Journal of Applied Ecology 2005
42, 4–12

FORUM

Information theory and hypothesis testing: a call for pluralism

PHILIP A. STEPHENS,*† STEVEN W. BUSKIRK,†
GREGORY D. HAYWARD†‡ and CARLOS MARTÍNEZ DEL RIO†

†*Department of Zoology and Physiology, University of Wyoming, PO Box 3166, Laramie, WY 82071, USA; and*
‡*USDA Forest Service, Rocky Mountain Region, PO Box 25127, Lakewood, CO 80225, USA*

Summary

1. A major paradigm shift is occurring in the approach of ecologists to statistical analysis. The use of the traditional approach of null-hypothesis testing has been questioned and an alternative, model selection by information–theoretic methods, has been strongly promoted and is now widely used. For certain types of analysis, information–theoretic approaches offer powerful and compelling advantages over null-hypothesis testing.
2. The benefits of information–theoretic methods are often framed as criticisms of null-hypothesis testing. We argue that many of these criticisms are neither irremediable nor always fair. Many are criticisms of the paradigm's application, rather than of its formulation. Information–theoretic methods are equally vulnerable to many such misuses. Care must be taken in the use of either approach but users of null-hypothesis tests, in particular, must greatly improve standards of reporting and interpretation.
3. Recent critiques have suggested that the distinction between experimental and observational studies defines the limits of the utility of null-hypothesis testing (with the paradigm being applicable to the former but not the latter). However, we believe that there are many situations in which observational data are collected that lend themselves to analysis under the null-hypothesis testing paradigm. We suggest that the applicability of the two analytical paradigms is more accurately defined by studies that assess univariate causality (when null-hypothesis testing is adequate) and those that assess multivariate patterns of causality (when information–theoretic methods are more suitable).
4. *Synthesis and applications*. Many ecologists are confused about the circumstances under which different inferential paradigms might apply. We address some of the major criticisms of the null-hypothesis testing paradigm, assess those criticisms in relation to the information–theoretic paradigm, propose methods for improving the use of null-hypothesis testing, and discuss situations in which the use of null-hypothesis testing would be appropriate. We urge instructors and practitioners of statistical methods to heighten awareness of the limitations of null-hypothesis testing and to use information–theoretic methods whenever prior evidence suggests that multiple research hypotheses are plausible. We contend, however, that by marginalizing the use of null-hypothesis testing, ecologists risk rejecting a powerful, informative and well-established analytical tool.

Key-words: AIC, likelihood, model selection, significance, statistical analysis

Journal of Applied Ecology (2005) **42**, 4–12
doi: 10.1111/j.1365-2664.2005.01002.x

Introduction

For much of the past century, Fisherian or 'frequentist' statistical approaches based on null-hypothesis testing (NHT) have been a central paradigm guiding experimental design and analysis in ecological research

*Present address and correspondence: Philip A. Stephens, Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK (fax +44 117 9287999; e-mail Philip.Stephens@bristol.ac.uk).

and many other sciences. Criticisms of this paradigm date back over six decades (see reviews in Carver 1978; Anderson, Burnham & Thompson 2000) but have gathered pace in the past decade. Recently, information–theoretic model comparison (ITMC) has been suggested as an alternative paradigm for statistical analysis (Anderson, Burnham & Thompson 2000; Burnham & Anderson 2002). Model-selection approaches are powerful tools and ITMC offers many advantages over NHT, especially where multiple hypotheses are plausible or multiple predictors are considered in combination (Johnson & Omland 2004). Quite rightly, ITMC is currently being used effectively in many areas of ecology and evolution and its use is increasing in others (Johnson & Omland 2004; Rushton, Ormerod & Kerby 2004).

Although ITMC is a welcome addition to the analytical arsenal of ecologists and evolutionary biologists, its merits are often framed largely as criticisms of NHT. For example, two of NHT's most severe critics have asserted that 'it should not be surprising that null-hypothesis testing is no longer very useful, considering that it was developed 70–80 years ago' and 'the usefulness of *P*-values is quite limited, and we continue to suggest that these procedures should be euthanized' (Anderson & Burnham 2002). The almost missionary zeal with which NHT has been vilified (Schmidt 1996) has led some to suggest that we are observing a major paradigm shift in our approach to statistical analyses (Guthery, Lusk & Peterson 2001; Rushton, Ormerod & Kerby 2004).

Wildlife biologists, in particular, seem to have embraced ITMC to the exclusion of NHT. As members of a department in which ecologists, evolutionary biologists and wildlife biologists coexist and interact intellectually, we have witnessed a remarkable change in the statistical tone of seminars given by wildlife biologists and their students, and in the manuscripts they submit. Among wildlife biologists, the use of NHT is now viewed as naive and, in many quarters, heretical. Presentations of hypotheses and associated significance levels are often accompanied by apologies for a practice viewed by many as outdated and inappropriate. ITMC seems to have become orthodoxy. Here, we call, not for the rejection of ITMC, but for its integration into the biologist's statistical toolbox, along with existing analytical tools. A key conclusion of Johnson & Omland's (2004) recent review of model selection approaches is that 'biologists must decide when it is most appropriate to use model selection and when it is most appropriate to use … inferences based on significance tests'. We wholeheartedly support this viewpoint but go further: we argue that many criticisms of NHT apply equally to ITMC; rigour in both approaches is essential. Moreover, in some cases there may be room to use both approaches; exploring the basis of different inferences gained from the two methodologies may well increase our understanding of the system under study.

