105

Ecological Applications, 5(2), 1995, pp. 401–410

SCALABLE DECISION RULES FOR ENVIRONMENTAL IMPACT STUDIES: EFFECT SIZE, TYPE I, AND TYPE II ERRORS[1]

BRUCE D. MAPSTONE[2]
Research and Monitoring Section, Great Barrier Reef Marine Park Authority, P.O. Box 1379, Townsville, Queensland 4810 Australia

Abstract. Assessments of environmental impacts are being subject to greater scientific and legal scrutiny than ever before. The application of traditional statistical decision-making criteria to questions of environmental impacts has become increasingly inadequate as society demands greater environmental accountability from economic development. In particular, impact assessment has inherited a preoccupation with Type I error rates that has pervaded ecological research, even though Type II errors are often equally severe in impact assessment. Estimation of Type II error rates and specification of critical effect sizes—or the magnitudes of impacts considered important—are mutually dependent. Consideration of Type II errors, therefore, requires the exact specification of an hypothesized impact, which is often difficult. Insistence on low rates of Type I error (e.g., $\alpha = 0.05$) typically means that equivalent rates of Type II error can be realized only when effect sizes (ES) are very large or when very many samples are taken.

Rather than adhering to a fixed, arbitrary, critical, Type I error rate, I propose a procedure by which the critical ES is given primacy. Statistical decision criteria are then selected according to the relative weighting of the perceived consequences of Type I or Type II errors. The critical Type I error rate is set by iteration to some multiple (k) of the estimated potential for Type II error, and the null hypothesis is rejected if that (variable) Type I probability is not exceeded. The value of k would be determined by the ratio of the consequences (e.g., costs) of Type II and Type I errors. The procedure focuses attention on the magnitudes of impacts considered important, and provides for statistical decisions based on the a priori consideration of the development and environmental costs of Type I and Type II errors. It also provides incentive for development proponents to support rigorous environmental monitoring.

Key words: effect size; environmental impact assessment; environmental management; environmental monitoring; statistical decisions; statistical power; Type I error; Type II error.

INTRODUCTION

Large-scale habitat destruction, local species extinctions, and the possibility of global environmental change have emphasized that anthropogenic impacts can exceed the environment's capacity to absorb them (Soulé 1991). Further, human activities that change the status of the environment from that which is considered productive or undisturbed impinge on the economic viability of many industries, ranging from fishing and agriculture to eco-tourism. The potential for environmental degradation as a result of development, and the subsequent economic, social, and political ramifications, are now routinely juxtaposed with the mainly economic and political consequences of impeding development and economic growth.

Historically, environmental impact assessment (EIA) simply involved the prediction of impacts of development, and wrangling over approval to proceed with development often focused on the soundness of those predictions (e.g., Buckley 1989*a*, *b*, 1990, Fairweather 1989, Lincoln-Smith 1991). An increasing tendency to regulate development, however, carries the implicit expectation that such predictions will be tested, that impacts will be measured, and that regulations will be enforced. Permission to develop is now often contingent on funding an environmental impact monitoring (EIM) program to measure (potential) environmental impacts, with the implicit expectation that real impacts can be and will be detected. Decisions about environmental impacts are being made empirically as well as prophetically, and inevitably the statistical and inferential bases for those decisions will be subject to increasingly thorough legal and scientific scrutiny (Millard 1987, Hoverman 1989, Christie 1990, Beder 1991, Buckley and McDonald 1991, Fairweather 1991, Jarrett 1991, MacDonald 1991, Martin and Berman 1991).

Decisions about environmental impacts have important consequences, whether they are correct or in error (Bernstein and Zalinski 1983, Andrew and Mapstone 1987, Millard 1987, Peterman 1990, Fairweather 1991, Faith et al. 1991, Lincoln-Smith 1991). For example, concluding that an impact has occurred may result in the cessation of work, the closing of a factory, or the

[1] Manuscript received 12 December 1993; revised 1 July 1994; accepted 11 July 1994.

[2] Present address: CRC: Reef Research Centre, Sir George Fisher Research Building, James Cook University, Townsville, Queensland 4811 Australia.

imposition of fishing quotas. If that conclusion was wrong (Type I error, α), the curtailment of development would have been unwarranted and the local economic and social hardship likely would have been unnecessary. On the other hand, concluding that no deleterious impact has occurred usually provides tacit support for continued development. An incorrect conclusion of "no impact" (a Type II error, β) might mean that serious environmental degradation occurred before the real impact was noticed, possibly resulting in the collapse of a fishery, catastrophic pollution, or extinctions of species. Fairweather (1991) and Peterman (1990) have suggested that a Type II error would be more costly than a Type I error, since, in addition to a more severe environmental impact, the same hardships and economic costs would eventually be incurred.

It follows that *any* erroneous conclusion (either a Type I or a Type II error) would be cause for concern, yet traditional inferential statistical decision making has been preoccupied with only Type I errors. The acceptable (or critical) level of Type I error (α_c) has been dictated by convention, with the result that α_c is treated as a constant. In this paper I suggest specific steps for evaluating new critical levels of α, and procedures for making decisions based on the joint consideration of a variable α_c and the probability of Type II error (β).

TABLE 1. The four alternative outcomes in tests of hypotheses, illustrating the relationships between rejection of H_0 and Type I errors and non-rejection of H_0 and Type II errors (modified after Toft and Shea 1983, Peterman 1990). Also shown (in parentheses) are the functional analogues of these outcomes in decisions about hypothesized environmental impacts. The "?"s are used to emphasize that untruth and/or rejection of a H_0 is not always indicative of an impact, though it would be prerequisite to inferring an impact from hypothesis-testing procedures.

