Vol. 279: 63–72, 2004 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published September 28

Assessing impacts of dredge spoil disposal using equivalence tests: implications of a precautionary (proof of safety) approach

Russell Cole[1,*], Graham McBride[2]

[1]National Institute of Water and Atmospheric Research, PO Box 893, Nelson, New Zealand
[2]National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand

ABSTRACT: Equivalence tests evaluate whether a treatment effect lies within or outside a predetermined equivalence interval. The equivalence interval might be determined in relation to a precleanup value, reference values, or unimpacted controls. Such tests are little known in ecology, but offer advantages over the common tests of point-null hypotheses. They come in 2 forms. The first, testing the equivalence hypothesis, constitutes a proof of hazard approach by postulating that a difference lies within the interval. The second, testing the inequivalence hypothesis, also known as bioequivalence testing, constitutes a proof of safety approach by postulating that a difference lies beyond the interval. The proof of safety option provides a formal mechanism for the implementation of the precautionary approach. We demonstrate the usage of these tests for a subtidal rocky reef dataset evaluating the impact of dredge spoil disposal at New Plymouth, New Zealand. Sampling of conspicuous subtidal organisms was done once before, and twice after spoil disposal. Abundances of subtidal organisms were compared between 6 sites in the predicted path of the dredge spoil and 6 distant control sites. Comparison of mean numbers of species and of individuals on and off the dumpground before and after disposal demonstrated that greater effort was required for proof of safety, though there was no proof of hazard. The most important conclusion was that much greater sampling effort than is common in ecological study was required to demonstrate safety. Because equivalence tests are readily calculated, test a realistic hypothesis, and provide an outcome that is directly interpretable in terms of a biological endpoint, we suggest that they should be more widely adopted.

KEY WORDS: Dredge spoil disposal · Equivalence tests · Proof of hazard · Proof of safety · Tests of null hypotheses

INTRODUCTION

Formal evaluation of evidence regarding biological impacts of human activities is an important role for marine ecologists (e.g. Underwood & Peterson 1988, Underwood 1997). The usual statistical tool used to evaluate such evidence in ecology is testing of point-null hypotheses in the frequentist paradigm (e.g. Underwood 1990, 1997, Fairweather 1991). Under such an approach, evidence is gathered to test the null hypothesis that there is no difference whatsoever between the population parameters considered (e.g. means). This hypothesis is rejected if the p-value for its test is less than the permissible Type I error risk (the probability of rejecting a true hypothesis, usually taken as $\alpha = 0.05$), and it is concluded that the treatments differ. In this way p-values for point-null tests are treated as evidence for or against research hypotheses. Even exhortations to publish statistically non-significant results appear to accept that the p-value from such a test is an appropriate criterion for judging the strength of the evidence gathered (e.g. Lortie & Dyer 1999). However, the interpretation of these p-values as evidence invokes serious inferential problems (see Appendix 1), not the least being that they tend to become ever smaller as the number of

*Email: r.cole@niwa.co.nz

samples is increased. This is because the p-value is calculated assuming the tested hypothesis to be true, but in fact it isn't — there will be some difference present, however small. This has tended to be discussed in the statistical (e.g. Gibbons & Pratt 1975, Berger & Sellke 1987, Sellke et al. 2001, McBride 2002), medical (Goodman & Royall 1988, Goodman 1999, Poole 2001) and sociological literatures (Morrison & Henkel 1970, Harlow et al. 1997), but it is increasingly discussed in some sectors of the ecological community (e.g. Shrader-Frechette & McCoy 1992, Mapstone 1995, Germano 1999, Johnson 1999, Quinn & Keough 2002). In particular, such p-values can serve well when comparing the strength of evidence from tests using the same number of data, but they are not comparable when the number of samples differs.

Comparisons of p-values cause difficulties for ecologists in several situations. As an example, rules have been offered for interpreting p-values in tables that arise from complex analysis of variance designs (e.g. Underwood 1991, 1992, 1993, 1997). In many situations, presentation of such tables is taken to summarise the data. However, as above, if different numbers of samples are used between factors or across studies, the relative sizes of p-values will not adequately summarise even the comparative evidence. In Cohen (1988), example 8.8 (p. 374–375) demonstrates this for 3 fixed factors with varying numbers of levels using 5% level tests. For a true effect size of 0.25, the 3 main effect tests have powers that range from 0.58 to 0.70. This varying power means that p-values will tend to differ between comparisons, merely as a result of the different numbers of samples between factors. The use of relative size of p-values as evidence is thus confounded by varying sample sizes. As a further example, Osenberg et al. (1999; their Fig. 1) report a meta-analysis of experiments investigating stonefly predator effects on densities of various taxa in streams. They show that the effect sizes of some of the statistically significant tests (i.e. those with $p \leq \alpha$) are relatively small compared to other statistically non-significant comparisons, demonstrating that the failure to reject the null hypothesis was a result of small power, rather than a small effect.

Such issues have given rise to discussions regarding negative results (Browman 1999, et seq.) and publication bias (Palmer 1999), indicating a need to have access to at least some studies in which a point-null hypothesis has not been rejected (i.e. α-censoring, where studies that fail to attain $p < 0.05$ tend not to be published). This can be manifest in calls to emphasize estimation (especially using confidence intervals) rather than testing (Gardner & Altman 1989, Hoenig & Heisey 2001). Indeed, Peterson et al. (2001) noted a move in ecological studies away from testing hypotheses toward estimation of the magnitude of effects. This approach permits some freedom for interpretation, and is always useful. However, it fails to provide a procedure for detecting ecologically important effects, as is often required, for example, in studies of environmental impact or for processing resource-use applications. Having expended considerable effort collecting and analysing the data, failing to reach a conclusion on the basis of those data is unsatisfactory. For those reasons, and because confidence interval widths depend (inter alia) upon the number of samples, we see the call for greater emphasis on estimation as desirable but incomplete.

