Ecology, statistics, and the art of misdiagnosis: The need for a paradigm shift

Joseph D. Germano

Abstract: This paper approaches ecological data analysis from a different vantage point and has implications for ecological risk assessment. Despite all the advances in theoretical ecology over the past four decades and the huge amounts of data that have been collected in various marine monitoring programs, we still do not know enough about how marine ecosystems function to be able to make valid predictions of impacts before they occur, accurately assess ecosystem "health," or perform valid risk assessments. Comparisons are made among the fields of psychology, social science, and ecology in terms of the applications of decision theory or approach to problem diagnosis. In all of these disciplines, researchers are dealing with phenomena whose mechanisms are poorly understood. One of the biggest impediments to the interpretation of ecological data and the advancement of our understanding about ecosystem function is the desire of marine scientists and policy regulators to cling to the ritual of null hypothesis significance testing (NHST) with mechanical dichotomous decisions around a sacred 0.05 criterion. The paper is divided into three main sections: first, a brief overview of common misunderstandings about NHST; second, why diagnosis of ecosystem health is and will be such a difficult task; and finally, some suggestions about alternative approaches for ecologists to improve our "diagnostic accuracy" by taking heed of lessons learned in the fields of clinical psychology and medical epidemiology.

Key words: statistical significance, Bayesian statistics, risk assessment.

Résumé : Dans ce travail, l'auteur aborde l'analyse des données écologiques avec une perception différente ayant des implications pour l'évaluation des risques environnmentaux. En dépit de tous les progrès de l'écologie théorique au cours des quatre dernières décennies, et de l'imposante quantité de données qui ont été récoltées dans divers programmes de suivi marin, nos connaissances sur le fonctionnement du milieu marin sont toujours insuffisantes pour permettre de faire des prédictions valables sur les impacts avant qu'ils ne surviennent, pour évaluer avec précision l'« état de santé » des écosystèmes, ou pour effectuer des évaluations de risques valides. L'auteur fait des comparaisons avec les domaines de la psychologie, des sciences sociales et de l'écologie en termes d'applications de la théorie de la décision ou de l'approche diagnostique des problèmes. Dans toutes ces disciplines, les chercheurs sont confrontés à des phénomènes dont les mécanismes ne sont que partiellement compris. Une des plus grandes embûches dans l'interprétation des données écologiques et l'avancement des connaissances sur le fonctionnement des écosystèmes prend racine dans la détermination des scientifiques marins et des responsables des règlements de coller de façon rituelle au test de signification basé sur l'hypothèse nulle (NHST), avec des décisions dichotomiques mécaniques autour du critère sacré de 0,05. Le travail comporte trois sections principales: d'abord, une brève revue de mauvaises compréhensions communes au sujet du NHST; deuxièmement, pourquoi le diagnostique de la santé des écosystèmes est et sera une tâche difficile; et finalement, quelques suggestions à propos d'approches alternatives permettant aux écologistes d'améliorer la précision des diagnostiques, en prenant note des leçons venant de champs d'activités tels que la psychologie clinique, et la médecine épidémiologique.

Mots clés : signification statistique, statistique bayésienne, évaluation de risques.

[Traduit par la Rédaction]

Introduction

The ideas in this paper are the result of more than 15 years of working a wide variety of applied problems in marine ecology. Most of the arguments to follow address marine environmental investigations as well as the emerging field of ecological risk assessment from a new vantage point. However, many of the ideas presented will be equally applicable to terrestrial or freshwater ecology.

For most of my academic and professional career, I have felt a vague discomfort that despite all the advances in theoretical ecology over the past four decades and the huge amounts of data that have been collected in various regional marine monitoring programs, we still do not know enough about how marine ecosystems function to be able to make valid predictions of impacts before they occur, accurately assess ecosystem "health," or perform valid risk assessments. The current situation with the decimation of commercial fishing stocks in the northwest Atlantic on both Georges Banks and the Grand Banks is a good example of our lack of knowledge or predictive power concerning ecosystem dynamics. While repeated trips down well-worn paths (e.g., collection and analysis of more benthic samples, repeated bioassay tests with unknown relevance to actual field impacts) may provide job security for those of us in the applied side of the field, it has become increasingly difficult for me to

Received June 29, 1999. Accepted November 12, 1999. Published on the web: December 10, 1999.

J.D. Germano. EVS Environment Consultants, 200 W. Mercer Street, Suite 403, Seattle, WA 98119, U.S.A. Telephone: (206) 217–9337; Fax: (206) 217–9343; e-mail: joeg@evs-eco.com

recommend to clients that we should keep running down the same blind alleys.

During the past eight years, most of the scientific papers I have read have been from clinical psychology journals, and they have served as a springboard for further readings in the fields of statistics, decision theory, and expert judgment. I have been struck by the similarity among the fields of psychology and social science with the field of ecology in terms of the applications of decision theory or approach to problem diagnosis. In both disciplines, researchers are dealing with phenomena whose mechanisms are poorly understood. Just as there are no hard and fast "laws of nature" in psychology or social science, we are also a long way from understanding how marine ecosystems function. It is just as difficult for a psychologist or psychiatrist to predict the occurrence of mental disease or diagnose character and personality disorders and for the clinician to prescribe the most effective course of treatment as it is for a marine scientist or risk assessor to predict when the next algal bloom or red tide will occur or diagnose ecosystem health and prescribe the most effective remediation technique.

With the recent advances in computer hardware and software coupled with the wealth of on-line information available on the Internet, ecologists now have easy access to large environmental databases with powerful workstations that will provide the means for achieving new insights into ecosystem structure and function (Germano, In press). However, fresh insights will only come about if we are willing to change our approach to environmental data analysis and not just apply these new tools as a way of doing the same old thing faster. Our current ability to diagnose ecosystem health accurately or to make valid predictions of recovery and (or) the effectiveness of alternative remediation treatments ranges from being extremely limited in many cases to nonexistent in others. Everyone will freely admit that ecological systems are extremely complex, and this simple statement has important implications for the development of ecological indicators (Bernstein 1990); worthwhile indicators cannot be developed without a valid model or thorough understanding of how ecosystems function. Psychologists have faced the same dilemma; people are complex, and it is difficult to find universal principles of human behavior. As Nunnally (1960) observed, this makes psychological research difficult and frustrating, and often this frustration has led to a "flight into statistics" (p. 649); the same can be said of ecological research over the past three decades.

However, we ecologists would do well to take heed of lessons learned in the fields of clinical psychology, decision theory, and medical epidemiology to structure a new approach to the design and interpretation of environmental monitoring projects, ecological research programs, and the emerging field of ecological risk assessment. While all of the arguments that follow are subject to revisions by future comments, criticism, and research, my main thesis is that we need to take a radically different approach to the interpretation of ecological data than those commonly employed today if we are to make any advances in our understanding of ecosystems and our ability to predict the impacts of our activities (and thereby have environmental regulations that are both protective and sensible). The majority of the ideas I will be presenting on expert judgment and statistical inference are hardly original. However, the bulk of the material exists in journals that few marine scientists or ecologists normally read. While it was a little embarrassing for me to discover that most of the articles that provided these insights were published 10–20 years before I even entered graduate school in the mid-1970s, I was not exposed to them during my academic or ensuing professional career, so I would not be surprised if some of the ideas appear as "new information" to other scientists and environmental policy regulators.

One of the biggest impediments to the interpretation of ecological data and the advancement of our understanding about ecosystem function is the desire of marine scientists (biologists, ecologists, chemists, toxicologists, etc.) and policy regulators to cling to the ritual of null hypothesis significance testing (NHST) with mechanical dichotomous decisions around a sacred 0.05 criterion. The continued blind application and misinterpretation of the "Fisherian" school of statistics (which all of us were taught in graduate school) appears to have stifled or limited our understanding of complex systems; psychologists and social scientists recognized these limitations years ago (e.g., Bakan 1966; Berkson 1938; Carver 1978; Lykken 1968; Meehl 1967; Rozeboom 1960), but few marine ecologists are aware of them. Many of the ideas I will present on the misapplication of NHST may either touch a few nerves or will be "what everyone knows." As Bakan (1966) stated so aptly in his wonderful review of the crisis of statistical testing in psychological research,

> To say it 'out loud' is, as it were, to assume the role of the child who pointed out that the emperor was really outfitted only in his underwear (p. 423).

This paper is not intended as a blanket criticism of ecology, statistics, or the test of significance when it can be appropriately used (which I have come to think of as a rare rather than routine situation). The real problem is when ecologists or regulators use NHST or just statistical significance to carry most of the burden of scientific inference; we need better approaches to both ecosystem diagnosis and the prediction of effective remediation alternatives for more realistic risk assessment studies. Most ecologists suffer from what Gould (1981) coined as "physics envy"; hence, the need to use numbers, multivariate statistics (the more complicated, the better!), and computers to convey the impression that indeed some hard science instead of qualitative or subjective judgment is being carried out (witness the continued practice of marine benthic ecologists calculating diversity indices from benthic data sets, a tribal ritual of data-processing handed down to successive cohorts in graduate school for the last three decades).

Suggesting alternative approaches to NHST is not an altogether novel concept for ecology and risk assessors (e.g., Reckhow and Chapra 1983; Reckhow 1990; Crane and Newman 1996; Hill 1996); however, the information about the problems with NHST as well as reasons for considering alternative approaches are scattered throughout a wide variety of journals.

Even though some ecologists are well aware of these problems, a large number of policy regulators, applied scientists (e.g., consultants), and graduate students are not; misunderstandings about what statistical significance testing can and cannot do are still widespread. As Stearns (1976) did with life-history tactics, I have attempted to organize a large body of literature and write the paper that I wanted to read, but could not find, when I started out on this path. The paper is divided into three main sections: first, a brief overview of common misunderstandings about NHST; second, why diagnosis of ecosystem health is and will be such a difficult task; and finally, some suggestions about alternative approaches for ecologists to improve our "diagnostic accuracy."

Part 1: Some common misconceptions about statistical significance

The aim of any test of significance is to obtain information concerning a characteristic of a *population* that is not observable directly (whether for practical or intrinsic reasons); what *is* observable is the *sample*. If the actual population values could be measured, there would be no need to do any statistical tests; we could simply measure the parameter of interest, make a comparison, and immediately know the answer. The test of significance is supposed to aid the researcher in making inferences from the observed sample to the unobserved population. The critical assumption involved in significance testing is that if the experiment or measurements are conducted properly, then *the characteristics of the population will have a determinative influence on the samples drawn from it*; for example, the mean of a population has a determinative influence on the mean of a sample drawn from it. Therefore, if $\boldsymbol{P}$, the population characteristic, has a determinative influence on $\boldsymbol{S}$, the sample characteristic, then there is some justification for the researcher to make inferences from $\boldsymbol{S}$ to $\boldsymbol{P}$ (Bakan 1966).

The literature on the misapplication of statistical significance and misunderstandings about NHST in the field of psychology is rich (e.g., Bakan 1966; Lykken 1968; Morrison and Henkel 1970; Carver 1978; Cohen 1994, and all the references cited in these publications); I would encourage the reader to consult these for more details if the overview presented below whets your appetite. While Bakan (1966) claimed that his chastisement of statistical significance "is hardly original" and compared it to the naked emperor of childhood fables, Cohen (1994) pointed out almost 30 years later that despite all prior published warnings, "this naked emperor has been shamelessly running around for a long time." (p. 997).

If these ideas were not original three decades ago, I certainly cannot claim any originality for them now, and it is equally sobering to read psychological literature 20–30 years after these convincing arguments were initially published and realize that the problems with NHST are still running rampant in this field where investigators are at least informed and aware they exist. While I have no illusions that making marine ecologists aware of the problems with NHST are going to cause an instant 180° change in their approach to monitoring design or data interpretation, an awareness of and admission that the problem even exists is the first step towards change. Also, it is important to keep in mind that statistical significance testing can involve more than one procedure because it has evolved from more than one source (Clark 1963); unfortunately, most introductory statistics texts (and hence most marine scientists) are confined to one procedure. So, with these disclaimers in mind, what exactly are the problems with NHST? I think they can be grouped under four main headings.