	Reality	
Decision	H_0 true (no impact)	H_0 false (Impact?)
Reject H_0 (Impact?)	Type I error (α)	No error–statistical power ($1 - \beta$)
Not reject H_0 (No impact detected)	No error ($1 - \alpha$)	Type II error (β)

COMPONENTS OF A DECISION

Under the hypothetico-deductive paradigm, empirical decisions typically rest on statistical tests of null hypotheses. This paradigm is adopted in many fields of science, and in EIM. The most common null hypothesis (H_0) is one of "no difference" or "no effect." Having collected our data, a statistical analysis provides us with a probability (α_0) of observing those data if the H_0 were in fact true. If this probability is less than some arbitrary critical level (α_c), almost always 0.05 in ecology, we reject H_0 in the belief that there is only a small probability (<0.05) of having done so incorrectly (a Type I error). If our associated probability is greater than the critical 0.05, we do not reject H_0. In EIM, non-rejection typically results in a conclusion of "no impact." Here, it is the probability of Type II error—the likelihood of the data under an alternative hypothesis (H_a)—that is important (Table 1).

Evaluation of β, however, cannot proceed without first stipulating (1) the critical Type I error rate (α_c) by which we failed to reject H_0, (2) the departure from H_0 represented by H_a that we would wish to detect if it were true, and (3) the sampling design and statistical model on which the test was based. The difference between H_0 and H_a has been termed the "effect size" (ES) (Winer 1971, Cohen 1988), and in EIM would be defined by the magnitude and form (Cohen 1988; ≈ "type" and "size," Bernstein and Zalinski 1983) of the *maximum* environmental impact that we would be prepared to tolerate in a particular case. This might roughly equate to the "limit of acceptable change" due to impacts of development. Sampling designs and statistical models for environmental impact studies have been discussed recently in considerable detail, and I will not discuss these issues here (see Green 1979, 1989, Bernstein and Zalinski 1983, Millard and Lettenmaier 1986, Stewart-Oaten et al. 1986, Millard 1987, Underwood and Peterson 1988, Warwick 1988, Faith et al. 1991, Underwood 1990, 1991*a*, *b*, Keough and Quinn 1991, Warwick and Clarke 1991; for further detail about the relations among sample size, variance, ES, α, and β see Andrew and Mapstone [1987], Bernstein and Zalinski [1983], Cohen [1988], Millard and Lettenmaier [1986], Peterman [1990], and Winer [1971]). My emphasis is on the criteria by which we make empirical decisions with the data from carefully designed EIM programs.

PROBLEMS WITH TRADITIONAL DECISIONS

Statistical decisions based only on Type I error rates are essentially decisions by a singular rule: Reject H_0 if $\alpha_0 < \alpha_c$. The singularity is emphasized by the tyranny of convention—everyone (in ecology) uses $\alpha_c = 0.05$! Being stipulated arbitrarily by us, the critical Type I error rate is not subject to other components of our research such as sample size and error variance. In rejecting H_0 by this rule, it is not incumbent upon us to worry about the magnitude of "statistically significant" differences (e.g., among means), even though that is perhaps the most interesting facet of our data: statistical significance has come to be treated almost synonymously with biological importance, even though no such relationship exists. Neither α nor β have any intrinsic meaning in terms of the biological variables we measure or the biological (or economic, political, etc.) importance of an outcome.

Toft and Shea (1983) and Peterman (1990) have pointed out that such an approach embraces an implicit, but unacknowledged, weighting of the importance attached to errors of Type I and Type II, a thesis supported by the weighting given to the two decision outcomes. In research, a "significant" result is cause for excitement, confidence, a manuscript, etc.; a nonsignificant result leads to despondency, a reevaluation of research direction, a conclusion of a non-result at best, or a failed experiment at worst. It follows that many authors are (perhaps unknowingly) content to consider Type I errors of far greater concern than Type II errors. This bias is often echoed in editorial judgements. In issues of environmental impacts, a "significant" result is cause for concern, action, litigation, controversy—no matter how small or inconsequential the impact; nonsignificance is sometimes controversial, often leads to complacency, but rarely precipitates action, concern, or litigation—even though large and important impacts might have been missed. Given the routine stipulation of small values of α_c, this constitutes a tacit prioritizing of development over environment and dictates a fundamental bias that permeates the entire design and decision-making process.

Further, with α_c fixed, negotiations over the costs of monitoring programs involve bartering, against increased costs of monitoring, the risks of failing to detect an impact (β) and/or the size of an impact that might be detectable with confidence. This equation provides little leverage in arguments with environmentally indifferent development proponents concerned to minimize costs of monitoring: their interests would be served best by a cheap and insensitive EIM (environmental impact monitoring) program. Here, notions of the importance of a detectable impact become entwined in a debate over costs and probabilities of error, with the result that the question of "how big an impact is tolerable?" often is translated to "how small an impact can the developer afford (economically) to worry about?"

The apparent immutability of the critical Type I error rate ($\alpha_c = 0.05$) might not present a problem if the Type II error rates typically realized in EIM were at least close to the α_c that drives decisions, or if consequential effect sizes (ESs) were detectable with reasonable certainty. In my experience, however, this is rarely the case—a view consistent with those of others who have addressed the neglect of statistical power in decision making (Toft and Shea 1983, Andrew and Mapstone 1987, Hayes 1987, Peterman 1990 and references therein, Fairweather 1991). In reviewing the unpublished reports of many EIM programs, I have found many in which the H_0 of no impact was not rejected, but very few where the likelihood of Type II error in doing so was <0.4 for an impact that would constitute an 80–100% change in the measured variables.