Another approach to the problems caused by α-censoring is to adjust α (Cascio & Zedeck 1983, Mapstone 1995). In Mapstone's procedure, a sampling programme is designed to regularly detect a minimum effect size (if such an effect actually exists) judged by experts to be of biological consequence. In analyzing a set of data, α and the Type II error risk β (of failing to reject a false hypothesis) are adjusted according to the ratio of costs associated with committing either error (the originators of hypothesis testing procedures Neyman & Pearson 1933, noted that the choice of error criteria has to do with the consequences of each error occurring). Inevitably, given the small numbers of samples used in marine ecological studies, this procedure will result in values of α larger than the usual 0.05, and accordingly this procedure will detect more effects than the traditional approach, especially because it forces the analyst to confront issues of minimum detectable effect size and permissible error risks before performing the test. However, it still tests an untenable (point-null) hypothesis — in which case, why test it at all?

In this paper we examine a different hypothesis testing approach that abandons the testing of point-null hypotheses and instead tests hypotheses stated in terms of equivalence intervals (Wellek 2002). The basic idea of such equivalence tests is that if a test suggests that the true difference is inside that interval, then a declaration of equivalence may be made. That is, it is recognised that there is a difference but that it can be small enough for the population parameters to be considered equivalent. These tests supply a formal mechanism by which the evidence for the tested hypothesis may actually be strengthened by the collection of more data (for a point-null hypothesis the overall outcome of increasing sample size is an inevitable weakening of the evidence for that hypothesis). We note that some working statisticians have the same view, e.g. '… we think equivalence testing approaches are underutilized. We often see examples where statisticians and non-statisticians are testing the wrong hypotheses, apparently stuck in a mode of thinking based on null hypotheses of no-difference' (Anderson & Hauck 1996).

Limiting the discussion to the 2-mean comparison, equivalence tests can examine either (1) the equivalence hypothesis, in which the true difference between means is postulated to lie within a prescribed equivalence interval, or (2) the inequivalence hypothesis, in which the true difference in means is postulated to lie beyond that interval. These tests provide a formal framework for demonstrating proof of hazard (1) or proof of safety (2). Therefore, tests of the latter hypothesis will be of particular interest to ecologists, since they provide a formal vehicle to implement the precautionary approach that traditional approaches have great difficulty in accommodating. For example, Dayton (1998) has advocated reversing the usual burden of proof in fisheries management—testing the inequivalence hypothesis enables this to be done in a straightforward manner; tests of point hypotheses do not (see especially the analysis of the power approach by Schuirmann 1987). Tests of the inequivalence hypothesis are now in routine use in applied fields where incorrect conclusions are matters of human life and death. In medical investigations (which are necessarily precautionary) of the efficacy of drug formulations, tests of point-null hypotheses have fallen from favour and equivalence procedures are mandated (http://www.fda.gov/cder/guidance/1716dft.htm). Despite their considerable advantages and their mandated use in drug tests, equivalence tests are seldom used in the ecological literature (see McDonald & Erickson 1993, Garrett 1997, McBride 1999, Cole et al. 2001 and MacKenzie & Kendall 2002 for non-medical examples). Here we demonstrate the use of equivalence tests in a dredge spoil disposal example. Our intention is to provide examples of the usage of tests of interval hypotheses, to underscore their advantages over tests of point-null hypotheses, and to demonstrate important consequences of their use.

MATERIALS AND METHODS

Study site. New Plymouth is situated in the province of Taranaki, on the west coast of the North Island of New Zealand (Fig. 1). The physical characteristics of this wave-swept coast are described in McComb et al. (1997) and McComb & Black (2001). Cole et al. (1999) describe a depauperate subtidal fauna of mobile invertebrates mainly comprising echinoderms and molluscs. Despite the dominance of coralline-covered areas, abundances of grazing invertebrates such as echinoids and herbivorous gastropods are much lower than those in northeastern New Zealand (summarised in Andrew 1988, Creese 1988).

Sampling. A sampling programme was instigated to assess impacts in the shallow subtidal of a trial subtidal disposal of 47 000 m^3 of sand (Fig. 1). Disposal occurred in late February and early March 1999 and details of the dispersal of the sediment can be found in McComb & Black (2001). Ecological sampling was done on 3 occasions, before (December 1998), and twice after (May 1999 and November 1999) spoil disposal. During May 1999 there was no underwater visibility at several sites, and no subtidal data could be obtained there; that incomplete subtidal survey is therefore omitted. A preliminary presentation of results is given in Cole et al. (1999).

Sampling was done at 12 sites on reefs, of which 6 lay close to and in the path of sediment from the disposal ground, and were designated impact sites, and 6 lay further away, and served as controls (Fig. 1). At each site, 5 randomly placed 1 m^2 quadrats were sampled for seaweeds and large mobile benthic invertebrates. Number of seaweed stipes (*Carpophyllum maschalocarpum* and *Ecklonia radiata*), and numbers of echinoids, starfish, and gastropods were counted in each quadrat. We analysed number of species and number of individuals, as abundances of individual subtidal species were low and/or highly variable.

Due to concerns that the sample sizes required to demonstrate equivalence or inequivalence may be prohibitively high, we undertook further sampling at New Plymouth. In December 2003, R. Cole recorded the number of mobile invertebrate species in 60 random 1 m^2 samples at each of 3 intertidal sites. Sampling of the 3 sites took a total of about 6 person hours.