Problem 1: The illusion of attaining improbability

The main problem, simply stated, is that statistical significance does not tell us what we want to know; however, because we are desperately trying to find out or prove what we want to know, we either ignore or misunderstand what NHST does and think it is telling us exactly what we want. What we are constantly trying to find out either through our research or monitoring studies is, "Given these data, is my research hypothesis ($\boldsymbol{H}_1$) true?" Most of us will recall that the *p* value from a statistical test (such as the *t* test or *F* test) is a *probability* or *proportion* of the time we can expect to find mean differences as large as or larger than the particular sized difference we get when we are sampling from the same population assumed under the null hypothesis. So, to rephrase our initial underlying desire, what we would like to know as a result of our statistical significance test is, "Given these data, what is the probability that my research hypothesis ($\boldsymbol{H}_1$) is true?" In our more sophisticated moments, we may realize that the statistical tests we are performing are testing the null hypothesis ($\boldsymbol{H}_0$), so we may unconsciously think the *p* value is telling us, "Given these data, what is the probability that the null hypothesis ($\boldsymbol{H}_0$) is true?" However, what the *p* value is telling us (which many investigators can correctly indicate if asked directly, even though they are misinterpreting NHST) is, "Given that $\boldsymbol{H}_0$ is true, what is the probability of these (or more extreme) data?" These four statements:

1. Given these data, is my research hypothesis ($\boldsymbol{H}_1$) true?
2. Given these data, what is the probability that my research hypothesis ($\boldsymbol{H}_1$) is true?
3. Given these data, what is the probability that the null hypothesis ($\boldsymbol{H}_0$) is true?
4. Given that the null hypothesis ($\boldsymbol{H}_0$) is true, what is the probability of these (or more extreme) data?

are *not* equivalent, as has been pointed out many times over the years by the investigators cited at the beginning of this section, and *all* that the *p* value indicates from NHST is the fourth statement. The implications of an investigator designing experiments or interpreting data while confusing these four are not trivial. Statistical significance (our convention of $p \leq 0.05$) simply means statistical rareness (Carver 1978). Results are considered "significant" because they would occur rarely in random sampling from a population under the conditions of

the null hypothesis. In other words, a statistically significant result means the probability is low we would get the result obtained *given that the null hypothesis is true* (Statement 4).

Given this stark reality, statistical significance, by itself, means little or nothing; as Carver (1978) pointed out in his excellent review article, the real problems occur when it is used to make inferences (a trend all too common these days in the world of environmental data interpretation, e.g., in the search for sediment quality criteria). The important contribution that Fisher made in his approach of statistical significance and null hypothesis rejection is that he recognized that it is rarely meaningful to set up any simple "***P*** implies ***S***" model for parameters in which we are interested. In the case of the mean, for example, it is rather that ***P*** has a determinative influence on the *frequency* of any specific ***S***. However, one experiment or measurement does not provide many values of ***S*** to study the frequency, it only gives *one* value of ***S***. The *sampling distribution* is conceived, which specifies the relative frequencies of all possible values of ***S***; then, with the help of an adopted level of significance, we can, in effect, say that ***S*** was false. That is, any ***S*** that fell in a region whose relative theoretical frequency under the null hypothesis was 5% would be *considered* false. IT IS IMPORTANT TO RECOGNIZE that one of *the* essential features of the Fisher approach is what could be termed the "once-ness" of the experiment; Fisher's inference model takes as critical that the experiment or measurement has been conducted *once*. If an ***S***, which has a low probability under the null hypothesis actually occurs, it is taken that the null hypothesis is false. As Fisher himself put it (1947, p. 14), why should the theoretically rare event under the null hypothesis actually occur to "us"? If it does, we take it that the null hypothesis is false. Basic to this is the idea that "the theoretically unusual does not happen to me" (Bakan 1966).

Cohen (1994) pointed out that a basic major problem of NHST arises from a misapplication of deductive syllogistic reasoning, or the "illusion of attaining improbability." The arguments of deductive reasoning below are taken from Cohen's (1994) paper and are a somewhat different presentation of the four statements above by their syllogistic derivatives. The following construct mimics the reasoning of the null hypothesis rejection:

Syllogism Type A: (direct or absolute logical implication)

> If the null hypothesis is correct, then this datum (*D*) cannot occur.
>
> It has, however, occurred.
>
> Therefore, the null hypothesis is false. (Cohen 1994; p. 998)

If this were the only reasoning that investigators used to interpret $\boldsymbol{H}_0$ testing, then it would be formally correct. This is an excellent example of what in Aristotelian logic is called a *modus tollens*, where the denial of the antecedent is derived from a denial of the consequent. What NHST does is to make the reasoning of denial probabilistic, as follows:

Syllogism Type B: (probabilistic logical implication)

> If the null hypothesis is correct, then these data are highly unlikely.
>
> These data have occurred.
>
> Therefore, the null hypothesis is highly unlikely. (Cohen 1994; p. 998)

The real crux of the problem is that by making it probabilistic, the reasoning becomes invalid. Cohen (1994) delightfully illustrated this point by providing the following syllogisms with formally correct *modus tollens* (this is another example of Syllogism Type A, based on what is hopefully a correct initial premise):

> If a person is a Martian, then he is not a member of Congress.
>
> This person is a member of Congress
>
> Therefore, he is not a Martian.

While this may sound reasonable, the following syllogism is also a direct or absolute logical implication, but is not reasonable because the major premise is wrong. However, the reasoning is the same as before and is still a formally correct *modus tollens* (Syllogism Type A, wrong initial premise):

> If a person is an American, then he is not a member of Congress (**WRONG!**)
>
> This person is a member of Congress.
>
> Therefore, he is not an American.

Now, if we make the major premise reasonable by changing it from being a direct or absolute logical implication to a probabilistic one, then the syllogism becomes formally incorrect and leads to a conclusion that is not sensible (Syllogism Type B, correct initial premise):

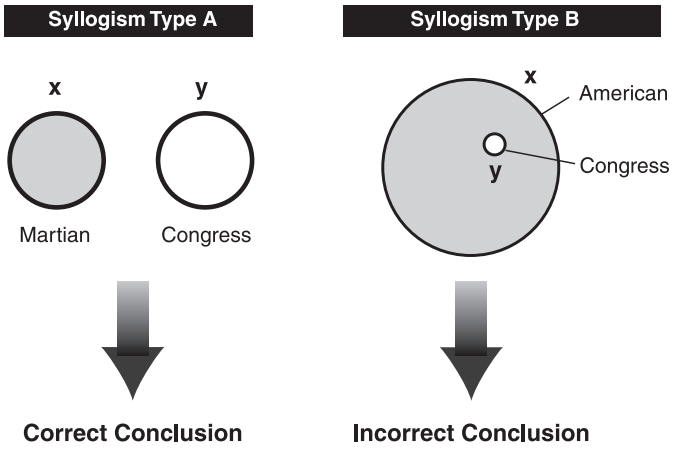
> If a person is an American, then he is probably not a member of Congress.
>
> This person is a member of Congress.
>
> Therefore, he is probably not an American. (Pollard and Richardson 1987).

This is the same syllogistic construct as:

> If $\boldsymbol{H}_0$ is true, then this result (statistical significance) would probably not occur.
>
> This result has occurred.
>
> Therefore, $\boldsymbol{H}_0$ is probably not true and so is rejected.

The difference in these syllogistic constructs is illustrated in Fig. 1. Syllogism Type A leads to a correct conclusion because the two parts of the initial premise are separate sets and do not overlap. Syllogism Type B leads to an incorrect conclusion because the two sets of the initial premise *do* overlap (one is the subset of the other), therefore the denial of the consequent

Fig. 1. Venn diagram of a correct (Type A) and incorrect (Type B) *modus tollens*.



does not equate to the denial of the antecedent. This "illusion of attaining improbability" is used in countless ecological journal articles; is the basis for the testing criteria in the U.S. Army Corps of Engineers and EPA's "Green book" (USEPA/USACE 1991) and "Inland testing manual" (USEPA/USACE 1998) for dredged material acceptability; and is also explicitly stated in many statistics textbooks. While this faulty reasoning is the underlying foundation for many a misinterpretation of the meaning of statistical significance, it unfortunately is only one of four major problems.

Problem 2: The "odds against chance" fantasy

Another common misinterpretation of the *p* value is seeing it as the probability that the research results were due to or caused by chance (Carver 1978). As mentioned previously, the *p* value is the probability of getting the research results when it is based on a probability of 1.00 that chance *did* cause the mean difference; these are not equivalent ideas. In a two-sample *t* test, a *p* value of 0.05 means that the odds are 1 in 20 of getting a mean difference this large or larger, and the odds are 19 in 20 of getting a mean difference this large or smaller *if and only if* the two samples are from the same population. We do not know how to estimate the odds that the null hypothesis is true, i.e., that the two samples *are* from the same population; the *p* values presented in statistical tables (e.g., Rohlf and Sokal 1969) are calculated based on a probability of 1.00 that the null hypothesis *is* true.

This is rather absurd if one takes the time to step back and think about it, because most researchers design experiments (monitoring programs, etc.) or use statistical tests to prove their research hypothesis, not the null hypothesis. As Edwards et al. (1963) point out, "in typical applications, one of the hypotheses — the null hypothesis — is known by all concerned to be false from the outset" (p. 214). This is one way of stating the "odds against chance" fantasy.

Cronbach and Snow (1977) present this odds against chance fantasy in terms of probability statements (which is a more useful framework for both discussing statistical results and viewing environmental data, and one that needs to become part of the normal landscape for environmental scientists):

> A *p* value reached by classical methods is not a summary of the data. Nor does the *p* value attached to a result tell how strong or dependable the particular result is... Writers and readers are all too likely to read 0.05 as $p(\boldsymbol{H} \mid \boldsymbol{E})$, "the probability that the ***H***ypothesis is true, given the ***E***vidence." As textbooks on statistics reiterate almost in vain, *p* is $(\boldsymbol{E} \mid \boldsymbol{H})$, the probability that this ***E***vidence would arise if the [null] hypothesis is true. Only Bayesian statistics yield statements about $p(\boldsymbol{H} \mid \boldsymbol{E})$ (p. 52).

This is without a doubt the most important and least understood principle of statistical significance testing. Carver (1978) used a wonderful example to drive this point home, which I want to repeat here:

> What is the probability of obtaining a dead person (label this part **D**) given that the person was hanged (label this part **H**); that is, in symbol form, what is $p(\mathbf{D}|\mathbf{H})$? Obviously, it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has been hanged (**H**) given that the person is dead (**D**); that is, what is $p(\mathbf{H}|\mathbf{D})$? This time the probability will undoubtedly be very low, perhaps 0.01 or lower. No one would be likely to make the mistake of substituting the first estimate (0.97) for the second (0.01); that is, to accept 0.97 as the probability that a person has been hanged given that the person is dead. Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with interpretations of statistical significance testing — by analogy, calculated estimates of $p(\mathbf{H}|\mathbf{D})$ are interpreted as if they were estimates of $p(\mathbf{D}|\mathbf{H})$, when they are clearly not the same.
>
> In statistical significance testing, we ask: what is the probability of obtaining a large mean difference (label this $\boldsymbol{D}'$) between two samples, if the two samples were obtained from the same populations (label this $\boldsymbol{H}_0$, the usual symbol for the null hypothesis); that is, what is $p(\boldsymbol{D}' \mid \boldsymbol{H}_0)$?... If we reverse the question to ask what is the probability that the two obtained groups were sampled from the same populations, we have the question that most people want to answer and assume they have answered when they calculate the *p* value from statistical significance testing. In essence, they are asking what the probability is that the null hypothesis, $\boldsymbol{H}_0$, is true, given the type of large mean difference we have obtained, or, what is $p(\boldsymbol{H}_0 \mid \boldsymbol{D}')$?

> The p value that was obtained from statistical significance testing, for example, $p(\boldsymbol{D}'|\ \boldsymbol{H}_0) = 0.05$, is used as an answer to the reverse question as well. This is a fantasy, however, because the p value that results from statistical significance testing is $p(\boldsymbol{D}'|\ \boldsymbol{H}_0)$, not $p(\boldsymbol{H}_0\ |\boldsymbol{D}')$ (pp. 384–385).

Going back to the four statements posed in the beginning of this section under Problem 1, the "odds against chance" fantasy is the belief that Statement 3 ≡ Statement 4, a mistake that readers of this article hopefully will no longer make.

Problem 3. The fantasy of p as an indicator of research hypothesis validity

To twist Carver's (1978) above statement a little further, I would contend that most investigators interpret the p value in statistical significance testing as an indicator that their research hypothesis ($\boldsymbol{H}_1$) is true; instead of making the mistake of thinking $p(\boldsymbol{H}_0\ |\boldsymbol{D}')$ is equivalent to $p(\boldsymbol{D}'|\ \boldsymbol{H}_0)$, they interpret $p(\boldsymbol{D}'|\ \boldsymbol{H}_0)$ as equivalent to $p(\boldsymbol{H}_1\ |\boldsymbol{D}')$ (again, referring back to the four initial statements in problem one, the mistake under this fantasy is that Statement 1 ≡ Statement 4). Those who succumb to this mistake also are likely to interpret the size of the p value as a reflection of the degree of the validity of the research hypothesis (or, the lower the p value, the "more highly significant" or "more valid" the research hypothesis is).