It is not our intention to provide an exhaustive review of all criticisms of NHT and its application. Rather, we hope to focus the attention of readers on some of the major arguments regarding the use and abuse of NHT and ITMC. At a time when many ecologists are becoming increasingly aware of ITMC, we hope to foster debate on the issues, bringing balance to the calls for an end to NHT. With that in mind, we consider six important criticisms of NHT (summarized in Table 1). We recognize that those criticisms are often (although not always) fair but we interpret them as a call to improve, rather than discard, NHT. We conclude that NHT still has considerable utility in both experimental and observational studies and that for some questions NHT is the more appropriate tool.

Null hypotheses are not always uninformative

A primary criticism of NHT is that the paradigm reduces research to a comparison between a typically meaningless, 'trivial' null hypothesis and a single alternative; these trivial hypotheses are known as 'silly nulls' (Anderson *et al*. 2001). First, information criteria offer no protection against trivial hypotheses, including silly a priori model constructions (Guthery, Lusk & Peterson 2001). Indeed, the a priori construction of an appropriate list of multiparameter candidate models (required for information–theoretic methods) demands far more insight from the researcher than a carefully framed series of dichotomous hypotheses (the essence of hypothesis testing). Pioneering research programmes often begin exploring ecological systems with insufficient knowledge to construct sound multiparameter models. Therefore constructing sound ITMC models may not be reasonable when first exploring an ecological system. Secondly, we contend that identifying classically trivial hypotheses in opposition to more plausible alternative ones often forces clarity in our questioning and statistical design. Thirdly, null hypotheses are often less trivial than they first appear. Substantiating null hypotheses (i.e. demonstrating that a factor has no apparent effect) may be very important and 'nature must be ripe with null effects that are [biologically] significant' (Guthery, Lusk & Peterson 2001) [Biological significance refers to the importance (in biological terms) of the measured effect size, in contrast to its statistical significance, defined as $1 - \alpha$ (where α is the probability that the data would have been observed had there been no treatment effect). For example, we may be able to show that two populations have a statistically significant, 2% difference in breeding frequency. Given the numerous other mechanisms that might affect population trajectories, however, we might well show that this is of extremely limited biological significance in terms of population dynamics.].

Null hypotheses may be especially interesting when: (i) the null and its alternative represent the full range of conceivable realities; (ii) we acknowledge that biological significance is of greater importance than statistical significance (Yoccoz 1991; Kirk 1996); and (iii) the null is framed more imaginatively than the standard 'there is no difference between *a* and *b*'.

Table 1. Some common criticisms* of NHT and their relationship with ITMC† (see sections of main text for further details)

Criticism	Relevance to NHT	Relevance to ITMC
Encourages trivial research questions (Johnson 1995; Nester 1996) and focuses attention on statistical (not biological) significance (Yoccoz 1991)	We disagree with the ubiquity of the first of these phenomena (but see Peters 1991) and argue that null hypotheses are often more important than they appear. More commonly, the statistical significance of an effect (i.e. the probability that it is likely to be real rather than a sampling artefact) is mistaken for its importance (i.e. its size). Tackling this requires a combination of better education and stricter journal editing	ITMC poses no obstacle to the examination of trivial hypotheses. Further, it requires substantially greater insight by researchers to ensure that candidate hypotheses adequately represent biologically relevant models and include a good range of possible approximations to the truth. Some researchers may confuse the idea that a model is the 'best' out of a range of candidate models, with the notion that it is a good model
Leads to arbitrary inferences (Johnson 1999) that are often poorly interpreted and incompletely reported Johnson 1999; Anderson, Burnham & Thompson 2000)	These criticisms are well supported. Reporting of NHT statistics must be improved, with a focus on effect sizes and full reporting of inferential statistics, regardless of their relationship to arbitrary cut-offs (e.g. $\alpha = 0{\cdot}05$). It is important to recognize that a failure to reject the null hypothesis is not evidence supporting the null hypothesis. Scientists must move away from the bias towards reporting only positive results, in order to avoid bias in subsequent meta-analyses	ITMC employs 'some simple rules of thumb' (Burnham & Anderson 2001) for assessing the relative merits of competing models. Whether these rules of thumb are any better supported than the $\alpha = 0{\cdot}05$ rule of thumb in NHT is unclear. Any use of confidence intervals implicitly invokes some arbitrary threshold of biological importance. Users of ITMC must also be careful to provide all necessary information
Inappropriate for analysis of observational data (Anderson, Burnham & Thompson 2000)	We disagree that this is always the case. NHT analysis of observational data may be helpful both when we have strong a priori grounds to suspect that one factor will explain an effect and when the question we are trying to answer is unaffected by other potential influences	ITMC offers no solution to the problem that unconsidered parameters may influence data collected in circumstances other than controlled experiments. However, ITMC is likely to be a more powerful tool for inference in systems where multiple factors may underlie an effect
Inappropriate for model selection (Burnham & Anderson 2002)	Where multiple hypotheses are plausible, we agree that a range of flaws exists in NHT approaches to comparing hypotheses. In such cases, ITMC represents a far more powerful inferential tool	Not applicable
May be subject to data dredging (Anderson, Burnham & Thompson 2000)	We accept this criticism of NHT but stress that data dredging may take different forms. A posteriori efforts to improve model fit by adding parameters are to be avoided but thorough and careful exploratory data analyses may well be important and revealing.	ITMC is equally prone to a related problem of 'model dredging'. Instances where researchers generate a few markedly different models based on substantially different parameters (and substantially different hypotheses) are relatively rare. More typical are nested combinations of a large number of parameters, representing subsets of a single overarching hypothesis
Cannot be used in concert with ITMC	We accept that the two inferential paradigms are very different but see no reason why both cannot be used to increase confidence in the findings of studies and to facilitate our understanding of cause and effect	Not applicable