Analysis. The nomenclature is introduced in Table 1. We present results for 3 types of hypothesis

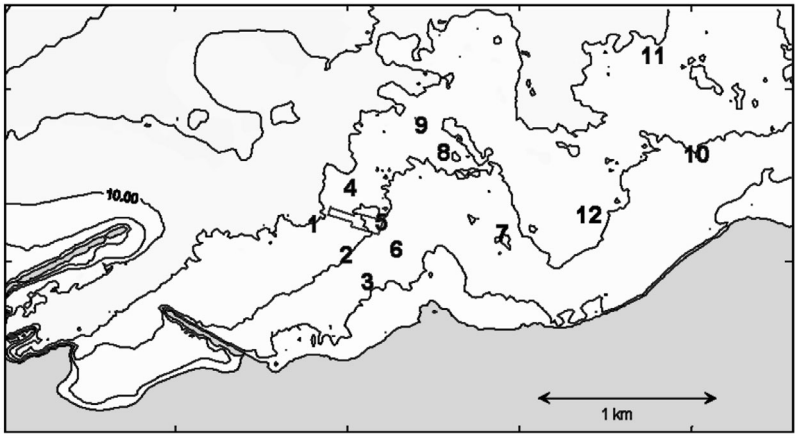

Fig. 1. Study area on the west coast of North Island, New Zealand (39° 02.5′ S, 174° 03′ E), with locations of sampling sites. Depth contours are at 5 m intervals. Sampling sites are labelled 1 to 12. Polygon surrounded by Sites 1, 2, 4 and 5 indicates spoil disposal area. Breakwaters at left of diagram enclose Port Taranaki. Sites 1 to 6 were designated as impacted, Sites 7 to 12 are distant controls

Table 1. Nomenclature

p	Probability of obtaining data at least as extreme as have been obtained assuming that the tested hypothesis was true
μ	Mean of a population
σ, σ^2	Joint common standard deviations, variances of 2 or more populations
CV	Coefficient of variation of a population (the population CV is σ/μ)
ES	Effect size, e.g. $(\mu_1-\mu_0)/\sigma$, a relative measure of the strength of a difference between population 1 and a control population compared with the natural variability. Note that if we define the limits on the equivalence interval in terms of change from the control sites' mean, i.e. as $\Delta = \pm (\mu_1-\mu_0)/\mu_0$, rather than as effect size limits, then equivalence probability calculations can only be calculated once the control sites' coefficient of variation is supplied (because $\Delta = (ES)(CV_0)$)

tests: (1) traditional 2-sided *t*-test for the difference between 2 means, (2) a confidence interval approach to test the equivalence hypothesis (McBride 1999), (3) the 2 one-sided tests (TOST) of inequivalence hypotheses (Schuirmann 1987) which is also a confidence interval approach. Tests of equivalence, inequivalence hypotheses and Bayesian posterior probabilities of a difference being within an equivalence interval, were carried out using basic Excel™ probability functions (McBride 1999). To demonstrate the behaviour of the equivalence tests, we present detection curves for all 3 types of tests. For cases (1) and (2) these are power curves; for case (3) this is the operational characteristic (OC) curve (the complement of a power curve). The abscissa used is the population effect size (ES), the true difference in means divided by their (unknown) common standard deviation. We used power analysis freeware available at www.niwa.co.nz/rc/prog/stats.

As noted above, equivalence tests can only be performed after the analyst states the width of the equivalence interval. Our best professional judgement (given the lack of information regarding the variability of fauna and flora in the area) is that a change of mean density of more than 50% of the natural variability (i.e. half the true standard deviation) is of biological importance. A change of 100% would definitely be of concern. So, for our equivalence tests, we have defined an equivalence interval with boundaries at an effect size of ±50% (at which the true difference in means is half the true standard deviation). In so doing we have sought to cater for issues in defining effects in terms of natural variability (Peterson 1993), i.e. smaller changes in other variables, such as light penetration, may be biologically important. We note the possibility that at some later date there may be a better understanding of impacts with respect to natural variability and the data could be re-analyzed with a different interval. We used a 20% equivalence interval for the analysis of species richness in intertidal sampling, because more information regarding those taxa is available regarding variability of those assemblages (Taranaki Regional Council unpubl. data).

RESULTS

Detection curves

Detection curves (Fig. 2) reveal the fundamentally different features of the 3 test procedures (*t*-tests for a point-null hypothesis, testing the equivalence hypothesis and testing the inequivalence hypothesis, all at the 5% level). For the *t*-test we see that at low sample size (i.e. n = 10) the test is permissive, not precautionary. That is, only large differences (well beyond the equivalence interval) will be regularly detected. This is even more pronounced for tests of the equivalence hypothesis. However, at large sample size the *t*-test becomes ultra-precautionary, often detecting effects well within the equivalence interval. In contrast, the test of the equivalence hypothesis regularly detects effects just a little beyond the equivalence interval. Consider now testing the inequivalence hypothesis. At n = 10, this has the greatest probability of detecting an effect—nearly 100% for all effect sizes, demonstrating why proof of safety is much more difficult than proof of hazard (Bross 1985). With 50 samples this procedure will still routinely detect effects—about 40% of the time if in fact there were only a miniscule effect size present. Note that the detection curves for the equivalence and inequivalence procedures pivot around the 5 and 95 percentile points at the edge of the equivalence region's relevant abscissa values (–50 and +50%) (except when the sample size is small, in which case the true significance level of the inequivalence test cannot be maintained at 5%, see Appendix 1). The difference between the equivalence test procedures is that tests of the inequivalence hypothesis (the proof of safety approach) have about a 95% chance of detecting an effect of ±50%, whereas for equivalence tests (the proof of hazard approach) that probability is about 5%. Thus, testing the inequivalence hypothesis at small α keeps the risk to the environment (the consumer's risk) small, and will often detect an effect even when that effect is some way within the equivalence interval. The reverse is true when testing the equivalence hypothesis at small