New Decision Rules

Resolving the above problems requires a substantial shift in approach to statistical decision making. An environmentally conservative approach might be to stipulate a low rate of Type II error, perhaps—though not necessarily—at the expense of elevated Type I error rates. Such an emphasis, however, leads to the consideration of α_c as a variable. Although other authors have suggested that the value of α_c should be carefully considered, implying that it should be considered as a variable at some stage (Bernstein and Zalinski 1983, Millard and Lettenmaier 1986, Oakes 1986, Andrew and Mapstone 1987, Millard 1987, Cohen 1988, Peterman 1990, Fairweather 1991), few have suggested an alternative strategy for decision making that incorporates a variable α_c (Cohen 1988, Peterman 1990) and few have adopted a non-standard α_c (e.g., Holt et al. 1987, Mapstone 1988, Mapstone et al. 1989). In most of these examples, one arbitrary value for α_c has simply been replaced with another.

If α_c is to be liberated, the single decision rule currently in use must be replaced with a sensible alternative. In so doing, it is important that we do not simply replace historic dogma ($\alpha_c = 0.05$) with contemporary chaos (anything goes) or the foundations of future dogma ($\beta = 0.05$). It is imperative also that any alternative decision rule(s) be specified and agreed a priori for any environmental impact monitoring (EIM) (or research) program—placing bets after the race has been run has always been illegal.

I suggest that changes to decision-making procedures must occur both before and after data collection. Two parallel, but initially independent, procedures should be followed prior to the collection of data. The first involves the choice of the level of impact (= critical effect size) that we want to detect (if it really occurs). The stipulation of an effect size (ES) is a biological (or chemical, physical, aesthetic, economic, etc.) decision, not simply a statistical or procedural decision, and involves a raft of judgements about the biological importance of an effect of a nominated magnitude or greater. As such, specification of ESs is perhaps the most critical aspect of environmental impact decisions, and possibly also of decisions in fundamental research, since this is where the importance of potential impacts (or alternative hypothesized outcomes) must be evaluated. Such evaluation should proceed in parallel with, but not subject to, negotiations over the costs of EIM and risks of erroneous decisions.

The second a priori requisite of the amended decision rules involves the derivation of critical levels of α and β. Evaluation of the relative consequences of Type I and Type II errors should play an instrumental part in determining the critical level of α that would lead to the rejection of a null hypothesis (see also Peterman 1990, Fairweather 1991). Setting a desired α_c in relation to the consequences of decision outcomes involves four steps:

1) Establish the relative importance of consequences (economic, political, environmental, social costs) of Type I and Type II errors. Designate the potential costs of Type I and Type II errors as C_I and C_{II}, respectively, and calculate $k = C_{II}/C_I$.

10372

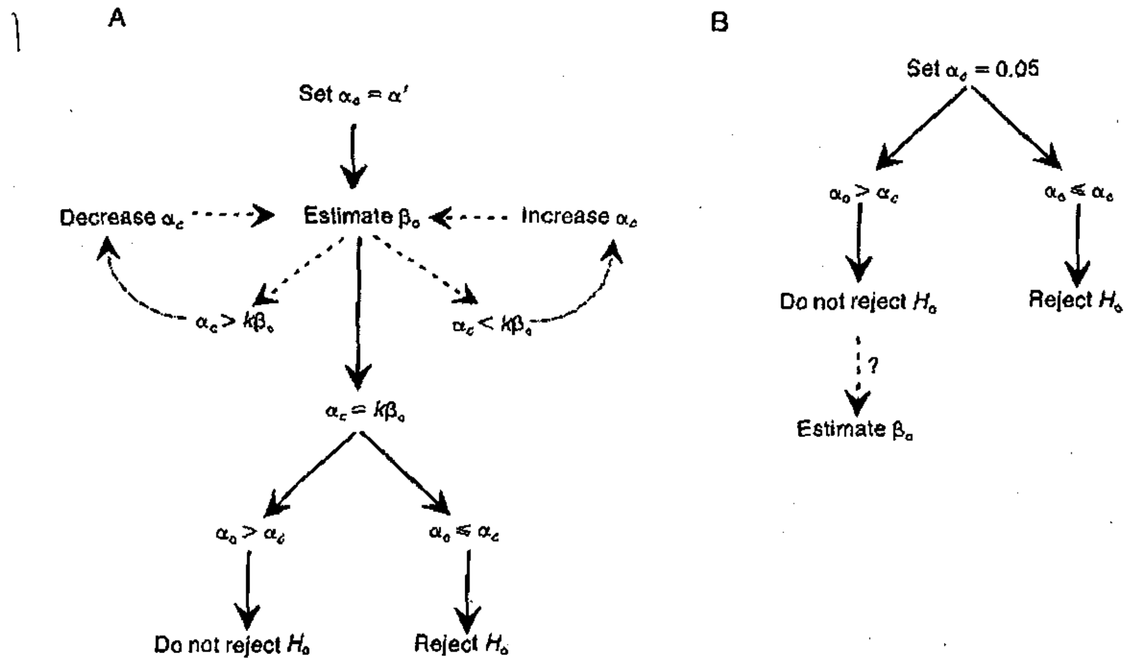

FIG. 1. Flow diagrams of the proposed (A) and conventional (B) decision procedures. Solid lines indicate the central components of a decision—i.e., choices or steps that must always be taken. Dashed lines in (A) indicate iterative steps that may be necessary in refining decision criteria. The "?" in (B) indicates that the estimation of β, the expected probability of a Type II error, is optional in conventional practice.

2) Set the *ratio* of critical Type I and Type II errors according to the relative costs of committing those errors—i.e., $\alpha_c/\beta = k = C_{II}/C_I$ (see also Kmietowicz and Pearman 1981, Peterman 1990), or more conveniently, $\alpha_c = k\beta$ (see later). This is an important and difficult task, and I suggest that in the absence of sufficient information to relate the costs of errors, C_I and C_{II} should be weighted equally—i.e., $k = C_{II}/C_I = 1$, $\alpha_c = \beta$—rather than the traditional strategy of tying down α_c and letting β roam (see also Peterman [1990] and references therein). Note that thus far the absolute level of neither potential error rate has been specified.