*Note, here we cite a few papers in which these issues are discussed extensively. References to further discussion of the problems can be found in those papers.
†We do not suggest that the criticisms of NHT apply to ITMC as it is proposed (on the contrary, advocates of ITMC have been very careful in drawing attention to misuses of the approach; Anderson & Burnham 2002). However, despite the relative infancy of the approach, some errors of method, presentation and interpretation have already begun to establish within ecology.

Two cited examples of trivial null hypotheses are 'the addition of nitrogen makes no difference to the growth of crops' (Anderson, Burnham & Thompson 2000) and 'lead poisoning makes no difference to the survival rate of ducks' (Guthery, Lusk & Peterson 2001). These examples raise several questions. First, if they really are trivial and implausible, how did we first discover the effects of nitrogen on crops or lead on wildfowl? Do humans possess innate knowledge of such effects? Silliness is context-specific: what is a silly hypothesis at one time might have been highly plausible a few years before. Secondly, are these null hypotheses always implausible (Hagen 1997), even under circumstances that cause them to appear obviously wrong? Under certain conditions, nitrogen may not be a limiting factor for crop growth; equally, lead in the environment (or its

uptake) may be so low, or other toxins so concentrated, that the effect of lead on mortality in ducks may be immaterial. Without demonstrating a measurable effect of nitrogen or lead that, furthermore, is highly unlikely to be the result of chance, should we indulge in fertilizing fields or removing lead pellets from duck habitat, both of which are expensive? The history of ecological management, and biological conservation in particular, is rich in examples of costly mistakes resulting from presumptuous acceptance of apparently 'obvious' hypotheses (Caughley 1994). We must not, by default, denigrate approaches that ask simple questions rigorously, using binary decision paths.

The criticism that null hypotheses are often trivial arises, we suggest, from three common misconceptions. The first of these is that the null should be interesting in its own right. Clearly, however, the NHT analysis structure dictates that the 'interesting' possibilities are framed as the alternative hypothesis. The null is, by definition, the converse of this and therefore often appears uninteresting. The important thing to notice is that the null is not independent but is part of a coupled statement (null and alternative) which defines the area of interest. The two statements should be viewed as a whole, not broken apart.

The second misconception concerns the purpose of the null. When researchers define a null hypothesis as a step towards designing a study, it is rarely their intent that the study should merely support or reject that null. The null and its alternative contain specificity about the parameter(s) or characteristic(s) of the populations to be compared. The null is defined to frame a question about effect sizes (Eberhardt 2003) and, in that context, serves as a yardstick against which to measure an effect. This purpose does not require that the null is plausible, merely that it is chosen to represent a suitable baseline against which to measure the effect of interest.

