1977, Winer et al. 1991). Once defined, power can properly be calculated *before* the experiment and the experiment designed to have large power (small β) for any pre-chosen level of α. Such a pre-designed experiment has none of the arm-waving associated with the performance described above!! Note that, in this case, and in many other cases, the effect size (the amount of increase in survival that should occur when predators are removed) is defined by the original observations (Underwood 1997a, b). It is the amount of difference observed that the model of predation was proposed to explain. There is no reason not to use the procedure properly.

Repeated experimentation. Of course, as any thinking biologist already knows, all of these issues would become less of an apparent problem if people were not so dependent on publishing results of single, unrepeated experiments. In the simplest case of an experiment with 3 treatments (predators removed, unmanipulated areas, controls for removal of predators [e.g. fences, cages]), each with $n = 5$ replicates, there may be a probability of Type II error (β) equal to 0.30 at a probability of Type I error (α) of 0.05 for some predetermined effect of predation. The experiment is done and fails to reject the null hypothesis at $p = 0.05$, but the observed effect is about the size of that specified from the observations. So, do the experiment again. Suppose, again, there is no significant effect. The probability of this being due to a Type I error is now 0.09 (i.e. β^2); power is 0.91. Do it a total of 3 times. If there is no significant difference over the 3 experiments, the probability of this being due to Type II errors is now 0.027—a very unlikely result.

Of course, if any of the experiments demonstrates as significant the predicted effect of predators, there are 2 results. Predation is sometimes important *and* some clues may have been provided as to the circumstances where or when such a process matters. There is a vast amount of extra information available from repeated experimentation (generality of circumstances; variation in intensity; consistency over seasons, etc.). These gains provide even more convincing rationales for repeated experimentation—in addition to drastically reducing the indecision about interpretations of failures to reject null hypotheses.

Editorial responsibility and progress in ecological science. Publication of so-called negative results requires at least 2 of 3 criteria to be met, to prevent the literature from being swamped:

(1) The basis for the study must be clearly articulated, so that interpretation of results in relation to observations, theories and hypotheses is transparent. This will ensure that the value of the failure to support predictions can be discerned.

(2) The power of the experiment (where this can be calculated) to detect the predicted (i.e. *specified a priori)* quantitative patterns is presented and discussed, so that readers can make an informed judgement about the meaning of the results.

OR (3) Where power is undefinable (because the hypotheses must, for some reason, be imprecise or very general), there must be adequate discussion of why the results can be accepted as a valid demonstration that a hypothesis is wrong. In addition, the steps that are going to be taken to determine whether or not the results are a robust demonstration of a rejection of theory, or are more likely to be a problem of inadequate sampling or other form of small power in the experiment, must be specified.

Clearly, ensuring conformity to these principles requires vigilance by referees, consistent decision making by editors, and proper attention to the nature of evidence by everyone concerned — particularly the authors.

Acknowledgements. This paper was prepared with support from the Australian Research Council through the Centre for Research on Ecological Impacts of Coastal Cities. I am grateful to Dr M. G. Chapman for her helpful comments.

LITERATURE CITED

Cohen J (1977) Statistical power analysis for the behavioural sciences. Academic Press, New York

Connor EF, Simberloff D (1986) Competition, scientific method and null models in ecology. Am Scient 75:155–162

Fairweather PG (1991) Statistical power and design requirements for environmental monitoring. Aust J Mar Freshw Res 42:555–568

Gray JS (1996) Environmental science and a precautionary approach revisited. Mar Pollut Bull 32:532–534

Heath OVS (1970) Investigation by experiment. Edward Arnold, London

Hume D (1779) Dialogues concerning natural religion 1779, 2nd edn, 1947. Nelson, London

Hurlbert SJ (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211

Mapstone BD (1995) Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. Ecol Appl 5:401–410

McDonald LL, Erickson WP (1994) Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed? In: Fletcher DJ, Manly BFJ (eds) Statistics in ecology and environmental monitoring. University of Otago Press, Dunedin, p 183–197

Medawar P (1969) Induction and intuition in scientific thought. Methuen, London

Popper KR (1969) Conjectures and refutations. Routledge & Kegan Paul, London

Simberloff D (1983) Competition theory, hypothesis testing, and other community ecological buzzwords. Am Nat 122: 626–635

Trusted J (1979) The logic of scientific inference. Macmillan, London

Underwood AJ (1990) Experiments in ecology and management: their logics, functions and interpretations. Aust J Ecol 15:365–389

Underwood AJ (1991) The logic of ecological experiments: a case history from studies of the distribution of macro-algae on rocky intertidal shores. J Mar Biol Assoc UK 71: 841–866

Underwood AJ (1997a) Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge

Underwood AJ (1997b) Environmental decision-making and the precautionary principle: what does this mean in environmental sampling practice? Landsc Urban Plan 37: 137–146

Winer BJ, Brown DR, Michels KM (1991) Statistical principles in experimental design, 3rd edn. McGraw-Hill, New York

Editorial responsibility: Howard Browman (Contributing Editor), Storebø, Norway

Submitted: October 29, 1999; Accepted: November 15, 1999
Proofs received from author(s): December 27, 1999