This point was illustrated to me a few years ago during an in-house seminar when my colleagues and I were discussing the findings from one of our monitoring cruises to study the impacts of open-water dredged material disposal. One of the staff scientists was presenting the results of sediment contaminant measurements obtained at a disposal mound with those at the reference stations; the results of the statistical test that was performed showed that one of the contaminants of interest was significantly different but not in the expected direction (i.e., the reference areas had higher concentrations than the disposal mound). Someone then asked, "Well, what was the p value? How significant was it? Was it just near 0.05 or was it really significant?" His comment reflected his interpretation that the size of the p value indicated the degree of validity of the research hypothesis, i.e., the lower the p value, such as $p \leq 0.001$, the more highly significant or valid the research hypothesis. At this point, I hope it is painfully clear that statistical significance does **not** reflect anything directly about the validity of the research hypothesis. Even if the null hypothesis can be rejected (and it always can, if a sufficient number of samples are taken), there is almost always more than one alternative hypotheses to be evaluated (who knows, one of them may even be your research hypothesis). Only after a thorough design and very rigorous theorizing (e.g., Platt 1964; Germano et al. 1994) along with multiple replications of the measurements in a variety of different settings can one say anything about the probability of the research hypothesis being true.

Problem 4. The fantasy of replicability or reliability

Similar to general theories about disturbance or succession in ecology or the relationship of sediment contaminant concentrations to organic carbon content (for hydrocarbons) or acid-volatile sulphides (for metals) in toxicology, most theories in psychology predict no more than the direction of a correlation, group difference, or treatment effect. The final major problem with the misinterpretation of NHST is another common misconception that the complement of the p value gives some indication of the replicability ($\boldsymbol{R}$) of the results, i.e., if $p = 0.05$, then $1 - p$ or 0.95 is the probability that the same mean difference would be repeated should the measurements or experiment be repeated. Again, the mistake here is thinking $p(\boldsymbol{R}\ |\boldsymbol{D}')$, the probability that the results are replicable or reliable, is equivalent to $p(\boldsymbol{D}'|\ \boldsymbol{H}_0)$. Bakan (1966) noted this error almost 30 years ago, and Lykken (1968) elaborated much further on this idea two years later when he pointed out that if statistical significance is used as a demonstration of "some empirical fact," then associated with this is a claim or confidence in the replicability of one's findings. Carver (1978) points out that there is nothing in the logic of statistics to allow a statistically significant result to be interpreted as any reflection of the probability of the result being replicated.

Lykken (1968) presents a detailed discussion of the relationship of statistical significance to the probability of a "successful" replication by distinguishing between three rather different methods of replication (literal, operational, and constructive replication). Literal replication is just what it sounds like, an *exact* duplication of everything, which realistically rarely happens; the closest one can come is asking the original investigator to simply run more subjects. Operational replication involves another investigator using the same "experimental recipe" by using conditions and procedures published in the "Methods" section of a journal article to see if similar results can be obtained. Constructive replication is where one deliberately avoids imitation of the first author's methods and has nothing more than a clear statement of the empirical "fact" that the first author claims to have established; the second investigator formulates their own method of sampling, measurement, data analysis, etc. to confirm this previously established "fact." The probability of replication decreases drastically as one goes from literal to operational to constructive replication; Lykken points out that because the null hypothesis is never strictly true, predictions about the direction of a correlation or treatment effect (e.g., the contaminant levels in the sediment will cause mortality or "adverse effects" on the population in the natural environment) have about a 50–50 chance of being confirmed even if our original theory is false, because statistical significance is merely a function of sample size.

Tversky and Kahneman (1971) classify this fantasy as a distortion on the law of large numbers (that very large samples will indeed be highly representative of the population from which they are drawn) and relabel it as a misappropriated belief in a supposed "law of small numbers." They contend (and my experience after 20 years in applied monitoring also confirms) that most investigators view a sample randomly drawn from a pop-

ulation as highly representative (i.e., similar to the population in *all* essential characteristics), so that any two samples drawn from a particular population would be more similar to one another (and to the population) than sampling theory predicts for small numbers. As an example, Tversky and Kahneman (1971) cite the results of a questionnaire they distributed at a meeting of the Mathematical Psychology Group of the American Psychological Association where they have described the findings of an experiment with 20 subjects having a significant result ($p < 0.05$) "which confirms your theory" (a statement which I hope the reader by now would recognize as more than a little misleading). They asked the respondents what the probability would be of getting significant results if another 10 subjects were run; most people thought the probability is somewhere around 0.85, when in fact it is around 0.48. Their basic premise is that most people have strong intuitions about random sampling, most of which are wrong (e.g., the common "gambler's fallacy" that deviations in one direction from the expected 0.5 of a random coin toss will soon be canceled by a corresponding deviation in the other direction), and that these are applied with unfortunate consequences to scientific investigations.

It is important to remember that the laws of chance are not an active, self-correcting process; deviations from a predicted outcome such as the expected 0.5 probability of getting a "heads" in a coin toss are not canceled as sampling proceeds, they are merely diluted. Tversky and Kahneman (1971) drive this home by the following illustrative example: the mean IQ of a population of eighth graders in a city is *known* to be 100; you select a random sample of 50 children, and the first child tested has an IQ of 150. What do you expect the mean IQ to be for the sample? A surprisingly large number of people believe that the expected IQ for the sample will be 100, based on the misbelief that a random process is self-correcting, when the correct answer is 101.

Tversky and Kahneman (1971) characterize the believers in the law of small numbers as investigators who

- gamble their research hypotheses on small samples without realizing the odds against them are unreasonably high; they overestimate power.
- have undue confidence in early trends and the stability of observed patterns (e.g., the number and identity of significant results); they overestimate significance.
- have unreasonably high expectations about the replicability of significant results; they underestimate the breadth of confidence intervals.
- rarely attribute a deviation of results from expectations to sampling variability, because they will find a causal "explanation" for any discrepancy; they rarely recognize sampling variation in action (Tversky and Kahneman 1971, p. 109).

Hence, their belief in the law of small numbers will always remain intact and continue to be the distorted lens through which they see the implications of their experimental results. A true believer, as Tversky and Kahneman (1971) point out,

> commits his multitude of sins against the logic of statistical inference in good faith...Thus, while the hasty rejection of the null hypothesis is gratifying, the rejection of a cherished hypothesis is aggravating, the true believer is subject to both. His intuitive expectations are governed by a consistent misperception of the world rather than by opportunistic wishful thinking (p. 110).

It is a fantasy to believe that statistical significance reflects anything about the degree of confidence in the replicability or reliability of results, i.e., that $p(\boldsymbol{R} \mid \boldsymbol{D}')$ is equivalent to $p(\boldsymbol{D}' \mid \boldsymbol{H}_0)$.

Part 1: Summary

The fantasies about statistical significance or obtaining a p value of 0.05 or less fall into four categories.

1. The illusion of attaining improbability, or thinking that denying a correct, probabilistic initial premise will result in a sensible conclusion.
2. The odds-against-chance fantasy, thinking that the p value is the probability that the results were caused by chance or that $1 - p$ represents the probability that the results were not caused by chance.
3. The fantasy of research hypothesis validity, that obtaining a p value of 0.05 or less says something about the research hypothesis instead of something about the rareness of the data given that the null hypothesis is true.
4. The fantasy of replicability, that obtaining a p value of 0.05 means that we can be 95% confident that the results are "reliable" or that the probability is 0.95 that the results will replicate.

Properly interpreted, the results of NHST give a p value that reflects the probability of obtaining mean differences of a given size under the null hypothesis (assuming it is true); the p value may be used to make a decision about accepting or rejecting the idea that chance caused the results. This is what statistical significance testing is — nothing more, nothing less (Carver 1978).

A wide variety of investigators (e.g., Edwards et al. 1963; Bakan 1966; Morrison and Henkel 1970; Cohen 1994) have emphasized repeatedly the unlikelihood of a null hypothesis ever being true in any population in nature. Why would we expect sediment contaminant levels at a dredged material disposal site to be the same as at the reference area any more than an educational researcher would expect all scores on a reading test from a fourth grade class in an economically depressed inner-city school to be the same as those from a fourth grade class in an economically privileged suburban Montessori school? Why should any correlation coefficient be *exactly* 0.00

in any natural population (the assumption on which all *p* values are calculated in statistical tables)? As Bakan (1966) pointed out, a quick glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature. In fact, in more cases than not, the issue that environmental investigators should be concerned with is not a Type I (α; false positive) but a Type II (β; false negative) error. If the null hypothesis is ever true, the probability of a statistical conclusion error is held to 5% by the convention of $\alpha = 0.05$. However, when the null hypothesis is false, the probability of error is β, and β can be quite large. It is sobering to realize that the probability of an erroneous conclusion in a statistical analysis is *not* necessarily limited to 0.05, but may easily range as high as 0.85 or more (Cohen 1962; Lipsey 1990; Germano 1991). Most investigators fail to keep in mind that the total probability of error in any experimental study is *either* α or β, *not* both or some combination of the two (Lipsey 1990). Low power is an all too common feature of many ecological and sediment and (or) water bioassay test results (Crane and Newman 1996; Forbes and Forbes 1994; Toft and Shea 1983), and the potential costs of Type II errors in these results on which environmental regulations are based can be substantially more serious than those from Type I errors (M'Gonigle et al. 1994).

Neyman and Pearson (1933) categorized the rejection of the null hypothesis in their seminal theoretical work as a function of five factors:

1. whether the test is one- or two-tailed (choice of investigator)
2. the standard deviation (a given of any particular situation)
3. the level of significance (choice of investigator, but a social norm of 0.05 is common in most scientific fields)
4. the amount of deviation from the null hypothesis (this is always unknown, but no matter how small, most likely always exists)
5. the number of observations (choice of investigator).

It is the dependency on this last factor that is the ultimate weak link; as Nunnally (1960) put it so aptly:

> ...if the null hypothesis is not rejected, it is usually because *N* is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data. (p. 643).

Berkson (1938) made the same observation almost 60 years ago when he stated,

> ...we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the *P* that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all (p. 527).

Hays (1963) emphasized the same point by stating, "Virtually any study can be made to show significant results if one uses enough subjects regardless of how nonsensical the content may be" (p. 326). Because one can always guarantee statistical significance by having a sample size large enough (regardless of the common misinterpretations of what the *p* value actually means), this is the ultimate irony for the unaware researcher (consultant, lawyer, regulator, etc.). Tyler (1931) made the important point that a statistically significant difference is not necessarily an important difference (a small mean difference from a research standpoint can be made statistically significant just by taking enough samples), and a difference that is not statistically significant may be an ecologically or scientifically important difference. However, witness our current practice based on our inability to interpret contaminant levels in invertebrate tissues; the common practice is to take samples from our area of concern (impact site) and at a "reference" area(s) and statistically compare the differences (often with no regard to the power of the statistical test, given the expense of individual sample analysis). If the difference is "statistically significant," then it is deemed that an adverse environmental impact has occurred.

When trying to reflect why we as scientists are so willingly blinded by what statistical significance testing can and cannot do, I feel that Bakan (1966) was correct by pointing out that making inductive generalizations is always risky (imagine the embarrassment and ridicule if we are proven wrong!). The reasons psychologists (and I think the same is true for ecologists) latched onto statistical significance testing so strongly are three-fold: because of the frustrations of dealing with such a complex subject (Nunnally 1960), the unconscious "physics-envy" pointed out in the beginning of this paper, and the apparent result of what "running the tests" that all of us were taught in our introductory cookbook statistics courses would provide, i.e., we would no longer have to engage in the dangerous business of making inferences ourselves, but we would let the statistical tests be our analytic analogues of inductive inference. By doing this, it would remove the burden of responsibility and the chance of being wrong from the shoulders of the investigator and place it on the test of significance. Two important goals are achieved by placing the contingency of conclusion on the $p < 0.05$ level of statistical significance.

1. There is an implied social agreement that 5% is good (e.g., see Cowles and Davis 1982), and 1% is even better.
2. By not making individual decisions about the level of significance, the investigator only has to report the *p* value as a "result" (and therefore a presumably "objective" measure of the degree of confidence in these results) (Bakan 1966).

Of course, the minor side issue as Bakan (1966) also pointed out is the stark reality that getting a statistically significant result is wholly contingent upon the number of observations; this has always been handled largely by ignoring this one crucial fact, because it does have the disconcerting effect of trivializing the results of most studies.