The third misconception relates to what constitutes a suitable baseline. An apparently widespread belief is that the only acceptable null is designated as 'there is no difference between *a* and *b*' or 'there is no effect of *a* on *b*' (Cohen 1994). Eberhardt (2003) stated that 'if the alternative hypothesis is something other than "no effect", then things get very complicated', but gave no reasons for the cause of this undesired complexity. We are not convinced of the accuracy of this statement. For example, when we use a two-sample *t*-test, we are testing a null hypothesis of the form H_0: $\mu_1 = \mu_2$, where μ_1 and μ_2 represent the means of populations 1 and 2, respectively. Typically, however, we are most interested not in whether there is a difference but, rather, in the magnitude of a potential difference. Thus, it would be equally valid to frame the null in terms of predetermined biological significance (Cohen 1994). For example, consider a common issue in environmental law. When an entity discharges effluent into a watercourse, there is typically a threshold increase in the pollutant above background (control), above which that entity will be found guilty of polluting the watercourse. If that threshold increase were, for example, 5 p.p.m., examining the company's guilt would be a perfect case for NHT and, moreover, for a null constructed on the basis of a consequential difference. Multiple samples could be taken from upstream and downstream of the discharge and we could test a null hypothesis of the form H_0: $\mu_1 + 5 = \mu_2$, where μ_1 is the mean concentration (in p.p.m) of that pollutant upstream of the discharge and μ_2 is the mean concentration downstream of the discharge. Appropriate sample sizes to confer suitable power on such a test (Goodman & Berlin 1994) could be determined in advance and standardized for particular questions. The probability with which H_A could be supported would be the probability of guilt, a value that could be determined by statute or common law. Given that comparators such as Akaike's information criterion (AIC) cannot be used to compare models of different data sets, it is hard to see how an ITMC approach could confront the same type of question with similar rigidity and clarity.

The data collected for the above example would be no different whether we were testing a null of the form H_0: $\mu_1 = \mu_2$ or of the form H_0: $\mu_1 + 5 = \mu_2$. However, nulls chosen with imaginative forethought may be far more plausible, more interesting and, most importantly, more closely focused on the biological significance of our question. Such nulls would also be robust to the criticism that a sufficient sample size will always permit their rejection. Finally, careful prior consideration to what effect size will be of practical importance will facilitate the use of power analysis, enabling researchers to determine in advance whether β (the probability of type II error) can be rendered acceptably low and, thus, whether their data collection is likely to be worthwhile (Cohen 1992).

Use, reporting and interpretation of NHT analyses can be improved

Three well-established criticisms of NHT are that thresholds for statistical significance are arbitrary, that statistical reporting is often uninformative and that the approach is open to abuse. Certainly, statistical reporting by ecologists can be woefully inadequate (Anderson, Burnham & Thompson 2000) and there is nothing sacred about the 0·05 level of α that guarantees an appropriate trade-off between acceptable type I and II error probabilities (a decision which, in many cases, requires some form of cost–benefit analysis of the risks posed to science or society by false positive or false negative results). Increasingly, however, ecologists and journal editors recognize that the reporting of effect sizes, their precision and associated *P*-values (whether 'significant' or 'non-significant') should be mandatory. Thoughtful consideration of the balance between type I and type II errors is also becoming more common. These arguments are a focus of the first edition of *Significance* (the newly launched journal of *The Royal Statistical Society*) (Reese 2004). Information criteria enthusiasts are leading the way by reporting criterion

scores for all tested models (and not discarding some that fall beyond an arbitrary threshold) but the need for better reporting of summary and inferential statistics applies to both NHT and ITMC approaches, a point that the proponents of ITMC have made clear (Anderson *et al*. 2001). We should be as critical of an experimental study that reports significance without reporting effect sizes and their precision, as of an observational one that merely ranks a series of alternative models and their AIC scores.

An important point relating to the arbitrary nature of 'significant' α-values is that whenever we give a confidence interval, we are (consciously or otherwise) employing an α-value to do so. Most commonly we use 95% confidence intervals, utilizing an α-level of 0·05. Whether or not *P*-values are given, any indication that two means differ because their confidence intervals do not overlap, or that a value is unlikely to be one or zero because these fall outside its confidence interval, is a frequentist approach, directly analogous to NHT (Efron & Tibshirani 1993). In spite of warning against mixing NHT with ITMC, Burnham & Anderson (2002) themselves provide several examples of where this may be useful. In particular, when assessing the merits of various competing models during ITMC, it is often necessary to use a variety of methods to make inferences about which model is best. One method is to look at the coefficients of parameters in the models and determine whether these differ from zero. Despite an ironic use of quotation marks around the word 'significant', this is precisely the approach used by Burnham & Anderson (2002). In spite of the emphasis on multiple methods for inference, there remains a danger that such approaches to ITMC could lead to equally arbitrary inferences to those made in NHT.

Finally, three insidious practices of frequentists, widespread in ecology, are equating (i) a failure to reject the null hypothesis 'there is no difference between treatments A and B', with the assertion that 'treatments A and B are the same' (Johnson 1999, 2002); (ii) the probability with which the data could have been obtained, given the null hypothesis, with the probability that the null hypothesis is true (Carver 1978; Cohen 1994); and (iii) poor support for the null hypothesis, with strong support for the alternative hypothesis (Carver 1978). The first of these is common where researchers wish to show that data from two treatments can be grouped for subsequent analysis. A simple, yet common, ploy for doing this is to test for differences between A and B with low power (high type II error probability). Clearly, in this case, the null and its alternative should be reversed and the onus should be on the researcher to prove that the new, non-trivial null (that A and B differ) can be rejected (i.e. the probability that A and B differ is sufficiently small that it can be treated as very unlikely). This approach of reversing hypotheses is the basis for equivalence tests (Robinson & Froese 2004), more commonly employed in analyses of pharmaceutical trials and studies of environmental toxicology.