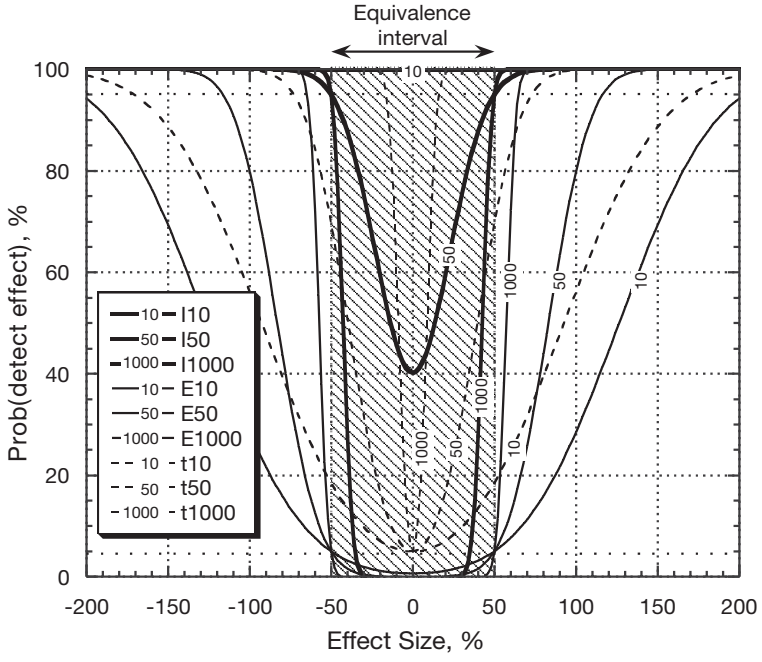

Fig. 2. Detection curves for tests of equivalence (curves with 'E' prefix), inequivalence hypotheses (curves with 'I' prefix) and a *t*-test of a point-null hypothesis (curves with 't' prefix) at 3 sample sizes (10, 50, 1000) and with true variance unknown. The probability (prob.) of concluding that there has been an impact is plotted as a function of effect size, where an impact is defined as causing a change in effect size greater than ±50%. Hatched area is the equivalence (no impact) interval

α; keeping α small protects the producer's risk, and so will regularly detect an effect only if that effect is some way beyond the equivalence interval. These distances within or beyond the interval decrease as the number of samples is increased, giving greater certainty of conclusions.

In summary, detection curves for the *t*-test move from being permissive (at low number of samples) to an increasingly precautionary stance as the number of samples is increased. That is, they move from a propensity to miss important effects to a propensity to detect the trivial. Equivalence test procedures do not exhibit this unfortunate behaviour, and remain consistent with their associated burdens-of-proof.

Subtidal sampling

Prior to disposal, the number of individuals and number of species were slightly higher at control sites than at impact sites (Fig. 3, Table 2). Site means were moderately variable, having SD/mean ratios of more than 75% in 3 of the 4 comparisons. Tests of point-null hypotheses indicated no statistically significant differences (Table 2), and tests of equivalence and inequivalence hypotheses retained their respective hypotheses for both variables (Table 2). The Bayesian posterior probabilities that the true difference between control and impact areas lay within the equivalence intervals for the before survey were 59 and 72% for number of individuals and number of taxa respectively.

After disposal, a similar pattern persisted as prior to disposal; numbers of individuals and species were slightly higher at control sites than at dumpground sites (Fig. 3). Again, there were no significant differences in the tests of null hypotheses, neither test of an equivalence hypothesis rejected its hypothesis, and the Bayesian posterior prob-

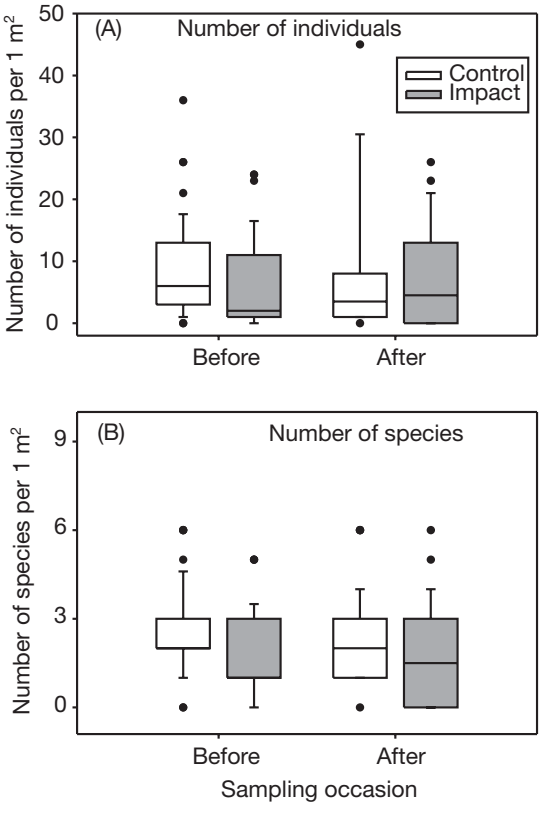

Fig. 3. Boxplots of abundances of (A) total numbers of individuals and (B) total numbers of species in subtidal sampling before (December 1998) and after (September 1999) spoil disposal. Data are the number of organisms or species in individual quadrats; thus, n = 30 (6 sites, 5 quadrats per site) for each bar. Medians for number of species before disposal at control and dumpground areas are 3 and 2, respectively (they coincide with 25th percentile lines). Whiskers extend to non-parametric 5th and 95th percentile values