It is more than sobering for us as ecologists to step back and reflect on the tortuous tangents and side paths that misunderstandings about NHST have led us over during the last four decades; while we can take comfort that we are not alone in our mistakes (indeed, the bulk of psychological and social science research show as much blind faith in ritualistic worship in the temple of $p < 0.05$), it is no excuse to continue the bad practice. Testing for statistical significance imparts an illusion of objectivity that simply does not exist; we would gain more insights from our data by abandoning the practice altogether, because the disadvantages consistently outweigh the advantages. Howson and Urbach (1991) pulled no punches when they stated, "Corroborating a hypothesis does not strengthen it, a significant result has no significance for the truth of the null hypothesis, and a 95 per cent confidence interval has no right to impart confidence, let alone 95 per cent's worth, to an estimate" (p. 373). Given our lack of predictive insight in ecology after more than four decades of repeated reliance on statistical significance, I find it hard to advocate continued use of techniques that give the impression of "hard science" in action but cannot really help investigators convert the ever-burgeoning amount of data that is accumulating into information or insight. There are alternative avenues available that can offer us new insights and provide a synthetic framework to allow accurate diagnosis of ecosystem health, validate predictive models, or assess the effectiveness of remediation measures; these will be presented in the sections that follow.

Part 2: The difficult task of accurate diagnosis of ecosystem health

As ecologists or consultants, scientists are often called upon to diagnose conditions of ecosystem health (e.g., Environmental Impact Report/Environmental Impact Assessment (EIR/EIA)) or make predictions about planned anthropogenic impacts (e.g., the Environmental Impact Statement (EIS) merry-go-round) or remediation effectiveness (Remedial Investigation/Feasibility Study (RI/FS)). To do this, they perform two basic functions: specify and (or) collect what they feel (or regulations stipulate) are the required data, and then interpret these data. There are a host of excellent references to help with the former (e.g., Green 1979, 1984; Eberhardt and Thomas 1991; Rose and Smith 1992); my interest is the diagnosis or interpretive function. It is in this area of ecosystem health "diagnosis" where we can benefit most from the experience gained in the fields of medical epidemiology and clinical psychology; if ecologists pay attention to the lessons learned in the areas of decision making and expert judgment in these fields, we can save ourselves a few decades of reinventing the wheel or from making the same mistakes they have committed.

Fig. 2. Prototype diagnostic model (human medical analogues in parentheses).

		ENVIRONMENTAL CONDITION (Outcome)	
MEASURED		X	Not X
VARIABLE	Present	**A**	**B**
(Symptom S)	Absent	**C**	**D**

Bernstein's (1990) quote of von Forster's comment is particularly relevant here: the "hard" sciences deal with the soft or easy problems, and the "softer" sciences, such as ecology, with the truly hard problems. As with a psychiatric or medical diagnosis, an ecologist is faced with a wealth of potential information about their patient or ecosystem. It is the job of the "expert" to decide what information is the most relevant, how to obtain it, how to integrate what is obtained, and how to relate it to what are often nebulous and ill-defined categories (Faust 1986*b*). As with the first section of this paper, I would refer the reader to the source articles on this subject for more details (e.g., Meehl 1954; Wiggins 1973, 1981; Arkes 1981; Faust 1984, 1986*a*, 1986*b*, 1989; Dawes et al. 1989, and references contained therein); once again, I will just present an overview of some of the more than sobering highlights from these studies.

Problem 1: The misestimation of covariation

The first and probably the most serious impediment for a marine scientist or ecologist in achieving a high diagnostic accuracy about ecosystem "health" is the inability to assess covariation accurately (Arkes 1981). The prototype diagnostic situation for a medical health model shown in Fig. 2 also is equally relevant for assessing any predictor of environmental disease or health (or the validity of an ecological indicator).

This 2×2 contingency matrix depicts the prototype situation for both a clinician assessing the relevance of a symptom for a predicted outcome and for a scientist exploring whether or not their new measured variable is a useful predictor of an environmental condition. The clinical psychologist, after testing a patient, may ask, "Is this profile on the Minnesota Multiphasic Personality Inventory diagnostic of an impending psychotic break?" just as the environmental scientist or policy regulator may ask after testing a particular sediment sample, "Does this amphipod bioassay validly predict adverse toxicity or harm to the community in the field?" (e.g., Spies 1989) or, to hit close to home, "Does the presence of this feeding void in a sediment profile image accurately predict the presence of a deposit feeding community?" (Rhoads and Germano 1982, 1986).

The transition from the medical model to the ecological model is quite easy: a clinician will note that some people have Symptom S while others do not. The clinician then attempts to determine if the presence of that particular symptom is diagnostic of some disease or future outcome. Past research in the

Fig. 3. Hypothetical data set from REMOTS® surveys.

IMAGED		HYPOXIC OR ANOXIC WATERS YES	NO
METHANE IN	Present	12	6
SEDIMENT	Absent	4	2

field of psychology (e.g., Smedslund 1963; Arkes and Harkness 1980; Arkes 1981) has shown that most people base their assessment of covariation largely on the number of instances in Cell A; when I thought about some of the conclusions I have made in the past about ecosystem function or about some of the things I have read in the literature, I realized that ecologists are no different from medical clinicians. I will use a variation on an example Arkes (1981) presented to illustrate this point. I have given talks on this topic at a variety of meetings and presented the following hypothetical situation to a number of audiences using a variable I am used to working with in sediment profile images: the presence of methane gas bubbles at depth in the sediment (these are readily visible in sediment profile images). The build-up of methane gas in the sediment is typically a result of excess organic loading and anaerobic decomposition (Rhoads and Germano 1982). I would like to see if this is a useful parameter for predicting whether or not an area will develop hypoxic or anoxic waters. After completing 24 surveys, the data sort out as illustrated in Fig. 3.

After presenting these data to an audience, I ask them to estimate the relationship between the measured variable (methane in the sediment) and predicted outcome (development of hypoxic or anoxic waters) on a contingency scale from 0 to 100, where 0 is no relation, 100 equals a complete relation. I give them the following four choices as interval ranges for the relationship and by a show of hands indicate which one they would choose:

1. 0 – 2
2. 2 – 20
3. 20 – 60
4. 60 – 100

When Arkes (1981) presented similar data to subjects and asked them to estimate a number, the mean was 64; whenever I have posed this situation to audiences, the majority of hands have always gone up on choice #3, with choice #4 a close second. Occasionally a few brave souls have chosen #2, and once a single individual chose #1. People are often surprised to find out that the actual relationship between this variable and the outcome given these data is zero, because the outcome (hypoxia) is twice as likely to occur regardless of whether or not the "symptom" (methane) is present. People are consistently fooled by the large magnitude in Cell A, and this produces a badly biased estimate of contingency.

If you just think for a moment how one might investigate a correlation of a supposed predictive variable, you will gain an instant appreciation of how unimportant Cells C and D (Fig. 2) seem to be. You are studying a few embayments where algal blooms occur every summer, and you suspect that a certain threshold level of nitrogen flux from the sediment may be diagnostic of the impending bloom. To verify this, you keep track of the nitrogen flux on a weekly basis in each of these embayments to see when the bloom occurs. Would you also consider keeping track of an equal number of embayments where algal blooms never occur or where extremely low flux levels of nitrogen have been reported? Consideration of the information in Cells C and D is mandatory for a scientist or clinician to determine if a measured variable and condition (or symptom and disease if you prefer the medical analogy) are related. Yet, these are the measurements that are often never made, or, if they are made, the results that are never reported (or get rejected for publication).

Problem 2: The bias of preconceived notions

This can also be classified as overreliance on confirmatory strategies. A classic study by Chapman and Chapman (1967) dramatically illustrated this problem in the field of psychology: drawings were randomly paired with personality traits presumably characteristic of the person who did the drawings. Clinicians, when asked to evaluate the drawings, fabricated illusory correlations between drawing features and personality traits. Prior associations by the evaluators warped the perception of incoming data, so much so that even when the true relation between a drawing feature and a trait was negative, it was seen and tallied as positive. A prior association or bias not only warps the perception of an existing correlation, it impedes the accurate processing of individual data (Arkes 1981). When researchers set out to test their hypotheses, more often than not they seek confirmatory evidence exclusively (i.e., only examples in Cell A in Fig. 2), and when disconfirmatory evidence is obtained, it is underweighted or dismissed (Faust 1986*b*). This point has been humorously illustrated in many of the fictional Sherlock Holmes movies where this bias is exaggerated to the hilt by the character of Inspector Lestrade. Lestrade will walk in on a murder scene, make an instant judgment about who the killer is, and then explain how every clue he subsequently finds supports his initial theory. This tendency to search for confirmatory evidence exclusively has already been pointed out as a shortcoming in ecology by Loehle (1987) and is what Rousseau (1992) has called "pathological science."

When an ecologist has an idea that a certain species is a good pollution indicator or that certain levels of a measured water column parameter are diagnostic of impending hypoxia, this "prior association" will be the lens through which they view all incoming data, even to the point where data from known polluted or hypoxic areas will produce the perception of a positive correlation when the true relation between the variable and the outcome was negative. If data do not agree with an investigator's preconceived notions, their natural instinct is to either ignore the data or, if recognized, dismiss its importance. After completing a study, it is always an instructive exercise to use the same data set to try to support exactly

the opposite of your initial hypotheses.

Investigators in the field of expert judgment have documented in several studies that all subsequent data gleaned after an initial tentative diagnosis is formed will be biased by those initial opinions; data consistent with the tentative diagnosis will be given added credence, while those inconsistent with the hypothesis will be disregarded (e.g., Chapman and Chapman 1967, 1969; Tversky and Kahneman 1974; Shweder 1977; Lord et al. 1979; Nisbett and Ross 1980). Unfortunately, even admitting to yourself that an initial opinion or hypothesis is merely tentative does not decrease its biasing influence. (Ross et al. 1977; Arkes 1981). The positive reinforcement of finding *any* evidence to put a tick mark in Cell A in Fig. 2 is so powerful that investigators will cling to the memory of any occurrence in Cell A while dismissing or not remembering data in Cells B or C (Lord et al. 1979). Marie Von Ebner-Eschenbach once quipped, "Even a stopped clock is right twice every day. After some years, it can boast of a long series of successes." The real problem with ecosystems, as with human personalities and behavior, is that there is a rich source of varied data that will allow any investigator to find some confirmatory evidence for almost any hypothesis, regardless of its validity (Faust 1986*b*); given enough data, all but the most outlandish diagnosis or hypothesis can appear obvious (Arkes 1981).

Problem 3: Lack of awareness and overconfidence

These two bad habits are the Scylla and Charybdis of the research scientist, senior consultant, or expert scientific witness. Work from a variety of investigators (e.g., Nisbett and Wilson 1977; Summers et al. 1969; Oskamp 1967) indicates we have negligible awareness of the factors that influence our judgment or final decision. Their studies demonstrate that when practitioners are presented with a variety of symptoms in a patient, the clinicians are unaware of the impact each symptom had on their final diagnosis. Sampling any number of reports in the "grey literature" about synthetic measures of sediment "quality" such as the sediment quality triad (Chapman et al. 1987) or the apparent effects threshold (AET) that has taken the regulatory community in the Pacific Northwest by storm (PTI 1988) will show more often than not that one or more of the variables contributing to the final index score are either equivocal or may be the key variable supporting the final conclusion about potential harm of chemical contamination. I cannot find any good reason to suspect why marine ecologists are any different than medical clinicians in this respect; I am sure we would demonstrate the same inability to explain accurately which data really influenced our final interpretation of ecosystem health or stress (was it the benthic community data, the presence of a particular species, the diversity index, the level of sediment contamination, or the bioassay results?).

Another factor that impedes accurate assessment is sometimes the serious overconfidence that scientists have in their final interpretation (often the same as their initial diagnosis). Numerous studies (e.g., Oskamp 1965; Dawes et al. 1989) have shown that providing an investigator with more information increases their confidence in their interpretation without necessarily increasing the accuracy of their judgement (in other words, selectively filtering the data to confirm a preconceived hypothesis; the initial diagnosis becomes a self-fulfilling prophesy). Even more discouraging was to read about studies that showed the most confident diagnosticians tended to be the least accurate (remember this next time you experience vague or contradictory symptoms and consult a doctor). Unfortunately, it is a basic human trait to disregard evidence that contradicts your current judgment; with selective filtering of confirmatory evidence and censorship of nonconfirmatory evidence, many hypotheses cannot fail to be well substantiated.