The second problem relates to ecologists' understanding of the logical basis of NHT and is far less easily treated. In a commentary on the subject of misapplication of NHT, Robinson & Wainer (2002) recommended that overcoming such problems will involve a mixture of enlightened journal editors and improved education of users of statistical procedures. Traditional NHT emphasizes type I error probabilities, whereas a more rigorous Fisherian approach emphasizes the importance of both error types and their associated probabilities.

The third abuse of NHT interpretation (that of equating a low *P*-value as strong evidence for the alternative hypothesis) is perhaps the most widespread of the three. The advice of Carver (1978) should be the focus of all scientists, students and practitioners alike: 'Even if the null hypothesis can be rejected, several other alternative or rival hypotheses still must be ruled out before the validity of the research hypothesis is confirmed. Only after rigorous theorizing, careful design of experiments, and multiple replications of the findings in varied situations should one contend that the probability is high that the research hypothesis is true'.

In summary, the practice of examining ecological data employing NHT includes far too many examples of sloppy implementation, poor reporting, arbitrary selection of significance levels and incorrect interpretation. However, these are not inherent problems with NHT nor is ITMC immune to similar abuse. Most commentaries on the subject (Carver 1978; Cohen 1994; Kirk 1996) have identified the highly informative properties of confidence intervals and have urged their increased use for all effects of interest. We note that modern, computationally intensive techniques are now available to derive confidence intervals for almost any population parameter (Efron & Tibshirani 1991) and we strongly recommend their use.

NHT analyses of observational data may be valuable

Burnham & Anderson (2002) concede that NHT has utility for the analysis of experimental data but caution that analyses of observational studies should be viewed largely as a problem of model selection. The basis of this argument is that in observational studies putative causative factors cannot be isolated by our sampling design. We suggest that: (i) unconsidered parameters that covary with examined parameters can cause problems with inference whether data are analysed using NHT or ITMC; and (ii) where there are strong a priori grounds to expect a single causative factor to be important and no indication that synergistic or confounding factors may covary, and where we are interested less in the underlying reasons for differences between two populations and more in determining whether a biologically significant difference exists, then NHT analyses of observational data may be valuable.

An example of where we have strong a priori grounds to expect that a single factor will explain an effect, is in

a study of the effect of the pollutant atrazine on demasculinization in a frog species. To examine this, we might identify a range of sites in which the species occurs, select a set of sites randomly, and measure both the concentrations of atrazine in surface waters and the proportion of males with developmental abnormalities. We might then test the hypothesis that the dose–response curve in the field would be non-linear and similar to that found in the laboratory, with the highest effect of atrazine found at intermediate levels (Hayes *et al*. 2003). Indeed, we might test the hypothesis that the levels that elicit the highest incidence of developmental abnormalities in male frogs are the same as those that do so in the laboratory.

It would be possible to begin our study of atrazine and frog development with multiple hypotheses that translate into a variety of candidate multiparameter models for comparison. These models may include linear and non-linear effects of a variety of factors, including atrazine concentration and, perhaps, surface water temperature, incident solar radiation, land use and others. However, Burnham & Anderson (2002) stressed that hypotheses should be generated a priori, usually on the basis of personal knowledge concerning the phenomenon. The case of atrazine and frog development is an excellent example of where prior evidence strongly links a single causative factor to a phenomenon. It seems reasonable to conduct a Fisherian study of the consequences of that single factor. If the results indicate that a large amount of the variation between sites in the incidence of developmental abnormality is explained by atrazine concentration, then the need for more complex, multiparameter analyses may be obviated. We may make recommendations on the use of atrazine in relation to conservation without recourse to ITMC methods.

The example given above is a specific case of a more general point. Many observational studies are undertaken because prior evidence indicates that a single factor plays an important role in some natural phenomenon. If these studies (which are ideally suited to analysis by NHT methods) indicate that the factor does indeed explain much of the observed variance, there may be no reason to indulge in more complex multiparameter analyses. Equally, if an NHT study indicates that the studied factor is a poor predictor of the phenomenon of interest, then we may wish to follow the study up with an ITMC approach.

A second situation where NHT is appropriate for evaluating observational data is where we are interested in the differences between two populations without necessarily needing to determine the causes of putative differences. For example, if a researcher had developed a model to assess the status of a population in one area, it may then be useful to apply it to a second area. If density regulation is a key component of that model, it may be important to determine whether territory size differs substantially between the two areas. If it does, elements of the model may need to be adapted for application to the second area. This is true regardless of the factors underlying differences in territoriality between the two areas.