3) Specify (a) the maximum risk that development would be unnecessarily interrupted because of a Type I error with which the development proponent(s) would be prepared to work; and (b) the maximum risk that real and important impacts would go undetected because of a Type II error with which managers would be prepared to live. From these starting values, the values for risks of errors are negotiated with reference to the agreed relation $\alpha_c = k\beta$. These negotiations determine a priori the *desired values* of both α_c and β (denoted α_c' and β′) that all parties would be prepared to accept as criteria for the design of a monitoring program. For example, if it is decided that $k = 1$ and a risk of Type I error of more than 0.05 is unacceptable, then this procedure will dictate that the risk of Type II error also should be 0.05 or less.

4) Design an EIM program that would be likely to realize the desirable Type II error rate (β′), given the effect size specified previously (and independently), and the desired critical Type I error rate, α_c'. Sample size and design would be the dependent variables in these calculations (e.g., see Millard and Lettenmaier 1986).

After collecting the data from such a program, statistical decisions would be driven by the agreed value(s) of *k*, given the critical ES of concern. Two procedures might be used to make those decisions. The first amounts to a decision against a critical α; the second entails a decision against the relative realized values of α and β.

In the first decision procedure, the actual critical Type I error rate (α_c) would be set by iteration, as follows (Fig. 1):

1) Set α_c to that desired a priori—i.e., $\alpha_c = \alpha_c'$;
2) Calculate the Type II error rate expected (β_0) for the nominated critical ES, if the null hypothesis was not rejected at that α_c;
3) If $\alpha_c < k\beta_0$ or $\alpha_c > k\beta_0$, adjust α_c (higher or lower, respectively) to reduce the inequality, and then repeat step (2) with this new α_c;
4) Iterate steps (2) and (3) until $\alpha_c = k\beta_0$;
5) Compare the observed probability of the data if H_0 were true (α_0) with the final value of α_c and reject H_0 if $\alpha_0 \leq \alpha_c$.

An alternative decision procedure differs from the first in that the decision rests on direct comparison of α_o and $k\beta_0$, as follows:

10373

1) Calculate the probability of the data if H_0 were true (α_0);
2) Calculate β_0 on the assumption that the H_0 was not rejected against a critical Type I error rate marginally less than α_0—i.e., $\alpha_c = \alpha_0 - \delta$ (δ very small);
3) If $\alpha_0 < k\beta_0$, reject H_0, otherwise, do not reject H_0.

This decision is based on the argument that had α_c been set at $\alpha_0 + \delta$, the null hypothesis would have been rejected (since $\alpha_0 < \alpha_c = \alpha_0 + \delta$) and the probability of Type I error would be α_c or less. If $k\beta_0 > \alpha_c = \alpha_0 + \delta$, the probability of error in rejecting H_0 would be of less concern than the probability of error in nonrejection against a critical value of $\alpha_c = \alpha_0 - \delta$. If $k\beta_0 < \alpha_0$, then nonrejection would be the option with least concern about error. Ambiguity would arise if $k\beta_0 = \alpha_0$ exactly. This is unlikely, but in that event either decision would carry the same (k-weighted) concern about the probability of error.

In both of the above alternatives, both error rates are considered variable. Consequently the critical *probabilities* upon which decisions will be made will vary from case to case, even though the decision *rule* might be constant.

In both cases it is essential to derive an absolute value for β' because it is the desire to perpetuate that value to the final decision that provides the main criterion for the design of the EIM program. In the first procedure, the value of α_c used in the hypothesis test will also depend on how successfully the EIM program has been designed to realize $\beta_0 = \beta'$, because $\alpha_c = k\beta_0$. Hence, minimizing the risk of frequent intervention in development is contingent upon minimizing the risk of failing to recognize real (*important*) impacts. In the second procedure, although the level of α_c is not set by the level of β_0, the decisions about the likely presence of impacts will depend on the relative values of α_0 and β_0. Frequent intervention will be avoided, therefore, by that EIM design that maximizes the potential for $k\beta_0 < \alpha_0$ for moderate and small values of α_0—i.e., a design which minimizes β_0.

It is important to note, however, that the values of β_0 and α_c ultimately realized are likely to differ from those considered desirable (β', α_c') because of the likelihood of estimation errors in the design of the EIM. Although we may design an EIM program to realize a specified $\beta_0 = \beta'$, such a design will be based on estimates of variance, etc., that will be unlikely to be realized exactly in the final data set (e.g., see McArdle et al. 1990). The consequences of not realizing the a priori desired probability of errors, however, are shared by both development proponents and managers. Had the traditional strategy of decision against a fixed α_c been followed, the consequences of inexact design would have reflected exclusively on the value of β_0, and been borne by the environment (Fig. 2).