Problem 4: Disregard or underuse of base rates

This appears to be one area where medical epidemiologists and clinical psychologists are light years ahead of marine ecologists in terms of awareness of how this can affect the diagnostic validity of any predictive variables. The value of *any* predictive diagnostic test (e.g., Acid Volatile Sulfides/Simultaneously Extracted Metals (AVS/SEM), AET, Effects Range Low/Effects Range Medium (ERL/ERM), standard acute toxicity bioassays, etc.) cannot be determined without at least knowing the base rate(s) for both the outcome(s) ("disease" in the medical model) to be identified *and* for both the false negative and false positive identifications (i.e., Cells B and C in Fig. 2). The tendency to overlook or underuse base rates is common even in the fields of medicine and psychology (e.g., Faust 1986*b*), so its complete absence in ecology and ecological risk assessment is not all that surprising. The clinical literature is full of convincing examples that point out the pitfalls of disregarding or not using base rates (e.g., Kahneman and Tversky 1973; Nisbett et al. 1976; Tversky and Kahneman 1978; Arkes 1981; Faust 1986*a*). Because this is usually a completely neglected aspect in marine ecological data interpretation, I will present two examples adapted from Arkes (1981) and Faust (1986*a*) that illustrate the seriousness of ignoring this aspect:

Example 1: The case of recurring algal blooms

You work for a large, international environmental consulting firm that specializes in ecological risk assessment and invests quite heavily in R&D for monitoring instrumentation. Your company has just developed a proprietary remote sensing technique that will predict the occurrence of algal blooms in response to eutrophic waters. Your assessment technique has the following characteristics:

- The technique gives a positive prediction in 95 out of 100 areas that do go eutrophic and develop algal blooms;
- The technique gives a negative prediction for 95 out of 100 areas that never develop algal blooms.

It is fair to say that if any company did develop such a tool, they would be able to make a mint in the consulting field; their investigators would (justifiably) feel extremely confident about the diagnostic power of the method (after all, it has only a 5% error rate for false positives and false negatives).

Now, for the past 5 years, there have been recurring problems with eutrophication and algal blooms destroying the tourist industry in the Mediterranean. Each year, an average of five embayments that were previously undetected out of 1000 along a coastline will go eutrophic. You have just been awarded a major contract by the government of Italy to survey an area of coastline that is critical for its economic value in attracting tourists to its beaches. If this proprietary assessment technique is used to survey a randomly selected embayment along this section of coastline and the prediction is positive, what is the probability that the area will experience an algal bloom during the next summer? Should you feel confident in recommending that your client spend additional funds to initiate any mitigating resource management actions right now as a result of your conclusions? The answer to this last question is no; the reason for this surprising conclusion is presented in the next section of the paper.

Example 2: The case of flounder liver lesions

Suppose the base rate of flounder liver lesions from sediment contamination in a population is 1 per 100 000. Suppose Toxicologist Jones just plays the base rates and always concludes the condition is never present without spending the money to do any histopathological analyses. However, Toxicologist Smith has jumped on the biomarker bandwagon and is willing to diagnose the condition. For argument's sake, let us say Smith never misses a true case of sediment-induced flounder liver lesions (i.e., he never gets a false negative); however, he does make a false positive identification for 1 in 1000 cases (i.e., he says the condition is present when it is not).

We should recognize before proceeding further that we have endowed Toxicologist Smith with remarkable diagnostic powers that exceed most analytical chemistry or toxicology laboratories (i.e., a 0% false negative and a 0.1% false positive rate). An extensive regional sampling survey is done (similar to EPA's EMAP program) and the results are applied to the 100 000 flounders that were caught. Toxicologist Jones misses the 1 case when the condition is actually present (a false negative error), and Smith misses the 0.1% $\times$ 99 999 cases in which the condition is not present, or makes about 100 false positive errors. Therefore, Smith is wrong 100 times more often than Jones. While the example is extreme, one of the points that Faust has made repeatedly in his various publications (1984, 1986*a*, 1986*b*, 1989) cannot be ignored: **unless a test can surpass the diagnostic hit rate achieved by the base rates alone, it will decrease instead of increase diagnostic accuracy**. Should we use sediment quality criteria that correctly predicts toxicity 90% of the time? The answer may seem obvious, but it cannot be determined unless one knows the base rates in the population of interest. Again, this is clear in the extreme case: the test will not help if the base rate is 100%.

Problem 5: Hindsight bias

Given a rich source of data (regardless of whether it is medical symptoms or environmental variables), almost any diagnosis can be supported unless it is truly outrageous. This is what Arkes (1981) termed "hindsight bias," a phenomenon initially documented by Fischhoff (1975). Fischhoff asked three groups of clinicians to read psychotherapy case histories and then judge the likelihood of four possible circumstances that may have followed therapy. One group was not told the outcome (foresight group), one group was told Outcome A would occur (hindsight group), another that Outcome B would occur (hindsight group). Of the two hindsight groups, those that were told in advance that Outcome A had occurred assigned probabilities to Outcome A that were 49% higher than the foresight group; the other hindsight group also predicted that Outcome B would have been easy for them to predict (the foresight group did not consider B to be a likely outcome). Arkes et al. (1986) found the same effect with medical diagnoses (presumably where the symptom–disease relationship is more exact than in psychology); a foresight group was shown an actual case history and asked to assign probabilities to four possible outcomes. Each of four hindsight groups were told that a different one of the four diagnoses were true; once again, the hindsight bias emerged and each of the hindsight groups assigned higher probabilities to the particular outcome they were told than the foresight group.

Given enough data, many diagnoses (or environmental impact conclusions) can appear obvious. Give the same benthic community data set to ten different investigators, and I feel confident you will get a wide range of final judgments about how "stressed" an area is. I have noticed the same phenomenon when training students on how to interpret results from a sediment profile camera survey; very often, I will give them a set of sediment profile photographs for which all the measurements are completed and the final report written. Their final interpretation of what is going on in an area is very different if they have read the report before they start work instead of doing the work "blind" and then comparing their results to the report. This problem is going to be even *more* acute as more ecologists and environmental scientists are involved in multidisciplinary studies that collect a wide variety of biological, physical, and chemical parameters — given enough data, any one of several conclusions can appear to be obvious when read in hindsight.

Part 2: Summary

The most common impediments that ecologists will face in finding valid predictors or ecological indicators to accurately diagnose ecosystem health are the same ones clinicians face in diagnosing medical or mental health in humans:

1. misestimation of covariation,
2. the bias of preconceived notions or initial hypotheses,
3. lack of awareness and overconfidence,

4. disregard or underuse of base rates,

5. hindsight bias.

One of the main weaknesses with the derivation of ERL/ERM's (Long and Morgan 1990; Long et al. 1998) as predictors of sediment quality is that only the "hit" data (sediments which demonstrate an effect in the paired bioassay test) are incorporated to derive the final sediment quality values (i.e., all the data from Cells B and D in Fig. 2 are discarded from the original data set before these sediment quality "predictors" are calculated). Because most researchers and investigators in ecology are unaware of the impacts of the above problems on the strength of the conclusions they reach, we still have a long way to go just to heighten awareness in the minds of both the "experts" who interpret the data and the regulators who enforce the implications of their own interpretations of these data. The evidence for the utility of the expert scientist or consultant as an accurate diagnostician given these problems is not overwhelming; the judgment literature is rich on the underperformance of the clinical diagnostician (e.g., Meehl 1954, 1973; Einhorn and Hogarth 1981; Pitz and Sachs 1984; Faust 1984, 1986*a*, 1986*b*, 1989; Dawes et al. 1989), so we would be wise not to put too much faith in our individual integrative or synthetic ability. Fortunately, there are a variety of trails out of this maze that have been blazed by investigators outside of the field of ecology, and we would do well to take advantage of these not only to examine our existing techniques, but also to point the way for new directions in the future. This topic is examined in the next section of this paper.

Part 3: Suggestions for improving the ecologist's diagnostic accuracy and alternative approaches to ecological data analysis

Lest the reader is thoroughly depressed by this point with no hope for changing our current behavior as scientists, it is now time for the "good news." Once again taking inspiration from the psychoanalytic model, I will suggest a variety of alternatives for reframing the way ecologists can view environmental data and problems of ecosystem health; it is up to the individual investigator whether or not they find any of these ideas useful or to determine which would be most appropriate in their particular situation. As I stated at the outset, while my intent is to hold a critical mirror up to view the current status of our discipline, I do not presume to offer a panacea that will solve all the problems we face. However, these alternative debiasing techniques and statistical methods appear to me to offer a more enlightened approach to dealing with the wealth of environmental data that drown most regulatory agencies.

Suggestion 1: Include ignored information: think Bayesian

Probably one of the most promising avenues for increasing the impact of usually ignored information (i.e., assessing the effects of covariation accurately, or incorporating the information from Cells B and C in Fig. 2) and avoiding the pitfalls of NHST is the application of Bayesian statistics. Again, I will just present an overview of some of the major points and refer the interested reader to the host of journal articles or textbooks for more details (e.g., Schlaifer 1961; Edwards et al. 1963; Galen and Gambino 1975; Lusted 1968; Schwartz et al. 1973; Arkes 1981; Iversen 1984; Berger 1985; Lee 1989; Press 1989). While Bayesian statistics are more popular and better known in the realm of business statistics for performing cost/benefit analyses, they also provide the appropriate framework for dealing with medical, psychological, and ecological data for decision analysis and validating predictive models. While Bayesian applications are not new to fisheries science (e.g., Punt and Hilborn 1997), and articles have been written in the past extolling their utility for ecology (e.g., Reckhow 1990) or risk assessment (e.g., Hill 1996), they are the exception rather than the rule for study design or data interpretation. My hope is that by making the same point as some of these other investigators but in a slightly different format, more people will understand their advantages and start using them in their own work.

A Bayesian analysis will allow us to accurately assess how useful our proprietary remote sensing technique for algal blooms (presented in the last section) really is. Bayes' theorem is a simple and fundamental fact about probability applied to a field of statistics: probability is orderly opinion, and inference from data is nothing other than the revision of such opinion in the light of relevant new information (Edwards et al. 1963). The prior odds are expressed as a ratio of the likelihood of the hypothesis being true divided by the likelihood of the hypothesis being false, or $p(\boldsymbol{H})/p(\bar{\boldsymbol{H}})$; once new information or data are obtained, the prior odds must be modified in light of this new information (Arkes 1981). To do this, the prior odds are multiplied by the likelihood ratio, which is the probability of obtaining that piece of datum if the hypothesis were true divided by the probability of obtaining the datum if the hypothesis were false, or $p(\boldsymbol{D} \mid \boldsymbol{H})/p(\boldsymbol{D} \mid \bar{\boldsymbol{H}})$ (these terms should look familiar from the discussion in the first section of the paper on NHST). When the prior odds are multiplied by the likelihood ratio, we get the posterior odds. The posterior odds are the probability that the hypothesis is true given this piece of information divided by the probability that the hypothesis is not true given this piece of information, or $p(\boldsymbol{H} \mid \boldsymbol{D})/p(\bar{\boldsymbol{H}} \mid \boldsymbol{D})$. Recall from the initial section on NHST, we are finally dealing with the information we want $[p(\boldsymbol{H} \mid \boldsymbol{D})]$ instead of what classical Fisherian statistics provides us $[p(\boldsymbol{D} \mid \boldsymbol{H}_0)]$. So, the whole formula is

Posterior odds = Likelihood ratio × Prior odds,

or,

$$[1] \qquad \frac{p(\boldsymbol{H}|\boldsymbol{D})}{p(\bar{\boldsymbol{H}}|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|\boldsymbol{H})}{p(\boldsymbol{D}|\bar{\boldsymbol{H}})} \times \frac{p(\boldsymbol{H})}{p(\bar{\boldsymbol{H}})}$$

So, let us use this construct to examine the situation posed in the last section of the paper. The prior odds of having an algal

bloom are

$$[2] \quad \frac{p(\boldsymbol{H}_{ab})}{p(\bar{\boldsymbol{H}}_{ab})} = \frac{5}{995}$$

because I stated initially that an average of 5 out of every 1000 embayments will experience this phenomenon. These are the odds before *any* measurements are made. The likelihood ratio is

$$[3] \quad \frac{p(\boldsymbol{D}|\boldsymbol{H}_{ab})}{p(\bar{\boldsymbol{H}}_{ab})} = \frac{95}{5}$$

because a positive assessment is given 95% of the time an algal bloom will occur and 5% of the time when a bloom never happens. The posterior odds are

$$[4] \quad \frac{p(\boldsymbol{H}_{ab}|\boldsymbol{D})}{p(\bar{\boldsymbol{H}}_{ab}|\boldsymbol{D})} = \frac{95}{5} \times \frac{5}{995} = \frac{475}{4975}$$

Because this fraction is an expression of odds, it means out of every 5450 positive indications with this assessment technique, only 475 will be from embayments that actually develop algal blooms and 4975 will be from embayments that never go eutrophic. Therefore, the probability that a positive test is from an embayment that actually will have an algal bloom is

$$[5] \quad \frac{475}{475 + 4975} = 9\%$$

If you had assumed your remote sensing "predictor" was giving you accurate advice and advocated the implementation of remediation measures, it is because you were probably impressed by the wonderfully high diagnosticity of the assessment technique. Unfortunately, if the prior odds are ignored (a reflection of the base rate or the commonness or rareness of the outcome you are trying to predict), the conclusions can be extremely misleading. What is even more distressing is that the confirmatory results (no blooms occurred) from treating the 91% of the lagoons that would have never gone eutrophic anyway with whatever remediation technology was applied would just reinforce your pre-existing bias: all the money your client spent on the remediation treatment obviously was worth it, because no blooms occurred; ergo, your assessment technique was right on the money and essential for diverting impending disaster in these 5450 embayments. The false positives would never be identified without a double-blind experimental design if the base rates (priors) are ignored. All too often we are impressed by the wonderfully high diagnosticity of certain assessment techniques (just look at what is going on in the field of sediment quality criteria and the application of bioassays using sensitive test species); these are reflected in the likelihood ratio. If there were some way to get scientists and regulators to pay attention to the base rates, there would be a phenomenal amount of money saved in toxicity tests as well as immense improvements in diagnostic accuracy of ecosystem health or stress.