We conclude that NHT is most likely to be of use in observational studies where we are interested in analysing univariate (or occasionally bivariate) causality, either because we have good reason to believe that a univariate model will explain much of the variation in a system, or where insufficient knowledge renders formulation of reasonable models a 'fishing expedition'. Importantly, there is no reason why the sampling distribution of the test statistic cannot be inferred from a post-hoc analysis of, for example, residuals from a regression line (e.g. Motulsky & Ransnas 1987, p. 370) or distribution of data collected. Furthermore, for observational data that do not conform to the null distribution of the test statistic, a variety of non-parametric methods is available to analyse these data using NHT. In particular, non-parametric bootstrapping approaches make no distributional assumptions beyond those indicated by the observed data, but have been shown to be very powerful (Efron & Tibshirani 1991, 1993).

NHT is inferior to ITMC for model selection

A variety of problems exists with using NHT to select between multiple competing hypotheses. These are well summarized by Burnham & Anderson (2002). In particular, making multiple comparisons with a tool designed explicitly for comparisons between a single null and its alternative is inappropriate. This is especially problematic because in NHT procedures used to select between models (using, for example, likelihood ratio tests in stepwise regression procedures), the extent of multiple comparisons is often implicit rather than explicit, and is not always clear. 'All subsets' approaches (where a statistical software package evaluates models, often numbering in the thousands or millions, constructed from every possible combination of measured variables) require little biologically motivated forethought on the part of the researcher and may, as a result, lead to the selection of spurious models of limited biological generality. In NHT model selection, α-levels influence which parameters are accepted or rejected from multiparameter models. A low α-level will lead to the adoption of a highly parsimonious model that, relative to poly-dimensional (or even infinite dimensional) reality, will be highly biased. A less stringent α-level will favour the identification of spurious treatment effects and the inclusion of spurious parameters. There is no satisfactory statistical basis for determining which α-level will lead to an appropriate trade-off between bias and variance, a problem of NHT to which ITMC is not vulnerable (Burnham & Anderson 2002). Finally, the statistical significance of models selected by NHT methods can be heavily influenced both by sample size and number of parameters in candidate models, and selected models may vary according to the precise process of parameter addition and

rejection (Derksen & Keselman 1992). Marginal parameters (with coefficient estimates and standard errors of a similar magnitude) will only be selected when data sets suggest a large absolute coefficient value; as such, the importance of these parameters is very likely to be biased (Burnham & Anderson 2002). Approaches based on maximum likelihood more typically produce unbiased estimators (Wackerly, Mendenhall & Schaeffer 1996). All of these problems with NHT are serious and well documented. We agree that studies which aim to identify the most informative from a variety of multivariate causal models are best analysed using ITMC's weight of evidence approach, and we strongly urge researchers and editors alike to consider the advantages of ITMC over NHT in model selection.

A priori hypotheses do not always reveal underlying phenomena

The proponents of ITMC are devoted champions of a priori hypothesis formulation and severe critics of data dredging (the practice of performing multiple a posteriori interrogations of data in order to determine whether unexpected hypotheses might explain patterns in the data). We absolutely agree that the practice of 'chasing' observed phenomena by iteratively fitting additional parameters to explanatory models is misdirected and very likely to lead to overfitted models with low generality (Ginzburg & Jensen 2004). However, biologists cannot even begin to explain every biological phenomenon in advance and some phenomena can, and probably should, surprise us. Lacking NHT, a list of plausible, multiparameter candidate models for evaluation (complete with interactions and non-linearities where appropriate) is left entirely to the intuition and qualitative imagination of the model builder. This would be a poor substitute for candidate models informed by exploratory data analysis (EDA) and there is no reason why the results of those analyses should not be published (Eberhardt 2003). Frair *et al*. (2004) give an example of how NHT-based exploratory analyses can be used to inform candidate models for subsequent ITMC analysis. Often the most interesting and fertile results of even the most carefully planned research are unexpected. Detailed exploration of the data using a range of graphical and statistical approaches may reveal unexpected patterns that lead to new hypotheses.

If only models that are defined a priori are deemed worthy of consideration, researchers may be restricted in their outlook or willingness to consider previously unthought of hypotheses, and might be prevented from detecting surprises in their data. Data dredging may be a 'poor approach for making reliable inferences about sampled populations' (Anderson, Burnham & Thompson 2000) but it can be a fertile source of novel hypotheses and plausible candidate models (Anderson & Burnham 2002). The proponents of ITMC seem to throw a blanket of condemnation on EDA and its emphasis on open-minded exploration (Hoaglin, Mosteller & Tukey 2000). Of course we must be wary of the statistical meaning of the 'inferences' that can be gleaned from data dredging, but we should not, as a matter of principle, disregard the insights that a creative and thorough data analysis can bring. There is a danger that a stifling adherence to rigid methodologies may arise from the emphasis placed on a priori hypotheses by proponents of ITMC.