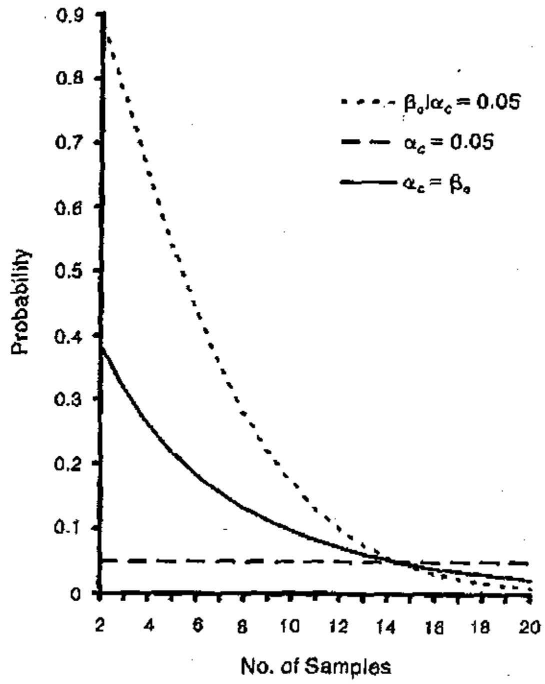

FIG. 2. Illustrative example of the effects of increasing sample size on the values of α_c, the critical (or acceptable) Type I error rate, and of β_0, the expected probability of Type II error, under the conventional and proposed statistical decision rules. Under conventional practice, the value of α_c (– – – –) does not change with increasing sample size, and all benefits of increased sample size reflect on the value of β_0 (- - - -). With the proposed rules, small samples mean equally high values of both α_c and β_0 (———), and increasing sample sizes reduces both. Effect size is constant throughout, and k = 1 (see *New decision rules*).

NEW DECISIONS IN PRACTICE: AN EXAMPLE

A (hypothetical) development was planned for a coastal region in southern Australia and effluent was to be discharged into the sea offshore. There was concern that effluent washing inshore would impact on a single rocky intertidal platform, which was a popular recreational area adjacent to a terrestrial national park. The major macro-organism-forming structure on the shore was the brown alga *Hormosira banksii*, and toxicological research had shown that this species was likely to die or suffer reduced reproductive success following exposure to high concentrations of the effluent. Thus, *H. banksii* was considered a sensitive indicator of local effects of the effluent.

Research had shown also that *H. banksii* readily recovered from depletion down to 30% cover, but greater disturbances had long-lasting effects, sometimes resulting in local extinctions for several years (Povey and Keough 1991). Prior data and a pilot study indicated that the average existing coverage of *H. banksii* was 84.75% (average of three locations), and had remained relatively stable for the previous 3 yr. Hence, the critical effect size considered important here was set at there being a 30% greater reduction in standing crop at the impact location than at the control locations. This reduction would take the cover at the location of potential impact from its existing cover of about 75% to about 45%, 1.5 times that which would cause lasting damage. It was considered that this level of disturbance, if detected, would be sufficient early warning from which to instigate reactive management.

Setting desired values for α and β

The ratio of the critical (acceptable) Type I error rate to the Type II error rate, α_c/β, considered desirable in future decisions about impacts of the development was set at 0.2 (i.e., $k = 0.2$). This ratio was a negotiated agreement between management agencies, local lobby groups, and the development company and was more conservative with respect to Type I error than Type II error because:

1) It had been demonstrated that existing technology produced effluent of low and benign concentrations, so real impacts were not expected;

2) The effluent constituents were not considered a health risk for humans;

3) If impacts of effluent discharge were inferred for the coast then effluent would have to be processed further than planned, at considerable extra cost;

4) Suspension of the development whilst additional effluent-processing mechanisms were installed would mean temporary stand-down or reduction in shifts for employees, resulting in financial loss to them, and lost production from the plant;

5) The rocky platform in question was important because it was a seaward extension of a terrestrial national park, but it was not unique in the region; and

6) There was concern that small business in the vicinity of the national park would suffer if people were deterred from visiting the coast as a result of perceived pollution problems, meaning that it was considered desirable to have at least a reasonable chance of detecting impacts before they became serious.

It was also agreed that the risk of Type I error should be kept low, because of the costs of erroneous intervention in what seemed likely to be an environmentally benign development. The desired level of α_c was set to 0.02, meaning that the desired level of β was 0.1. Hence, the environmental impact monitoring (EIM) program for *H. banksii* was designed to realize statistical power ($\rho = 1 - \beta$) of $\geq$0.9 to detect a decline (or change) in average cover of *H. banksii* at the (potential) impact location of 30% less than that which occurred on average at the control locations. Such an impact, if it occurred, would be expected to occur within any or all of the first 3 yr post-start-up, given the expected frequency with which diluted effluent might reach the impact location ($\leq$12 times/yr).

Estimation of required number of control locations

In this case, three years of data were available for *H. banksii* from three nearby rock platforms, providing estimates of the interaction between locations and years in the absence of effluent impacts. From these data, the behavior of *H. banksii* was modelled at l hypothetical control locations and one impact location over 3 yr prior to effluent discharge, and during 3 yr of effluent exposure to derive estimates of the appropriate error variances for tests of the null hypothesis (no effect of effluent on *H. banksii*) in each of the 1st, 2nd, and 3rd yr post-start-up. The required number of locations, l, to be sampled to allow decisions against $\alpha_c = 0.02$ such that $\beta_0 \leq 5 \cdot \alpha_c$ when *H. banksii* cover dropped by 30% more at the impact location than at the average of the controls in any one of the three years was derived by iteratively:

1) choosing a value for l (the number of control locations sampled);
2) deriving the critical value of F at $\alpha_c = 0.02$ (when the null hypothesis was true) with the df appropriate to sampling l control locations;
3) calculating the probability (β_0) of observing that value of F or less when an impact of −30% at the impact location existed;

until $\beta \leq 0.1$. The results are presented in Table 2.

TABLE 2. Expected potential for Type II error (β) given an impact of effluent in the 1st yr post-start-up that caused reduction in cover of *Hormosira banksii* by 30% more at the impact location than at control locations, when l locations were sampled and $\alpha_c = 0.02$. Estimates derived from 3 yr of pilot and 3 yr of prior data, and from revision of the environmental impact monitoring (EIM) program during the 1st and 2nd yr of baseline sampling, are given. The last line of the table shows the values of α_c* and β_0† expected for hypothesis tests after 1 yr of operation if $\beta_0 = 5 \cdot \alpha_c$ and sufficient locations were sampled to ensure $\alpha_c \leq 0.02$ and $\beta_0 \leq 0.1$.