Table 2. Statistics from the comparisons of number of individuals and number of species for the subtidal environment before and after spoil disposal. Data are based on 6 site means for each treatment. Bayesian probability (prob.) is the posterior probability (as %) that the true difference is within the equivalence interval (using uniform priors). H: hypothesis

	Before		After	
	Control	Impact	Control	Impact
Number of individuals				
Means	8.33	5.83	9.30	7.07
SD	7.47	6.69	12.55	6.91
CV	89.6	114.8	135.0	97.8
H: no difference	Not significant		Not significant	
H: inequivalence	Inequivalent		Inequivalent	
H: equivalence	Equivalent		Equivalent	
Bayesian prob.	59		52	
Number of species				
Means	2.43	1.6	2.13	1.70
SD	0.91	1.26	1.01	1.50
CV	37.3	78.7	47.3	88.0
H: no difference	Not significant		Not significant	
H: inequivalence	Inequivalent		Inequivalent	
H: equivalence	Equivalent		Equivalent	
Bayesian prob.	72		76	

abilities were roughly similar to those before disposal (52% for number of individuals, 76% for number of species) (Table 2).

The failure to reject any tested hypothesis suggests an examination of the adequacy of sample sizes to demonstrate safety. For example, how many samples would be necessary to infer safety, using ±50% effect size equivalence intervals, if in fact the true effect was vanishingly small (i.e. effect size ~ 0)? Assuming sampling from a normal distribution, and an α level of 0.05, calculations indicate that 60 sites per treatment would give a probability of about 70% to detect no effect, and that with 80 sites per treatment this probability rises to about 87%. These results indicate that many more than our 12 sites (6 impacted sites round the dumpground, and 6 more distant control sites) would be necessary to be confident that we might reliably demonstrate absence of effect (defining effect as being larger than the interval, not ES ~ 0). (Note that for the among-site means, CVs of the tested variables ranged from 37 to 135%. Taking a typical CV as 100%, the ±50% effect size limits on the equivalence interval are also ±50% of the control sites mean, see Table 1).

Intertidal sampling, December 2003

The hypothesis of inequivalence of species richness could be rejected in favour of equivalence for Sites 1 and 3 (Bayesian posterior probability [BPP] that the difference is within the equivalence interval >99.9%), whereas both those sites were not equivalent to Site 2 (BPP Site 1 vs 2 = 2.43%, [BPP] Site 2 vs 3 = 0.07%) (Fig. 4). Thus, it is clear that although appropriate sample sizes are higher than for tests of point-null hypotheses, they are not impossibly large, and clear demonstrations of equivalence and inequivalence may be attained.

DISCUSSION

There are 2 major differences between the conventional approach of testing point-null hypotheses and equivalence tests. The first, and more general, distinction is that equivalence tests have to do with an interval hypothesis, which immediately adds realism compared to tests of point hypotheses. The statistical literature contains numerous expressions of doubt regarding the utility of testing an hypothesis that we know *a priori* to be untrue for any real population of continuous variables (e.g. Goodman & Royall 1988, Cohen 1990), and also for most discrete variables. Of course the means of 2 or more field populations differ, it merely remains to be determined whether the sample size was large enough, or the variability small enough, to render that difference statistically significant and if so, whether that difference is environmentally significant. Krebs (1989, p. 8) says that confusing these 2 items is 'The greatest mistake an ecologist can make in the routine use of statistics'. The

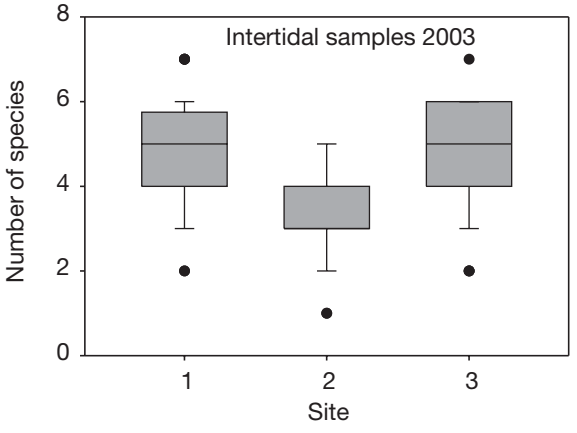

Fig. 4. Boxplots of numbers of mobile invertebrates in intertidal sampling in December 2003. Data are the number of species in individual quadrats (n = 60) at each of 3 sites in New Plymouth. Whiskers extend to 10th and 90th percentiles; box spans 25th and 75th percentiles; median indicated by horizontal line

second point of distinction is that the outcome of a test of an inequivalence hypothesis is exactly that required to implement the precautionary approach, i.e. assume the presence of an important effect unless data convincingly demonstrate otherwise. In contrast, cumbersome procedures are required to render the p-value of a point-null hypothesis meaningful (e.g. Schuirmann 1987, Dixon 1998); even then interpretation of its result can be problematic, as it can be overly precautionary (particularly if variability is small). In testing the inequivalence hypothesis the means are assumed to differ by an important amount, unless data are convincing to the contrary (proof of safety). This immediately weights the decision in favour of the environment, and reverses the usual burden of proof as favoured by Dayton (1998) and Gerrodette et al. (2002). In that case, a developer whose activity poses a risk to the environment must demonstrate safety. They cannot collect only a few samples, fail to reject a null hypothesis, and then conclude that there has been no impact. Stronger, more thorough, impact assessments must result, and the onus falls upon those undertaking activities that could damage the environment to show that their activities are not harmful. One inevitable consequence of demonstrating proof of safety rather than proof of hazard is a substantially increased sampling effort, as we have shown here (Fig. 2). This consequence is also known in the statistical literature (e.g. Bross 1985), and has also been demonstrated in toxicology by Stallard & Whitehead (1996) using bioequivalence procedures.