Analyses that employ both frequentist and information–theoretical approaches may be revealing

Anderson and colleagues warned that researchers should not mix ITMC and NHT, as this involves mixing differing inferential paradigms (Anderson *et al*. 2001; Anderson & Burnham 2002; Burnham & Anderson 2002). We agree that, given the broadly different circumstances under which the two methods are most appropriate, this will often be good advice. However, we suggest that there are three reasons why this prohibition should not be taken as absolute. First, as we have already discussed, NHT can be used to both bolster confidence in the coefficients of models (see the section on statistical reporting, above) and inform candidate models for ITMC analysis (see the comments on a priori hypotheses above).

Secondly, users of ITMC often judge competing hypotheses in terms of differences in their information criteria scores, e.g. Δ_i. In contrast to the absolute measures given by NHT statistical analysis, Δ_i is obviously a relative parameter, meaningful only in terms of its relation to the best model examined. Some authors have suggested that concentrating on the information criteria of models may be less important than more pragmatic concerns of accuracy (Chatfield 1995) and it appears that accepting a 'best' model or set of models on the basis of AIC may not always result in the selection of the most useful model (C. Meyer & M. Ben-David, pers. comm.) or even of an adequate model. More traditional approaches, such as judging models on the basis of goodness of fit or classification success, along with testing of residuals, may be of more use to those for whom classification success of a model is more important than knowing it is the best model of those evaluated. For resource selection models, in particular, a range of methods is available to judge the quality of the best model. These include the use of kappa statistics (Boyce *et al*. 2002) and relative operating characteristic curves (Pearce & Ferrier 2000).

Finally, if our objective is to maximize our understanding of a system and develop a model that best approximates reality (within the boundaries of supportable parameters), there is an argument that we should use whatever means are available to do so. If conventional NHT approaches and ITMC analyses, respectively, imply that different models best describe a phenomenon, understanding why the results of these two approaches differ will help to isolate the assumptions inherent in our analyses, and may lead to a more

comprehensive understanding of the system. If, by contrast, both methodologies suggest one model to be pre-eminent, this will boost our confidence in the robustness of the model we select. Such a comparison (between ITMC methods and hierarchical partitioning) underlies the confidence of Gibson *et al*. (2004) in identifying the relative importance of variables underlying habitat selection in the rufous bristlebird *Dasyornis broadbenti*. A combination of information theory and likelihood ratio tests was also used to develop powerful methods to analyse factors underlying observed time-series of population dynamics (Dennis & Taper 1994; Dennis & Otten 2000).

Conclusions

Ecological research strives for an ideal: the development of predictive mechanistic models that can be applied outside the spatial or temporal context within which they are parameterized, i.e. models of general predictive utility. However, much research will only help to identify one small piece in that puzzle. For example, we might wish to build a comprehensive model to predict habitat use in a certain species, X. This does not mean that research that aims to determine whether X preferentially preys on prey species A or prey species B is worthless. That finding may, ultimately, allow the development of a mechanistic understanding of where X is likely to be found. Furthermore, that finding (assuming effect sizes, their precision and *P*-values are reported adequately) may well be based on a hypothesis test of the null 'Species X does not preferentially select prey of species A or B'. There is a danger that an emphasis on full, explanatory models may lead us to overlook the fact that simple questions can be both interesting and informative.

Many criticisms of NHT stem from sloppy application and reporting by scientists employing the approach; these criticisms, although fair, are not irremediable. Effective application of ITMC requires similar vigilance executing the critical steps in the process: model formulation, analysis, model selection and interpretation. We contend that many biologists currently employing ITMC have limited training in the subtle art of translating biological hypotheses into suitable statistical models, despite the fact that successful application of ITMC depends critically on this step. Therefore, the pressure to adopt ITMC and dispense with NHT in wildlife biology may result in a new era dominated by investigations searching for the 'best' model among an array of biologically meaningless candidates. We will have substituted model dredging for data dredging.

Anderson, Burnham & Thompson (2000) stated that 'the fundamental problem with null-hypothesis testing is not that it is wrong (it is not), but that it is uninformative in most cases, and of relatively little use in model and variable selection'. We contend that selecting hypotheses with care and improving our reporting of NHT statistics will help to ensure that NHT is not uninformative. Furthermore, advocating ITMC need not automatically involve denigrating NHT. When multiple causal factors are considered, ITMC is clearly more useful than NHT, avoiding many of the pitfalls implicit in the supposed comparison of two (and only two) complementary hypotheses. However, NHT remains a valuable tool for investigating univariate causality. Both approaches deserve a place in the statistical toolbox available to researchers in ecology and evolution. As Johnson & Omland (2004) observed, the two approaches are appropriate in different circumstances. It is up to us, as biologists, to recognize those circumstances and to make the most of both tools.

Acknowledgements

This manuscript began life during lunchtime discussions involving a range of ecologists, physiologists and wildlife biologists at the University of Wyoming. We would like to thank all of the participants of those discussions for their input at the time and for subsequent suggestions and constructive criticism. We also thank Merav Ben-David, Rudy King and Chris Nations for constructive and insightful comments on an earlier draft of this manuscript. P. A. Stephens was supported by funding from the USDA Forest Service International Programs.