	Pilot data		Baseline sampling data			
			Year 1		Year 2	
l	α_c	β_0	α_c	β_0	α_c	β_0
3	0.020	0.599	0.020	0.816	0.020	0.716
4	0.020	0.239	0.020	0.630	0.020	0.427
5	0.020	0.061	0.020	0.436	0.020	0.200
6			0.020	0.278	0.020	0.079
7			0.020	0.167		
8			0.020	0.095		
	0.017	0.083	0.0195	0.0977	0.018	0.091

* α_c = the critical (or acceptable) Type I error rate.
† β_0 = the expected probability of Type II error.

Review of design during baseline period

In each of the 1st and 2nd yr of the baseline study, the design of the EIM program was reviewed. The reviews entailed remodelling the dynamics of *H. banksii*, based on the additional data available from the actual control and impact locations being sampled (Table 2).

Serious consideration was given to expanding the EIM in this case, since after the first year of baseline monitoring it was estimated that eight control locations should be sampled to realize the desired rates for Type

I or Type II error. Continuing with the original design of only five control locations would mean that the potential error rates would increase by about 2.5 times, and the power of the program to detect the nominated critical level of impact would drop from ≈90% to ≈75%. The decision whether to increase the number of control locations was driven by comparing the certain costs of increasing monitoring (by 67%/yr) with the increased potential for costs resulting from erroneous decisions about impacts (increase in $\alpha_c \times$ cost of an alleged impact). In this case it was decided to increase the scope of the EMP to keep low the risks of erroneous intervention in development, and maintain confidence of detecting critical impacts.

The final decision

Analyses of the data during the first 3 yr post-startup indicated that the desired rates of Type I and Type II error were realized as a result of sampling at the increased number of control locations. The results showed that the estimates from the 1st yr review of the design were about right. Sampling eight control locations produced hypothesis tests with $\alpha_c = 0.0206$, 0.0226, and 0.0218, and $\beta_0 = 0.1028$, 0.1132, 0.1091 in each of the three years, respectively. In this case, however, no impact was inferred ($\alpha_0 > 0.15$ in all years).

Advantages of Liberating Alpha

The liberation of α_c and adoption of a scalable decision rule has several advantages. Firstly, trading a variable α_c' against a variable β' means that attention is clearly focused on the consequences of both decision outcomes. This forces the explicit a priori evaluation of the relative importance of Type I errors and Type II errors, and hence direct comparison of the risks the interested parties are prepared to take in proceeding with a development (given a nominated level of environmental impact monitoring, EIM). Here, better monitoring means a lower probability of Type II error *and* a lower probability of Type I error. Both are essentially arbitrary limits of confidence in an outcome with which we are prepared to live, and in this process they are evaluated as such.

Secondly, cost savings in monitoring will be traded against costs of frequent intervention in development that might arise from a high α_c, tied to a high β_0 by the relation $\alpha_c = k\beta_0$. Lower risk of erroneous interference is achieved only by realizing a lower value for β_0, given a nominated threshold of impact—i.e., by better monitoring (Fig. 2). Further, given that it is likely to be more difficult to prove an error of Type I (for which a developer might seek compensation) than an error of Type II (for which a developer might have to pay reparation), it is clearly in the developer's best interest to promote rigorous monitoring.

Thirdly, the stipulation of effect size (ES), the crucial environmental variable, is removed from the bargaining table when the scope and costs of EIM are being negotiated. The limits of acceptable impact should be stipulated independently by those with relevant expertise, not by those responsible for the economic success of a proposed development. At no point should the critical ES in a chosen variable be adjusted because of the cost of designing an EIM program that would be likely to detect that effect, although such logistic and financial considerations might influence the choice of variables to monitor. If it is impossible to stipulate what level of perturbation would be considered important for particular variables, then monitoring those variables should be seen as information gathering, exploration, provision of data for the future—not as vehicles for testing hypotheses about impacts. The above procedures precipitate these issues a priori, whereas conventional procedures may never stimulate consideration of them.

Fourthly, the suggested procedures allow different decision criteria (α_c, β) to be attached to different scales of impact and/or different variables measured in EIM. Where EIM incorporates assessment of impacts at a hierarchy of spatial or temporal scales (Mapstone et al. 1989, Underwood 1991*b*), the consequences of impacts at each scale can be evaluated independently, and decision criteria set accordingly. For example, it might be considered unimportant that an impact of specified magnitude was missed on a very small scale (such as a single small site, or for a period of only 1 wk), and the maximum acceptable level of β might be set at a relatively large value (e.g., 0.3). An impact of similar, or different, magnitude (the ES might also be scale dependent) at a larger scale, such as over an entire bay or effective for 1 mo, might be considered far more important and, accordingly, the maximum acceptable level of β might be set to a very low level. Thus, the significance criteria by which H_0s were rejected might vary among the terms within a single multi-factorial ANOVA. Similarly, it might be considered crucial that a real impact for one variable (e.g., levels of organochlorines) was detected, but of only marginal concern that an impact on another variable (e.g., water clarity) was missed. Different decision criteria would be applied to different variables, and their sampling intensities determined accordingly.

Fifthly, the above procedures explicitly identify a set of priorities in assessing environmental impacts. The independent establishment of a critical ES—the maximum "acceptable" impact—takes highest priority. Developments will always have impacts on the environment, but many will be trivial. The limits to impacts, which might vary among cases, should be determined in advance by reference to the local environment. Establishing the relative importance of Type I and Type II errors, and the levels of certainty ($1 - \alpha$ or $1 - \beta$) desired in decisions, are also high priorities and should be assessed relative to stipulated critical ESs.