Our recommended approach receives little mention in ecological texts (e.g. Quinn & Keough 2002, but see Kingsford & Battershill 1998); so what evidence is there that such methods are useful? The test of the inequivalence hypothesis adopted here is required by US FDA (http://www.fda.gov/cder/guidance/1716dft.htm) and EU (http://www.emea.eu.int/pdfs/human/ewp/140198en.pdf) drug trials. If human health is sufficiently important to adopt equivalence procedures, could not the environment be given similar protection? Other barriers to the adoption of the methods include the availability of software. Most software packages have numerous procedures for testing null hypotheses, but lack protocols for tests of interval hypotheses. We list targeted software packages for carrying out equivalence tests in Appendix 1, and are aware of macros circulating to carry out such tests in popular statistical packages. Wellek (2002) contains a webpage (http://www.zi-mannheim.de/wktsheq) with code for most of the numerous analyses he gives, including comparisons of multiple means. All of our analyses were carried out using code written in Excel™. The major barrier to the adoption of the equivalence testing approach appears to be one of familiarity, rather than lack of software. One of our prime objectives here is to draw attention to the large number of difficulties with the widely used p-value approach.

Some workers state that they cannot use equivalence procedures because they are unable or unwilling to specify an equivalence interval; they cannot indicate what an effect of real consequence might be, and so test a point-null hypothesis. In such situations, the appropriate course of action is to estimate the magnitude of the effect, rather than carry out a test with no context. We consider this to be a point of considerable importance. Presenting the result of any statistical hypothesis test, in which the size of a biologically important effect is unknown, gives an aura of precision where little exists. If a researcher cannot specify how big an effect of practical consequence might be, why carry out a test of an untenable hypothesis? Is it merely because one doesn't have to state an important effect size? In that case, what does statistical significance actually mean?

When this study was designed R. Cole had not encountered tests of equivalence hypotheses. We subsequently set a relatively lenient symmetrical 50% equivalence interval as the criterion for our decisions *a posteriori*, as in our experience populations of subtidal organisms can fluctuate considerably. That interval is probably a reasonable first attempt, but as the relevant populations become better known, more stringent or lenient equivalence intervals could be examined. Resource managers are frequently able to dictate monitoring standards that must be met, and therefore we suggest they adopt equivalence protocols. Although the medical derivation of equivalence procedures has probably directed attention toward the comparison of multiple treatments with a single control treatment, in environmental studies it is more usual to compare multiple control sites with a single impacted site (e.g. Underwood 1992, 1994). Such experimental designs are well-suited to the use of equivalence procedures.

Our approach has been that of global safety, where all treatments must be shown to be safe (i.e. inequivalence hypothesis rejected in favour of equivalence for all sites in our case). This is a very demanding condition, and one that is likely to be expensive to demonstrate (see sample size calculations). Hauschke & Hothorn (1998) also consider partial safety, in which at least one comparison shows safety. This more lenient criterion is less demanding of a developer, but is also more hazardous to the environment. We consider that neither approach is necessarily the most useful for point-source impacts in spatially structured situations. In assessments such as ours, there are likely to be clear spatial gradients in the severity of the impact, since the source of impact is clearly identified (the disposal area), and it is effectively a point-source impact. The large sample sizes required by the inequivalence pro-

cedure also offer the opportunity to map the severity of the impact in detail. In such cases, an added benefit of the equivalence testing approach is that, where clear definitions of impact and control are not possible, mapping the extent and severity of impact will be done in more detail. Such assessments of impact could be incorporated into predictions of mixing zones and zones of acceptable impact.

Our findings have important implications for survey design. To demonstrate no effect with the inequivalence procedure, powerful tests and therefore high replication are required. Intra-site variation does not influence the power of the test for an effect of disposal. To obtain a powerful test for the effect of disposal, it is best to minimise replication within sites and maximise the number of sites within each level of disposal. With data such as these (which are not particularly variable compared to abundances of plankton or many reef fishes), sample sizes at least an order of magnitude larger than are usual currently will be required to determine whether there has been an impact, even for the relatively large effects we postulate to be of consequence. That finding is particularly characteristic of inequivalence test procedures (e.g. Bross 1985, our Fig. 2). In our experience, such sample sizes are not prohibitive, and use of innovative technology can further reduce sampling costs.

Acknowledgements. Thanks to R. Dickson, B. Gray, P. McComb, M. McLean, C. Mundy, J. Sait, D. Tindale, and N. Alcock for assistance with fieldwork, P. Atkinson and J. McLean for project supervision, D. J. Morrisey and R. B. Taylor for comments on the manuscript. Partial funding for this work was obtained from Westgate Transport Limited, and the New Zealand Foundation for Research, Science and Technology (contract C01X0215, Sustainability of aquatic systems and water resources).