References

Anderson, D.R. & Burnham, K.R. (2002) Avoiding pitfalls when using information–theoretic methods. *Journal of Wildlife Management*, **66**, 912–918.

Anderson, D.R., Burnham, K.P. & Thompson, W.L. (2000) Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, **64**, 912–923.

Anderson, D.R., Link, W.A., Johnson, D.H. & Burnham, K.P. (2001) Suggestions for presenting the results of data analyses. *Journal of Wildlife Management*, **65**, 373–378.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Burnham, K.P. & Anderson, D.R. (2001) Kullback–Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, **28**, 111–119.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information–Theoretic Approach*, 2nd edn. Springer-Verlag, New York, NY.

Carver, R.P. (1978) The case against statistical significance testing. *Harvard Educational Review*, **48**, 378–399.

Caughley, G. (1994) Directions in conservation biology. *Journal of Animal Ecology*, **63**, 215–244.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A, Statistics in Society*, **158**, 419–466.

Cohen, J. (1992) A power primer. *Psychological Bulletin*, **112**, 155–159.

Cohen, J. (1994) The earth is round ($P < 0{\cdot}05$). *American Psychologist*, **49**, 997–1003.

Dennis, B. & Otten, M.R.M. (2000) Joint effects of density dependence and rainfall on abundance of San Joaquin kit fox. *Journal of Wildlife Management*, **64**, 388–400.

Dennis, B. & Taper, M.L. (1994) Density dependence in time series observations of natural populations: estimation and testing. *Ecological Monographs*, **64**, 205–224.

Derksen, S. & Keselman, H.J. (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, **45**, 265–282.

Eberhardt, L.L. (2003) What should we do about hypothesis testing? *Journal of Wildlife Management*, **67**, 241–247.

Efron, B. & Tibshirani, R. (1991) Statistical data analysis in the computer age. *Science*, **253**, 390–395.

Efron, B. & Tibshirani, R. (1993) *An Introduction to the Boostrap*. Chapman & Hall, New York, NY.

Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2004) Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology*, **41**, 201–212.

Gibson, L.A., Wilson, B.A., Cahill, D.M. & Hill, J. (2004) Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology*, **41**, 213–223.

Ginzburg, L.R. & Jensen, C.X.J. (2004) Rules of thumb for judging ecological theories. *Trends in Ecology and Evolution*, **19**, 121–126.

Goodman, S.N. & Berlin, J.A. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200–206.

Guthery, F.S., Lusk, J.J. & Peterson, M.J. (2001) The fall of the null hypothesis: liabilities and opportunities. *Journal of Wildlife Management*, **65**, 379–384.

Hagen, R.L. (1997) In praise of the null hypothesis statistical test. *American Psychologist*, **52**, 15–24.

Hayes, T., Haston, K., Tsui, M., Hoang, A., Haeffele, C. & Vonk, A. (2003) Atrazine-induced hermaphroditism at 0·1 ppb in American leopard frogs (*Rana pipiens*): laboratory and field evidence. *Environmental Health Perspectives*, **111**, 568–575.

Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (2000) *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, New York, NY.

Johnson, D.H. (1995) Statistical sirens: the allure of nonparametrics. *Ecology*, **76**, 1998–2000.

Johnson, D.H. (1999) The insignificance of statistical significance testing. *Journal of Wildlife Management*, **63**, 763–772.

Johnson, D.H. (2002) The role of hypothesis testing in wildlife science. *Journal of Wildlife Management*, **66**, 272–276.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Kirk, R.E. (1996) Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, **56**, 746–759.

Meyer, C.B., Ben-David, M. & Herreman, J.K. (in press) Resource selection functions: a cautionary note on model selection. *Ecology*, in press.

Motulsky, H.J. & Ransnas, L.A. (1987) Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *Faseb Journal*, **1**, 365–374.

Nester, M.R. (1996) An applied statistician's creed. *Applied Statistics, Journal of the Royal Statistical Society Series C*, **45**, 401–410.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Peters, R.H. (1991) *A Critique for Ecology*. Cambridge University Press, Cambridge, UK.

Reese, R.A. (2004) Does significance matter? *Significance*, **1**, 39–40.

Robinson, A.P. & Froese, R.E. (2004) Model validation using equivalence tests. *Ecological Modelling*, **176**, 349–358.

Robinson, D.H. & Wainer, H. (2002) On the past and future of null hypothesis significance testing. *Journal of Wildlife Management*, **66**, 263–271.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species' distributions? *Journal of Applied Ecology*, **41**, 193–200.

Schmidt, F.L. (1996) Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, **1**, 115–129.

Wackerly, D.D., Mendenhall, W. & Schaeffer, R.L. (1996) *Mathematical Statistics with Applications*, 5th edn. Wadsworth Publishing Co., Belmont, CA.

Yoccoz, N.G. (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72**, 106–111.

Received 10 August 2004; final copy received 22 August 2004