Finally, the costs of monitoring are evaluated against the risks involved in assessments of impact, again, without impinging on the size of impact that is considered ecologically (or by some other cost-independent criterion) important. I suggest that these priorities provide a fair basis for balancing the burden of proof (Peterman 1990) between developer and environmental manager.

Some Problems with a Variable α

The determination of critical effect size

Effect size (ES) is defined by two components: the form(s) and magnitude(s) (Cohen 1988) of the impact(s) we seek to detect if they occur. The form of impact involves deciding whether we are concerned with changes in means and/or variances at impact sites relative to control sites (Green 1979, 1989, Stewart-Oaten et al. 1986, Underwood 1991*a, b*), deciding at what scales we expect impacts might occur, and specifying which means (or groups of means) differ from which others. The magnitude of an impact is a measure of the amount by which means or variances change. Cohen (1988) discusses the need to specify exactly both components of an effect to sensibly construct alternative hypotheses. Specifying both aspects of effect size (ES) will be difficult for complex environmental impact monitoring (EIM) designs, particularly when impacts of interest are measured by interaction terms in analyses.

For each variable being monitored, we must answer the questions "how much anthropogenic disturbance is acceptable?," and "what amount of development-related change should precipitate management action?" The two answers will not always be similar, since with some variables (e.g., concentrations of persistent or toxic pollutants) we might wish to initiate management action at levels of change well below those that would be considered the limits for environmental or human well-being. For such variables we might wish EIM to provide the basis for proactive rather than reactive management. Even when limits of acceptable change are specified by regulations (e.g., U.S. Environmental Protection Agency regulations), we will still be faced with deciding at what point we would wish to know if that level were being approached, in the interests of ensuring that it was never reached. The above procedures highlight the shortcomings of our current understanding of environmental impacts, how we measure them, their importance, and how we manage them. Clearly, more research in this field is required.

Two further points should be noted here. Firstly, although I have routinely referred to ES in terms of "ecological" importance, other criteria for critical ESs might be important. For example, aesthetic or economic consequences of environmental impacts might result in the stipulation of more stringent critical ESs than would arise from consideration of ecological consequences alone. Two examples illustrate this point: (1) In areas where eco-tourism is important (e.g., the Great Barrier Reef), impacts that might be considered relatively minor ecologically might represent important degradations in aesthetic quality and reduced tourism; and (2) Where environmental impacts impinge on a fishery, either by pollution or stock reduction, the magnitude of impacts that precipitated economic losses to the fishing industry might be smaller and considered more critical than those which would cause substantive ecological effects. In both examples, it would be the economic costs to third parties that would most influence the choice of a critical ES.

Secondly, it must be emphasized that the critical ES is not a categorical distinction between "impact" and "no impact," but simply provides a cut-off point on a continuous relation between ES and β (or statistical power). A monitoring program will have lower power to detect smaller than "critical" impacts that have slighter consequences. The critical ES merely specifies where the (increasing) consequences of impact become unacceptable.

Weighting α_c and β

The weighting of Type I and Type II errors is unfamiliar to most scientists, managers, and development proponents, although the need to do so has been discussed for some time (see Kmietowicz and Pearman 1981, Oakes 1986, Peterman 1990). If α_c becomes large as a consequence of liberating α_c and demanding low values for β, there is likely to be increased intervention in development, and more frequent litigation of decisions from EIM programs. Avoiding such action will entail greater costs of EIM, but those costs will become increasingly critical (for the proponent) as the economic scale of development decreases. Further difficulty arises when the costs of Type I and Type II errors are measured in different currencies—e.g., money vs. genetic diversity. It is likely that economic rationalist attempts to value all costs in monetary terms will favor short-term development interests, especially where entire communities are dependent on a nominated development. Such issues make the evaluation of α_c'/β' particularly critical. Again, these issues can be comfortably ignored within traditional decision-making practice, but are critical elements of the procedures I suggest above.

Administrative reality

Pragmatists will say that the above procedures involve considerably more work than existing practice and that it will be difficult to agree upon and sustain decision criteria throughout the life of an EIM exercise. The vagaries of political priorities, management changes, and new information are likely to mean that what was considered appropriate initially will be repeatedly challenged, debated, and possibly changed—particularly where developments and EIM span several years. The above procedures precipitate such debate, but the issues are not new. In current practice, issues such as the importance of an observed impact, the adequacy of monitoring, or the basis of a decision are fought after the event, when revision, adjustment, or improvement is impossible. The likelihood that these debates would be brought forward is an advantage of the above procedures, but will require diligent legal management.

Conclusion

Conventional decision-making practices in respect of environmental impacts perpetuate an inherently one-sided perspective of significance. Masked by the security of a well-established convention of statistical decision making is a suite of difficult inferential and epistemological problems that have real, tangible implications. The revised approach to statistical decision making I suggest here does not solve these problems, but focuses attention on them and highlights the need for urgent attention to them. More importantly, the procedures I suggest provide a mechanism by which the burden of limitations of current practice is shared between potential environmental assailant and environmental defender rather than being borne solely by the latter.

Acknowledgments

The ideas expressed here benefitted from discussions with Neil Andrew, Howard Choat, Geoff Jones, Sally McNeill, Helene Marsh, Jamie Oliver, Kath Shurcliff, and Tony Underwood. I am grateful to Neil Andrew, Sally Hollbrook, Geoff Jones, Craig Osenberg, Russ Schmitt, and two anonymous reviewers for comments on earlier manuscripts. Mick Keough, A. Povey, and Gerry Quin kindly provided some real data from which I constructed the cited example of "new decision rules in practice." The main theme of the paper was presented at the 2nd International Temperate Reef Symposium (Auckland 1992). I acknowledge the support of the Australian Research Council (for a National Research Fellowship), the Great Barrier Reef Marine Park Authority, and the CSIRO (Pulp and Paper Research Project).