LITERATURE CITED

Anderson S, Hauck WW (1996) Comment on Berger RL & Hsu JC, Bioequivalence trials, intersection-union tests and equivalence confidence sets. Stat Sci 11:303

Andrew NL (1988) Ecological aspects of the common sea urchin *Evechinus chloroticus*, in northern New Zealand. NZ J Mar Freshw Res 22:415–426

Atherton Skaff PJ, Sloan JA (1998) Design and analysis of equivalence clinical trials via the SAS® system. In: Proc Twenty-Third Annual SAS® Users Group International Conf. SAS Institute, Cary, NC, p 1166–1171

Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of p-values and evidence. J Am Stat Assoc 82:112–122, Discussion 123–139

Bross ID (1985) Why proof of safety is much more difficult than proof of hazard. Biometrics 41:785–793

Browman HI (1999) Negative results. The uncertain position, status and impact of negative results in marine ecology: philosophical and practical considerations. Mar Ecol Prog Ser 191:301–309

Cascio WF, Zedeck S (1983) Open a new window in rational research planning: adjust α to maximise statistical power. Personnel Psychol 36:517–526

Casella G, Berger RL (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J Am Stat Assoc 82:106–111, Discussion 123–139

Cohen J (1988) Statistical power analysis for the behavioural sciences, 2nd edn. Academic Press, New York

Cohen J (1990) Things I have learned (so far). Am Psychol 45: 1304–1312

Cole RG, McComb P, Sait J (1999) Effects of nearshore sand disposal on subtidal and intertidal organisms at New Plymouth, New Zealand. Coasts & Ports '99 Challenges and directions for the new century. Proc 14th Australas Coast Ocean Engineering Conf and the 7th Australas Port and Harbour Conf Vol. 1, National Committee on Coastal and Ocean Engineering, Institution of Engineers, Barton, p 129–134

Cole RG, McBride G, Healy TR (2001) Equivalence tests and sedimentary data: dredge spoil disposal at Pine Harbour Marina, Auckland. J Coast Res Spec Issue 34:611–622

Creese RG (1988) Ecology of molluscan grazers and their interactions with marine algae in north-eastern New Zealand: a review. NZ J Mar Freshw Res 22:427–444

Dayton PK (1998) Reversal of the burden of proof in fisheries management. Science 279:821–822

Dixon PM (1998) Assessing effect and no effect with equivalence tests. In: Newman MC, Strojan CL (eds) Risk assessment: logic and measurement. Ann Arbor Press, Chelsea, MI, p 275–301

Fairweather PG (1991) Statistical power and design requirements for environmental monitoring. Aust J Mar Freshw Res 42:555–567

Gardner MJ, Altman DG (1989) Statistics with confidence: confidence intervals and statistical guidelines. British Medical Journal Publishers, London

Garrett KA (1997) Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. Phytopathology 87: 372–374

Germano JD (1999) Ecology, statistics, and the art of misdiagnosis: the need for a paradigm shift. Environ Rev 7: 167–190

Gerrodette T, Dayton PK, Macinko S, Fogarty MJ (2002) Precautionary management of marine fisheries: moving beyond burden of proof. Bull Mar Sci 70:657–668

Gibbons JD, Pratt JW (1975) p-values: interpretation and methodology. Am Stat 29:20–25

Goodman SN (1993) p-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 137:485–496

Goodman SN (1999) Toward evidence-based medical statistics. 1: the p-value fallacy. Ann Intern Med 130:995–1004

Goodman SN, Royall R (1988) Evidence and scientific research. Am J Public Health 78:1568–1574

Harlow LL, Muliak SA, Steiger JH (eds). (1997) What if there were no significance tests? Lawrence Erlbaum, Mahwah, NJ

Hauschke D, Hothorn LA (1998) Safety assessment in toxicological studies: proof of safety versus proof of hazard. In: Chow SC, Liu JP (eds) Design and analysis of animal studies in pharmaceutical development. Marcel Dekker, Hong Kong, p 197–225

Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55:19–24

Johnson DH (1999) The insignificance of statistical significance testing. J Wildl Manag 63:763–772