Suggestion 2: Entertain alternative hypotheses

Aldous Huxley is credited with stating that "The tragedy of science is that frequently a beautiful hypothesis is slain by an ugly fact." (Preston 1981). One of the most basic concepts that most ecologists and environmental scientists are aware of but relatively few (if any of us) uphold in actual practice is to seriously consider alternative hypotheses. One of the main problems with NHST is that it assumes a totally binary system of alternatives, either the null hypothesis or the research hypothesis (and, as readers know by this point, it tells us absolutely nothing about the research hypothesis); it is extremely rare to come upon a phenomenon in nature where there are only two possible explanations (i.e., your research hypothesis or the null hypothesis of no change or difference). It is more likely that a continuum of explanatory hypotheses between these two choices exist.

Investigators would profit enormously from being self-critical as well as objective while developing as broad an hypothesis set as possible. The essence of decision analysis arguments put forward by Raiffa (1968) and others is that when people make decisions under uncertainty, they will examine the range of possible outcomes of each decision in terms of costs (or some other measure); they will then weight each possible outcome in terms of how likely they expect that outcome to occur. In other words, people make choices by comparing the expected payoffs from their range of choices, where the expected utility of each choice is a sum or average of possible outcomes, each weighted by the odds of its occurrence. If this is indeed an accurate depiction of decision making, then, in essence, no possible outcomes are actually "rejected," especially if an outcome is very costly. Instead, people might assign a particularly costly outcome a low probability, but they still could worry about it as a possibility.

In the last section of the paper where the problems of preconceived notions and hindsight bias were discussed, one good training exercise to heighten awareness of potential bias is that once you have explained how outcome "x" might have been expected given the existing data, then attempt to explain how outcome "y" (one of your several alternative hypotheses) can be supported by the same set of data instead of outcome "x." Problem solvers since Benjamin Franklin (Wickelgren 1974) have suggested that decision making is improved if one ensures that all alternatives are given substantial consideration; even modest efforts in this approach (i.e., listing one or two reasons why you think your conclusion is incorrect and one or two reasons to support other interpretations) have been shown to pay substantial dividends (Faust 1986*b*; Fischhoff 1982; Koriat et al. 1980; Slovic and Fischhoff 1977). Studies in the medical field have shown that the most accurate diagnosticians tend to arrive at their final diagnosis later than do less accurate clinicians (Elstein et al. 1978); I cannot imagine any reason why the same would not be true for ecologists or environmental scientists. Premature formulation results in the biased processing of subsequent data. It appears that a valid way to improve accuracy and reduce bias is to entertain alternative hypotheses for a long period of time (Arkes 1981).

Suggestion 3: Critically examine predictor variables

When doing any risk assessment or environmental impact study, the ecologist (consultant, academician, etc.) is more often than not faced with a myriad of multidisciplinary data. Like benthic ecologists who are routinely faced with making sense of large 3-dimensional data matrices (station locations × species × abundance), environmental scientists frequently turn to multivariate statistical techniques such as classification and ordination (e.g., Pielou 1984) as a way of organizing, distilling, and making sense out of a large volume of data. While every statistical text always cautions practitioners about the connection between correlation and causation, as Gould (1981) has pointed out, much of the fascination with statistics is embedded in the gut feeling that abstract measures summarizing large tables of data express some mysterious truth or something more real and fundamental than the data themselves. Gould (1981) gives the humorous illustration of factor analysis of a 5×5 correlation matrix of his age, the population of Mexico, the price of Swiss cheese, his pet turtle's weight, and the average distance between galaxies during the past 10 years yielding a strong first principal component (p. 250).

My intent is not to just emphasize these cautionary restrictions on multivariate techniques, but instead to highlight research in the field of expert judgment on predictive variables (e.g., Goldberg 1968; Dawes 1979; Faust 1986*a*, 1986*b*). The real problem is not so much about how to weight or interrelate different variables (contrary to what the literature in benthic ecology would indicate, e.g., Erman and Helm 1971; Field 1971; Hughes and Thomas 1971; Cassie 1972; Chardy et al. 1976; Culp and Davies 1980; Poore and Mobley 1980; Kohn and Riggs 1982; Williams et al. 1982), but, more precisely, which variables should be used for consideration. The recommended course of action may seem counter-intuitive, because conventional wisdom would dictate that the more information one has at one's disposal to integrate, the better the results. Results to date indicate that improving judgmental accuracy is usually more an exercise in exclusion than one of inclusion (Faust 1989).

A limited set of *valid* predictors (ca. 3 or 4), if simply added together and not weighted, is as predictive or nearly as predictive as optimally weighted variables (Goldberg 1968; Dawes 1979; Faust 1986*a*, 1986*b*). Psychologists, like benthic ecologists and risk assessors, are more likely to err by overincluding predictors than by not optimally weighting the most valid predictors. An especially sobering fact for those embarking on risk assessments is that including more than the two or three most valid variables in the prediction formula only minimally increases predictive accuracy (Faust 1986*a*), and expanding the list of variables for inclusion often decreases predictive accuracy. The recent resurgence of interest in the minimalist approach of bounded rationality decision models also supports the contention that better decisions can be made with a few simple predictors (Bower 1999). The ceiling on predictive accuracy is usually approached once two, three, or four of the most valid variables have been identified (Faust 1989); the trap most benthic ecologists fall into is that because they have taken the time to identify *all* the animals in a sample, they feel compelled to include all these data in their final analyses (Germano 1985). This type of error is magnified by techniques such as the sediment quality triad and the AET, which include benthic community summary indices as part of their final index value. If an individual variable is not a valid, reliable predictor, the greatest likelihood is that it will decrease the overall accuracy of any conclusion reached. Faust (1986*a*, 1986*b*) has summarized three useful tests to determine which variables should be included for diagnostic predictions:

Test 1: Is there a true association?

To determine if a variable is a valid predictor, it is essential to have an accurate assessment of covariation: the environmental outcome must occur more often when the variable is present than when it is absent. To determine this, one must have information from *all* four cells in Fig. 2; if not, any conclusions about the validity of the predictor under consideration must be considered tentative.

Test 2: Does the measured variable increase diagnostic accuracy?

While a predictor may pass Test 1, it still may not pass Test 2. The predictor must surpass the diagnostic hit rate achieved by using the base rates alone; therefore, the frequency of false identifications must be lower than the frequency of the environmental condition. If it is not, and false negative errors are no more costly than false positives, then one should not use the predictor and just rely on the base rates. Using the base rates alone to support predictions means that if the condition occurs more than 50% of the time, one should always say *yes* (condition present); if the condition occurs less than 50% of the time, always say *no*. For high- and low-frequency outcomes (i.e., very common or very rare occurrences), most predictors do not improve upon the accuracy rate achieved just by using base rates alone (the example of the flounder liver lesions in the last section).

Test 3: Does the measured variable produce incremental validity?

This is the final acid test, for a predictor can pass Tests 1 and 2 and still fail this last one. This can occur when the predictor is redundant with other predictive variables of higher validity (or may add so little incremental validity that it is not worth the effort to obtain). Adding a variable of low or modest validity to two or three variables of greater validity will usually decrease or, at the very least, not alter judgment accuracy.

Faust (1986*b*) points out that very few of the diagnostic tests used in clinical psychology pass all three of these tests (many do not pass the first one). Very few (if any) environmental predictors have ever been subjected to any of these tests; witness the confusion and controversy during the last few years over endocrine disruptors (Kaiser 1996). The rote practice of risk assessors automatically calculating Hazard Indices and Hazard Quotients (wholly dependent upon the num-

ber of chemicals of concern measured or identified) would most likely be eliminated entirely if these supposed predictors were subjected to the above tests. If environmental scientists routinely applied these rules, I am fairly confident that both the U.S. Army Corps' and EPA's "Green book" and "Inland testing manual" for dredged material testing would look quite different from what it does today.

Suggestion 4: Decrease reliance on memory

Eliminating the fallibility of recall is one of the most simple but effective strategies, and one that I am reminded of daily as I get older. Arkes and Harkness (1980) report that unpresented symptoms consistent with a diagnosis tended to be remembered as having been presented; conversely, in some circumstances, presented symptoms that were inconsistent with the diagnosis were not remembered as having been presented (the bias of preconceived notions). Unfortunately, one tends to remember the facts supportive of any particular hypothesis and to forget those inconsistent with the hypothesis.

Another problem of relying on memory instead of data records as far as overcoming the common tendency to misestimate covariation was highlighted by Ward and Jenkins (1965). Estimates of covariation in their tested subjects were grossly incorrect when individual pieces of data were presented one at a time (taxing subjects' short-term memory); when box-score summaries of all four cells shown in Fig. 2 were presented simultaneously, estimates were much more accurate.

With the wealth of environmental data to which most regulators and scientists have access, we would be much better served by simply looking things up instead of relying on our memory; this assumes that the data are organized and easily accessible, and the current trend toward database compilation and information management is a critical, necessary step in the right direction for scientists and environmental resource managers to eliminate these types of errors. As Arkes (1981) stated,

> ...a more humble view of one's own memory will result in less of a need to be humble about the accuracy of one's judgment (p. 329).

Suggestion 5: Increase reliance on actuarial methods

The scientist or consultant called in to do a risk assessment study or to be an expert witness in any sort of environmental litigation often is confronted with an overwhelming array of data. The real issue here is how well any individual, regardless of his or her professional background, education, or training can meet this demand. A considerable amount of research has been done on the merits of clinical versus actuarial judgment, starting with the first introduction to this topic by Meehl (1954) over four decades ago and numerous studies since that time (for reviews, see Meehl 1965; Sawyer 1966; Wiggins 1973, 1981). In the clinical method, the decision-maker (e.g., regulator, scientist, consultant, expert witness) combines or processes information in their head; in the actuarial or statistical method, the human judge is eliminated and conclusions are based solely on empirically established relations between the data collected or available and the outcome or condition of interest (Dawes et al. 1989).

The results of studies comparing clinical to actuarial methods are not only disappointing but often alarming; actuarial methods have consistently outperformed clinical judgment. There are more than 100 studies that have compared the accuracy of clinical to actuarial judgement encompassing a wide range of diagnostic and predictive tasks; based on a recent review, Dawes et al. (1989) stated "in virtually every one of these studies, the actuarial method has equaled or surpassed the clinical method, sometimes slightly and sometimes substantially" (p. 1669). In one of the early and now classic studies in the field of psychology, Goldberg (1959) showed that the judgment effectiveness of experts (psychologists) in distinguishing patients with and without organic brain damage by interpreting test results failed to surpass that of a group of secretaries and barely exceeded chance levels. A variety of judgment studies since that time have shown low clinician performance (e.g., Goldberg 1968; Einhorn 1972; Slovic and MacPhillamy 1974). Dawes (1971) found that an actuarial method based on a single variable was more accurate at predicting success in graduate school than the clinical judgments of the admissions committee working with data from multiple sources. There are several reasons for this, many of which have been touched on above (e.g., the clinician's difficulty in distinguishing between predictive and nonpredictive variables, the bias of preconceived notions), but the essential problem appears to be that human beings are not computers: individuals have difficulty handling more than two or three variables, so they are not effective at mental retention, weighting, and (or) organization of data (Faust 1984, 1989).