Literature Cited

Andrew, N. L., and B. D. Mapstone. 1987. Sampling and the description of spatial pattern in marine ecology. Oceanography and Marine Biology Annual Reviews 25:39–89.

Beder, S. 1991. The many meanings of means. Search 22: 88–90.

Bernstein, B. B., and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. Journal of Environmental Management 16:35–43.

Buckley, R. 1989*a*. Precision in environmental impact prediction: first national environmental audit, Australia. Resource and Environment Study Number 2. Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia.

———. 1989*b*. What's wrong with EIA? Search 20:146–147.

———. 1990. Environmental science and environmental management. Search 21:14–16.

Buckley, R., and J. McDonald. 1991. Science and law—the nature of evidence. Search 22:94–95.

Christie, E. 1990. Science, law and environmental litigation. Search 21:258–260.

Cohen, J. 1988. Statistical power analysis for the behavioural sciences. Second edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Fairweather, P. G. 1989. Environmental impact assessment—where is the science in EIA? Search 20:141–144.

———. 1991. Statistical power and design requirements for environmental monitoring. Australian Journal of Marine and Freshwater Research 42:555–567.

Faith, D. P., C. L. Humphrey, and P. L. Dostine. 1991. Statistical power and BACI designs in biological monitoring: comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate communities in Rockhole Mine Creek, Northern Territory, Australia. Australian Journal of Marine and Freshwater Research 42:589–602.

Green, R. H. 1979. Sampling design and statistical methods for environmental scientists. John Wiley & Sons, New York, New York, USA.

———. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research 50: 195–205.

Hayes, J. P. 1987. The positive approach to negative results in toxicology studies. Ecotoxicology and Environmental Safety 14:73–77.

Holt, R. S., T. Gerrodette, and J. B. Cologne. 1987. Research vessel survey design for monitoring dolphin abundance in the eastern tropical Pacific. U.S. Fish and Wildlife Service Fisheries Bulletin of the US, 85:435–446.

Hoverman, J. S. 1989. Scientists must punt or be damned. Search 20:147–148.

Jarrett, R. G. 1991. Means to an end: a statistical adjudication. Search 22:92–93.

Keough, M. J., and G. P. Quinn. 1991. Causality and the choice of measurements for detecting human impacts on marine environments. Australian Journal of Marine and Freshwater Research 42:539–554.

Kmietowicz, Z. W., and A. D. Pearman. 1981. Decision theory and incomplete knowledge. Gower, Hampshire, England.

Lincoln-Smith, M. P. 1991. Environmental impact assessment: the roles of predicting and monitoring the extent of impacts. Australian Journal of Marine and Freshwater Research 42:603–614.

MacDonald, R. 1991. Statistics in the real world. Search 22: 91–92.

Mapstone, B. D. 1988. The determination of patterns in the abundance of *Pomacentrus moluccensis* Bleeker on the southern Great Barrier Reef. Dissertation. University of Sydney, Sydney, New South Wales, Australia.

Mapstone, B. D., J. H. Choat, R. L. Cumming, and W. G. Oxley. 1989. The fringing reefs of Magnetic Island: benthic biota and sedimentation. Research Publication Number 13. Great Barrier Reef Marine Park Authority, Townsville, Queensland, Australia.

Martin, R., and P. A. Berman. 1991. Environmental research in Australia. Search 22:95–96.

McArdle, B. H., K. J. Gaston, and J. H. Lawton. 1990. Variation in the size of animal populations: patterns, problems, and artefacts. Journal of Animal Ecology 59:439–454.

Millard, S. P. 1987. Environmental monitoring, statistics, and the law: room for improvement. American Statistician 41: 249–253.

Millard, S. P., and D. P. Lettenmaier. 1986. Optimal design of biological sampling programs using the analysis of variance. Estuarine and Coastal Shelf Science 22:637–656.

Oakes, M. 1986. Statistical inference: a commentary for the social and behavioural sciences. John Wiley & Sons, Chichester, England.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47:2–15.

10378

Povey, A., and M. J. Keough. 1991. Effects of trampling on plant and animal populations on rocky shores. Oikos **61**: 355–368.

Soulé, M. E. 1991. Conservation: tactics for a constant crisis. Science **253**:744–750.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology **67**:929–940.

Toft, C. A., and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist **122**:618–625.

Underwood, A. J. 1990. Experiments in ecology and management: their logics, functions and interpretations. Australian Journal of Ecology **15**:365–389.

———. 1991*a*. Biological monitoring for human impact: how little it can achieve. Pages 105–223 *in* R. V. Hyne, editor. Proceedings of the 29th Congress of the Australian Society of Limnology, Jabiru, Northern Territory, 1990. Office of the Supervising Scientist Alligator Rivers Region, Australian Government Publishing Service, Canberra, New South Wales, Australia.

———. 1991*b*. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian Journal of Marine and Freshwater Research **42**:569–587.

Underwood, A. J., and C. H. Peterson. 1988. Towards an ecological framework for investigating pollution. Marine Ecology Progress Series **46**:227–234.

Warwick, R. M. 1988. The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. Marine Pollution Bulletin **19**:259–268.

Warwick, R. M., and K. R. Clarke. 1991. A comparison of some methods for analysing changes in benthic community structure. Journal of the Marine Biological Association of the United Kingdom **71**:225–244.

Winer, B. J. 1971. Statistical principles in experimental design. McGraw-Hill, New York, New York, USA.