Kingsford MJ, Battershill CN (1998) Procedures for establish-

ing a study. In: Kingsford MJ, Battershill CN (eds) Studying temperate marine environments. A handbook for ecologists. Cambridge University Press, Christchurch, p 29–48
Krebs CJ (1989) Ecological methodology. Harper & Row, New York
Lortie CJ, Dyer AR (1999) Over-interpretation: avoiding the stigma of non-significant results. Oikos 78:183–184
MacKenzie DI, Kendall, WL (2002) How should detection probability be incorporated into estimates of relative abundance? Ecology 83:2387–2393
Mapstone BD (1995) Scalable decision rules for environmental impact studies: effect size, Type I and Type II errors. Ecol Appl 5:401–410
McBride GB (1999) Equivalence tests can enhance environmental science and management. Aust NZ J Stat 41:19–29
McBride GB (2002) Statistical methods helping and hindering environmental science and management. J Agric Biol Environ Stat 7:300–305
McComb PJ, Black KP (2001) Dynamics of a nearshore dredged-sand mound on a rocky, high-energy coast. J Coast Res Spec Issue 34:550–563
McComb P, Black K, Atkinson P, Bell R, Healy T (1997) High-resolution wave transformations on a coast with complex bathymetry. Proc 1996 Pacific Coasts and Ports Conf, Christchurch, Vol 2. Centre for Advancd Engineering, University of Canterbury, Christchurch, p 995–1000
McDonald LL, Erickson WP (1993) 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
Morrison DE, Henkel RE (eds) (1970) The significance test controversy — a reader. Aldine, Chicago
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A 231:289–337
Osenberg CW, Sarnelle O, Cooper SD, Holt RD (1999) Resolving ecological questions through meta-analysis: goals, metrics and models. Ecology 80:1105–1117
Palmer AR (1999) Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. Am Nat 154:220–233
Peterson CH (1993) Improvement of environmental impact analysis by application of principles derived from manipulative ecology: lessons from coastal marine case histories. Aust J Ecol 18:21–52
Peterson CH, McDonald LL, Green RH, Erickson WP (2001) Sampling design begets conclusions: the statistical basis for detection of injury to and recovery of shoreline communities after the 'Exxon Valdez' oil spill. Mar Ecol Prog Ser 210:255–283
Poole C (2001) Low p-values or narrow confidence intervals: which are more durable? Epidemiology 12:291–294
Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge
Royall RM (1997) Statistical evidence. A likelihood paradigm. Chapman & Hall/CRC, Boca Raton
Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15:657–680
Sellke T, Bayarri, MJ, Berger JO (2001) Calibration of *p*-values for testing precise null hypotheses. Am Stat 55:62–71
Shrader-Frechette KS, McCoy ED (1992) Statistics, costs and rationality in ecological inference. Trends Ecol Evol 7: 96–99
Stallard N, Whitehead A (1996) An alternative approach to the analysis of animal carcinogenicity studies. Reg Toxicol Pharmacol 23:244–248
Underwood AJ (1990) Experiments in ecology and management: their logics, functions and interpretations. Aust J Ecol 15:365–389
Underwood AJ (1991) Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Aust J Mar Freshw Res 42: 569–587
Underwood AJ (1992) Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. J Exp Mar Biol Ecol 161:145–178
Underwood AJ (1993) The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world. Aust J Ecol 18:99–116
Underwood AJ (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecol Appl 4:3–15
Underwood AJ (1997) Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge
Underwood AJ, Peterson CH (1988) Towards an ecological framework for investigating pollution. Mar Ecol Prog Ser 46:227–234
Wellek S (2002) Testing statistical hypotheses of equivalence. Chapman & Hall/CRC, Boca Raton
Wonnacott TH, Wonnacott RJ (1977) Introductory statistics, 3rd edn. Wiley, New York

Appendix 1. p-values and hypothesis tests

A p-value is defined as the probability of obtaining data at least as extreme as have been obtained assuming that the tested hypothesis was true (e.g. Wonnacott & Wonnacott 1977). The most common hypothesis tested is a 2-sided point-null, positing exact equality between population parameters, e.g. means. There are 2 other options for the form of tested hypotheses. The first is 'one-sided', positing an inequality; for example that the mean of one population is less than that for a second population. The second (intermediate) hypothesis form is that of a 2-sided interval — that a difference (e.g. between means) lies within or beyond an interval, such as the equivalence tests discussed in this paper.

The behaviour of p-values is fundamentally different for tests of these types of hypotheses. For a 2-sided point-null, p tends to become ever smaller as the number of samples (the sample size) increases, as can be shown mathematically. This is because the hypothesis tested in fact cannot be expected to be true, even though it is assumed to be true when calculating the p-value. To see that the hypothesis cannot be true, note that it poses an exact equality, but for continuous variables there will virtually always be some difference between the tested parameters, however small. Consequently, if p is not small (i.e. less than the significance level) one should never infer that a point-null hypothesis is valid, by accepting it. This was always stated by R. A. Fisher (the original proponent of significance tests; see Goodman 1993, Royall 1997). Though not widely understood, procedures of acceptance of such hypotheses, in notions of hypothesis tests advanced by Neyman & Pearson (1933), were proposed for the purposes of a guide to behaviour, i.e. to guide a decision (what should I do?) rather than as a means of inference (what should I believe?), as elaborated by Royall (1997).

In contrast, the p-value for 1-sided tests and for 2-sided equivalence tests can rise or fall as the sample size increases, and the tested hypothesis (which is no longer null) can be accepted for inference.

An interesting consequence of the above behaviour of p concerns Bayesian hypothesis probabilities (i.e. the probability of a hypothesis being true, obtained by using new data to update a prior probability). For a 1-sided test the Bayesian probability and p may be very close to one another (Casella & Berger 1987), whereas for a 2-sided point-null these 2 probabilities are nearly always quite different. In the intermediate case of 2-sided tests of interval hypotheses (equivalence tests) p-values are often rather close to the appropriate Bayesian probability.

Finally, for interval tests with unknown variance, UMPU (Uniformly Most Powerful Unbiased) tests do not exist; the *t*-test is UMPU however. A consequence of this is that the true level of significance of the equivalence tests presented herein may be less than the nominal significance level (α), even though the test's size is α. This is particularly the case at small numbers of samples. That is, the power of the test at the edge of the critical region may be less than the desired 5%, as seen in the inequivalence test in Fig. 2 for n = 10 samples.

Software for calculating equivalence tests available at:
http://www.statsol.ie/equivtest/equivtest.htm
http://www.studysize.com/index.htm
http://pages.prodigy.net/johnsonp12/cd1and2.html
http://www.summitpk.com/pksolutions/pksolutions.htm
http://www.zi-mannheim.de/wktsheq

Dixon (1998) and Atherton-Skaff & Sloan (1998; http://www2.sas.com/proceedings/sugi23/Stats/p218.pdf) show how to calculate tests of equivalence with generic statistical software, and Atherton-Skaff & Sloan (1998) and http://niwa.co.nz/rc/prog/stats/news provide software for power analysis procedures.

Editorial responsibility: Otto Kinne (Editor), Oldendorf/Luhe, Germany

Submitted: December 2, 2002; Accepted: April 6, 2004
Proofs received from author(s): September 7, 2004