While the scientist involved in risk assessment studies may contend that careful attention to the relative importance of variables (or proper weighting) is crucial to a final interpretation, not only do individuals have difficulty in assigning such weights (Faust 1984, 1989), but Dawes and Corrigan (1974) have shown that the value of optimal weighting is often overestimated; it usually provides negligible advantage over equal unit weights. In most cases, predictions on equal unit weights correlate highly with those based on optimal weights; even randomly assigned weights generally result in predictive accuracy that approaches optimal weighting of variables (see Dawes 1979 for discussion of these issues). These findings suggest that the derivation of assessment techniques such as the sediment quality triad or AET could bear re-examination. Meehl (1986) addressed subjective weighting of variables as follows:

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments...from empirical studies...for believing

> that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently (p. 372).

After citing the above passage, Dawes et al. (1989) expanded this line of thought in response to those who justified the clinical or human judgment approach because they were not dealing with simple additive models:

> Suppose instead that the supermarket pricing rule were, "Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price"; would the clerk and customer eyeball that any better? Worse, almost certainly (p. 1672).

The authors illustrate quite dramatically that if human judgment performs poorly with simple additive models, then it should not be expected to do better with mental models requiring more complex manipulation of the input variables. A considerable body of research in the fields of psychology, medicine, and decision-making indicate that properly developed actuarial methods are more accurate at both diagnosis and prediction than the clinical method, even when the scientist or decision-maker has access to equal or greater amounts of information. It appears that we as environmental scientists or ecologists would make greater advances in assessing environmental health or predicting risk if we spent more time developing proper actuarial methods so that the constant reliance on an individual regional regulator's "best professional judgment" so common in regulatory guidance documents (the "Green book" and "Inland testing manual") would become a thing of the past.

Suggestion 6: Recognize predictive uncertainty

Although the public at large often looks to the scientist to "provide the facts" so that decisions can be made with absolute certainty based on known outcomes, environmental problems are not chemical reactions. Prediction contains a certain element of chance, often more than most people recognize or scientists are willing to admit. Whether scientists like to admit it or not, errors are unavoidable. Because of the limitations of our knowledge of ecosystem function and the unexpected results that can occur from the interaction of human activities with natural ecological processes, the predicted "outcomes" are often times a best guess. If one defines an experiment as "an action whose outcome we cannot predict precisely or specify beforehand" (i.e., the disposal of dredged material in the natural system), then an alternative frame of reference is to view environmental monitoring or remedial investigations as feedback mechanisms providing data about the outcome of experiments (Bernstein and Zalinski 1986). This admission of fallibility does not detract from the underlying basic value of the study or the development of better assessment and predictive techniques; it merely requires letting go of the illusion of certainty (Holling 1978).

Because chance is unavoidable in prediction, there are two general suggestions to deal with this sobering fact (Faust 1986*b*). First, one should not abandon good predictors because they are fallible (Arkes et al. 1986); it is unrealistic to think that we will ever find predictors that only score "hits" in Cells A and D in Fig. 2 (i.e., no false positives or negatives). This can cause scientists to dismiss state-of-the-art predictors when examined in light of the information presented in this paper because they do not meet their preconceived notions of what scientific standards should be. There are useful nonparametric methods available for estimating prediction error (Efron and Gong 1983), so the magnitude of errors associated with different predictors can be compared. One should have a better predictor before one gives up a good one (Faust 1986*b*).

Second, scientists and consultant experts alike should acknowledge the uncertainty associated with their conclusions, so that when decision makers receive these results, they can be conservative when appropriate; it is rare to find a predictive situation that justifies extreme confidence. Obviously, if the cost of errors is grave (acute human health risk), then one should be cautious and err in that direction (a false positive is much more desirable in this case than a false negative).

Part 3: Summary

In one of his papers on the application of human judgment to clinical practice, Faust (1986*b*) stated that two types of insight are needed for improvement, which I believe apply to ecologists as well. The first is an accurate assessment of the current problems in the field; investigators need to educate themselves about the situations in which errors are likely and the sources of errors. The second is that investigators must recognize the limits of self-recognition. Just having an awareness or admitting that one does make errors does not protect one against them; many of our judgment habits are so ingrained that they occur without our recognition and may be only partly accessible to conscious awareness. Ecologists, just like medical clinicians, need to learn when not to rely on intuitive judgments and when to use cross-validation methods and actuarial techniques that surpass their unaided decision powers.

Six recommendations were made to improve diagnostic accuracy and to suggest alternative approaches to ecological risk assessment:

1. Use Bayesian statistical methods
2. Systematically evaluate outcomes from a broad hypothesis set
3. Critically examine predictor variables using the three recommended tests
4. Decrease reliance on memory
5. Increase reliance on actuarial methods
6. Recognize predictive uncertainty

Suggestion numbers 1, 3, and 5 would constitute more of a paradigm shift in a basic approach for ecologists and risk assessors, and they are all linked to one another. Valid actuarial methods can be derived once accurate predictor variables are found, and inferences derived from Bayesian statistical methods far outweigh those from a classical or Fisherian approach for finding accurate predictors.

These six approaches would provide a new route to discovering the untapped information contained within the enormous environmental databases that have been compiled by EPA, National Oceanic and Atmospheric Administration (NOAA), or the U.S. Army Corps of Engineers (Ocean Data Evaluation System (ODES), Storage and Retrieval (STORET), National Status and Trends (NS&T), Environmental Residue-Effects Database (ERED), etc.). The advances in electronic imaging workstations combined with the wealth of data that already exists would allow investigators to have access to and sort variables of interest rapidly and efficiently so that they could accurately examine covariation for proposed predictors or criteria. Alternative hypotheses could be formulated and tested with a wealth of on-line data; the usefulness of any suggested predictors could then be assessed with Bayesian methods. Finally, the derivation of any predictive variable or measurement technique could then be followed by cross-validation studies; this type of follow-up is essential, because a predictive assessment technique should be shown to work where it is needed, i.e., in cases where the outcome is unknown. Crane and Newman (1996) argue that progress in environmental toxicology is more likely by a pluralistic approach that is embedded in a systematic structure, and they point out that, " . . . the sole use of classical statistical techniques may overlook useful contributions from Bayesian theory" (p.120). I would suggest that the six recommendations outlined above provide such a pluralistic approach, but I would discourage altogether the use of classical statistical techniques (NHST) for the reasons pointed out earlier.

It would be unreasonable to think that ecologists will embrace these approaches any more rapidly or willingly than mental-health practitioners have over the past 40 years; aside from the normal resistance or inertia to preserve the status quo and to do things the way they have always been done, i.e., the way ecologists were trained or taught, these approaches can, on first glance, appear threatening. Once valid predictors are established and reliable actuarial methods developed, why would anyone require the scientist or expert to interpret the data? A common anti-actuarial argument to prevent this frightening potential endpoint from occurring is that "group statistics" do not apply to single events, and, therefore, the expert is still required to interpret the unique aspects of a particular case. As Dawes et al. (1989) point out, this is really a misconception that ignores basic principles of probability. While individual events or situations will no doubt exhibit unique features, they typically share common features with other events and (or) situations that would permit the previously established predictor variables to be tallied and an outcome predicted at a specified power. The authors drive this point home by stating,

> An advocate of this anti-actuarial position would have to maintain, for the sake of logical consistency, that if one is forced to play Russian roulette a single time and is allowed to select a gun with one or five bullets in the chamber, the uniqueness of the event makes the choice arbitrary.

The one item in the above list that would have the biggest impact on improving our ability to assess ecosystem health (interpret bioaccumulation results, assess sediment quality criteria, perform valid risk assessments, etc.) would be to abandon NHST and switch to Bayesian models. There are two crucial rules that should be evident from the earlier example of the algal bloom diagnosis (Arkes 1981):

1. If the prior odds are x/y, then the likelihood ratio must be larger than y/x for the hypothesis to be correct more than 50% of the time

2. The lower the prior odds, the greater the likelihood ratio has to be to justify that hypothesis.

For example, if the prior odds are $\frac{1}{2}$, a likelihood ratio of 4/1 makes the hypothesis likely ($p = 0.67$). If the condition or outcome you are trying to assess is rather rare, i.e., 1/100, then the same likelihood ratio puts your hypothesis in the long shot category ($p = 0.04$). It is ludicrous to think we will improve our diagnostic accuracy in ecological investigations or interpret results in a meaningful fashion if we continue to ignore base rates; it is important to always keep in mind that posterior odds are essentially a contest between prior odds and the likelihood ratio (therefore, base rates cannot be ignored). This also will lead to a much more accurate assessment of covariation in the environmental predictors we are trying to validate.

The need to acknowledge the base rates and take into account the prior odds brings to the forefront the main criticism to which people point as the reason not to use Bayesian statistical methods. Bayesian analysis requires a prior distribution for the unknown parameter being studied. Most investigators, when pressed, would contend that they have no idea what the prior odds are for the particular phenomenon they are studying (this is, in fact, the main reason they are collecting data in the first place). Therefore, the more "objective" approach of classical (Fisherian) statistics is justified as appropriate for scientific investigations. However, stating that one has no idea, or that one value is just as good as any other value, implies the notion of a rectangular distribution for the possible range of values for the unknown parameter (Iverson 1984). Even the rectangular distribution embodies some information about the parameter, and it also represents a step up from having no idea at all. There are problems associated with using the rectangular distribution (one of a class known as noninformative priors) in Bayesian analysis for the prior odds, and the reader is referred to texts that deal with this issue in more detail (e.g., Raiffa 1968; Schmitt 1969; Iverson 1984; Berger 1985). However, there usually is some available prior information, because research is never done in a vacuum (if absolutely nothing was known about a parameter, then no one would have ever thought

of doing research in the first place). Whatever prior information does exist, regardless of how vague, usually can be expressed in an informative rather than a noninformative prior distribution. Edwards et al. (1963) point out that prior distributions are often quite vague, and it is most likely this vagueness that has discouraged investigators from adopting Bayesian techniques over classical statistical methods. The authors capture the irony of this discomfort by paraphrasing de Finetti's (1959, p. 19) reflections on the inherent construct of the Bayesian objectionists' arguments:

> We see that it is not secure to build on sand. Take away the sand, we shall build on the void.

The advantages of Bayesian methods over classical methods for data inference are numerous (Edwards et al. 1963; Efron 1986; Berger and Berry 1988; Reckhow 1990, 1994; Howson and Urbach 1991). The attractiveness of Bayesian methods resides in its parallel construct with research methodology. Research is cumulative, and while classical statistics are based on the "once-ness" of the experiment, Bayesian methods permit the use of knowledge or results from earlier research in the formulation of results from new research. However, critics continually counter that subjective prior distributions have no place in an objective scientific analysis. Very few statistical analyses even approximate objectivity. Both Iverson (1984) and Berger (1985) point out that classical statistics are just as subjective as Bayesian analysis, and that subjectivity is expressed in the choice of the significance level and the choice of the statistical model used for the final analysis (which can have a much greater impact on the outcome than the choice of a prior distribution). Box (1980) expressed surprise at how expressing probabilities for prior beliefs has been thought of as a trait peculiar to Bayesian inference, and that

> this seems to come from the curious idea that an outright assumption does not count as a prior belief.

Good (1973) was much more blunt by stating,

> The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science.

The use of priors in Bayesian methods brings the subjective aspects of the analysis out in the open; the analyst is forced to express personal opinions and biases, and each reader can make their own assessment about the reasonableness of the priors. Anyone is free to apply their own set of priors on the data if they so desire to see how it affects the posterior distribution, but for large samples, the prior distributions typically have little or no effect on the posterior distribution (principle of stable estimation; see Iverson 1984 for more details).

While there is not a universally accepted way for assigning prior probabilities (Jefferys and Berger 1992), there do exist well-developed techniques for calculating prior odds. Unfortunately, many of them are not for the statistically unsophisticated; a quick glance at the formulas in Chap. 3 of Berger (1985) will confirm this. Instead of shunning a Bayesian approach because these calculations may be more advanced than most ecological investigators are used to dealing with (they are not one of the statistical function choices in an Excel® spread-sheet or an option in Statistical Package for the Social Sciences (SPSS) routines), it would be much more fruitful to enlist the help and collaboration of statisticians trained in Bayesian methods. Our task as ecologists, risk assessors, or environmental scientists would be to acquire a rudimentary understanding of Bayesian methods to provide the statistician with the information they need to perform their analyses (see Winkler 1967, for a good example of how nonstatisticians were guided in constructing their priors) and then interpret the results in a relevant context for the particular ecosystem under investigation.

I would venture that developing actuarial methods through these approaches, rather than making the individual expert obsolete, would have just the opposite effect. This shift in investigative focus in the field of ecology and risk assessment would increase the demand for trained investigators and open up a diversity of new avenues for research aimed at establishing valid predictors of environmental health; these, in turn, would necessarily lead to new levels of understanding about as well as new models of how ecosystems function (which in turn would lead to new lines of research, etc.). While it would require some initial retraining of some of us grey-hairs in the field, we can take solace from remembering that Andrés Segovia took his first guitar lesson at the age of forty-five, and some of the most brilliant lectures on ecology delivered at Yale were done by G.E. Hutchinson when he was in his eighties.

Discussion

It is sobering to reflect on the possible ramifications of some of the viewpoints expressed in this paper. A positive impact (and hopefully the primary one) will be to provide investigators with a new set of lenses through which they can view existing databases or design new investigations to discover insights about ecosystem structure and function. The current state of ambiguity in predicting bioeffects from suggested sediment quality guidelines with relatively low reliability, e.g., false negatives $\leq$ 25% (Long et al. 1998), would be drastically revised once investigators adjusted their conclusions by paying attention to base rates and accurately assessed covariation to decide if these proposed guidelines really are valid predictors of environmental health. It would also prevent sediment criteria from being derived by methods that ignore procedures covered earlier that would increase the diagnostic accuracy of these predictors. Another positive impact would be a complete revision for the way ecological risk assessments are done, with a justifiable dismissal of conclusions reached from calculating questionable or unvalidated metrics such as Hazard Indices or Hazard Quotients (a classic example of admitting insufficient

evidence as sufficient; see Faust 1986*a* for further details).

The downside is that anyone can use the arguments presented earlier in Part 1 of this paper to derail any conclusions based on results from classical statistical tests structured around the sacred *p* value of 0.05. This could range from dismissing results from toxicity or bioaccumulation testing in the wide variety of permit applications, enforcement actions, or court cases dealing with contaminated sediment management to determining injury in NRDA cases. A great deal of environmental case law is unfortunately structured around and based upon misunderstandings about statistical significance. For example, in Ohio versus Department of Interior, 880 F.2d 432 (D.C. Cir. 1989), the court analyzed what type of causal link must be shown between a substance release and an injury to establish National Resource Damage (NRD) liability. The court upheld the Department of the Interior (DOI) Rules that state that to show biological injury, a trustee must show that a number of acceptance criteria are met, including that the biological response identified differs in a "statistically significant" way from the condition of similar organisms in a control area; as pointed out earlier, this is easily achieved merely by taking a sufficient number of samples. It is also difficult to ignore the irrelevance of standard statistical significance testing because of "odds against chance" fantasy outlined earlier, as well as the minor detail that standard statistical tests tell us absolutely nothing about the validity of our research hypothesis; all of these arguments would make it quite easy for any skillful lawyer to repudiate any claims of environmental injury based on results from standard tests of statistical significance. Standard statistical techniques do have a proper place and can be improved tremendously by the application of modern methods (see Rand 1996, 1997, 1998); the main point of this paper is that they have provided more heat than light as inferential discriminators to discover valid predictors of environmental health.

Classical procedures are very asymmetric (in Fisher's inference model, the choices are *reject* and *inconclusive*, whereas in the decision theory approach, the choices are *reject* and *accept*). Bayesian procedures, in stark contrast, can strengthen as well as weaken a null hypothesis. In classical approaches, if the null hypothesis is rejected, the alternative is willingly embraced, whereas if it is not rejected, it remains in a "limbo of suspended disbelief" (Edwards et al. 1963). Ecological investigators who feel more comfortable or insist on using classical techniques would fare much better by distinguishing between "sharp" versus "loose" null hypotheses; as pointed out earlier, there is usually little reason to believe a sharp null hypothesis actually exists in nature, so logic would suggest that we should not test a hypothesis of zero relationship when we want to establish that a relationship actually does exist; as Mohr (1990) points out, rejecting a claim of precisely zero is not very informative. Not every choice is dichotomous, corresponding to a simple partition between two hypotheses, because gradations of difference within each hypothesis can be important. The times in life are relatively few when we must choose between exactly two acts, one appropriate to the null hypothesis and the other to its alternative. People would not save for their retirement if they believed they would die before they stopped working, and they would not buy life insurance if they believed they would not. Many intermediate acts, or bet hedging, is possible with most situations, and this continuum of relationships that can be addressed with Bayesian models is the more appropriate one for ecological risk assessment (Hill 1996). Instead of always employing sharp null hypotheses, the investigator using classical methods should frame the test by saying, "I believe there is a relationship in the population of at least such-and-such magnitude." The practical problem is that it forces the investigator to think about and supply meaningful numbers to computer programs (and it would also require that these computer programs *ask* for these input parameters instead of automatically throwing zero in the formula for *t*). The Bayesian analyst is unlikely to consider a sharp null hypothesis as often as someone using classical techniques; it would make no sense unless the null hypothesis deserves some special initial credence or the Bayesian's prior distribution has a sharp spike around the null hypothesis value.

The popularity of statistical significance testing would decline if researchers and regulators recognized that it is *not* a predictor of replicability of research data. As Stevens (1971) pointed out over 25 years ago,

> In the long run, scientists tend to believe only those results that they can reproduce. There appears to be no better option than to await the outcome of replications. It is probably fair to say that statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything (p. 440).

Berkson (1938) has suggested the use of the

> interocular traumatic test: you know what the data mean when the conclusion hits you between the eyes (Edwards et al. 1963).

As Dominic DiToro so aptly stated at a national EPA conference on sediments in 1996,

> Statistics should be used like a drunk uses a lamp post: more for support than illumination.

Ecological researchers and environmental risk assessors need to overcome the misplaced belief that if their treatment of subject is mathematical, it is therefore precise and valid.

A paradigm shift in how ecological data are assessed would not only allow the development of more precise "ecological laws" (*sensu* Loehle 1988) and point the way to valid predictors of environmental health, but it would also likely affect our perception of ecosystem function. Whether or not particular data are used will affect what we observe actually "exists" in the particular system we are studying. For example, most people in the United States are incorrect in their answers about which way Lincoln's profile faces on a penny, because they do not normally use this information. This example illustrates

the difference between perceiving and observing; perceiving is prior to observation. Observed information can be reported verbally and processed, while perceived data that are not observed cannot be reported or factored into conclusions or results. Just because something is perceived does not mean that it is observed, and something can also be observed without appreciating its significance in terms of how it affects ecosystem function (in familiar terms for benthic ecologists, just what do all those nematodes, ostracods, or unidentified oligochaetes really mean when you count them in a benthic sample?).

Scientific "truth" has always been a relative term, as any historian of science is quick to point out: today's scientific "facts" are looked upon with bemused curiosity by future generations in relation to the contemporary accepted scientific constructs, similar to how parents chuckle at a child's internal logic of "where babies come from" in relation to known human reproductive physiology. Kuhn (1962) pointed out over 30 years ago that major changes in thinking or "paradigm shifts" have come about because of both a willingness of people to question the status quo and to invest a concerted effort in alternative approaches, not just because a better theory has arrived on the scene. Science will progress only "after the new paradigm has been developed, accepted, and exploited" (p. 156). The change in direction that new theories dictate is not simply determined by internal logic, but by other societal factors that come into play to determine the community's choice of a "best theory" (Keller 1985). The implications of this, as Keller has pointed out in her outstanding work, is that not only different collections of facts or different focal points of scientific attention are consistent with what we call science, but so are different organizations of knowledge and interpretations of the world.

What is being proposed in this paper is not just a mere substitution of Bayesian methods for NHST; anyone agreeing with the thoughts presented in this manuscript who quickly runs out and starts reading material on Bayesian statistics with the hope of finding a new basis for automatic inference will no doubt be disappointed (remember, it is only one of the six suggested techniques for improving diagnostic accuracy). In his landmark text, Raiffa (1968) pointed out that students in their first course in statistics learn that they must constantly balance between making an error of the first kind (rejecting the null hypothesis when it is true) with the error of the second kind (accepting the null hypothesis when it is false). He credited Tukey with suggesting that all too often, practitioners make errors of a third kind (solving the wrong problem), and he nominates a candidate for errors of the fourth kind: solving the right problem too late. My hope is that if ecologists and risk assessors adopt the approaches outlined in this paper, we can stop making the first three kinds of errors in our routine investigations and finally develop accurate predictors of ecosystem health. Risk assessments and environmental regulations will only make sense if they are based on valid environmental predictors, and therein lies our hope for not committing errors of the fourth kind.

Acknowledgements

Valuable feedback on earlier drafts of this manuscript was provided by Peter Chapman, Magda Havas, Barbara Hecker, Ryan Hill, Lorraine Read, Donald Rhoads, and an anonymous reviewer. While many of the ideas presented in this paper were first developed as the basis for an invited talk at the COST 647 Symposium at the University of Kiel in the fall of 1990, I owe the biggest debt of gratitude for the ideas presented here to two people. The first is Donald Rhoads, who was my advisor in graduate school, and who has been a close friend and colleague for the past 20 years. It was he who first exposed a naive biologist to looking at benthic ecosystems from an alternative, more expansive point of view (animal–sediment–fluid interactions), and many of the understandings I have gained about benthic ecosystems through our investigations with REMOTS® technology have been a result of discussions and collaborations with him. The second is my wife Marilyn, who through her graduate education and research in clinical psychology not only exposed me to all the relevant statistical literature, but who also has been a midwife (literally and figuratively) during the past 30 years for my intellectual, emotional, and spiritual development. Her professional career as a midwife and clinical psychologist has been both an inspiration and model for me in my recent forays into the field of environmental mediation. I am deeply indebted to both of these individuals and consider myself fortunate to have had such invaluable professional and personal partners.

References

Arkes, H.E. 1981. Impediments to accurate clinical judgment and possible ways to minimize their impact. J. Consult. Clin. Psychol. **49**: 323–330.

Arkes, H.E., and Harkness, A.R. 1980. The effect of making a diagnosis on subsequent recognition of symptoms. J. Exp. Psychol. Hum. Learn. Mem. **6**: 568–575.

Arkes, H.E., Wortmann, R.L., Saville, R.D., and Harkness, A.R. 1981. The hindsight bias among physicians weighing the likelihood of a diagnosis. J. Appl. Psychol. **66**: 252–254.

Arkes, H.E., Dawes, R.M., and Christensen, C. 1986. Factors influencing the use of a decision rule in a probabilistic task. Behav. Hum. Decis. Processes, **37**: 93–110.

Bakan, D. 1966. The test of significance in psychological research. Psychol. Bull. **66**: 423–437.

Berger, J.O. 1985. Statistical decision theory and Bayesian analysis. 2nd ed. Springer-Verlag, New York.

Berger, J.O., and Berry, D.A. 1988. Statistical analysis and the illusion of objectivity. Am. Sci. **76**: 159–165.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. J. Am. Stat. Assoc. **33**: 526–542.

Bernstein, B. 1990. A framework for trend detection: coupling ecological and managerial perspectives. Presented at the International Symposium on Ecological Indicators, Fort Lauderdale, Fla., October 16–19, 1990.

Bernstein, B.B., and Zalinski, J. 1986. A philosophy for effective monitoring. *In* Proceedings of IEEE OCEANS '86 Conference, Washington, D.C., September 23–25, 1986. pp. 1024–1029.

Bower, B. 1999. Simple minds, smart choices. Sci. News, **155**: 348–349.

Box, G.E.P. 1980. Sampling and Bayes inference in scientific modeling and robustness (with discussion). J. R. Stat. Soc. Ser. A, **143**: 383–430.

Carver, R.P. 1978. The case against statistical significance testing. Harv. Educ. Rev. **48**: 378–399.

Cassie, R.M. 1972. Fauna and sediments of an intertidal mud-flat: an alternative multivariate analysis. J. Mar. Biol. Ecol. **9**: 55–64.

Chapman, L.J., and Chapman, J.P. 1967. Genesis of popular but erroneous psychodiagnostic observations. J. Abnorm. Psychol. **72**: 193–204.

Chapman, L.J., and Chapman, J.P. 1969. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. J. Abnorm. Psychol. **74**: 271–280.

Chapman, P.M., Dexter, R.M., and Long, E.R. 1987. Synoptic measures of sediment contamination, toxicity and infaunal community composition (the sediment quality triad) in San Francisco Bay. Mar. Ecol. Prog. Ser. **37**: 75–96.

Chardy, P., Glémarec, M., and Laurec, A. 1976. Application of inertia methods to benthic marine ecology: practical implications of the basic options. Estuarine Coastal Mar. Sci. **4**: 179–205.

Clark, C.A. 1963. Hypothesis testing in relation to statistical methodology. Rev. Educ. Res. **33**: 455–473.

Cohen, J. 1962. The statistical power of abnormal-social psychological research: a review. J. Abnorm. Soc. Psychol. **65**: 145–153.

Cohen, J. 1994. The earth is round ($p < .05$). Am. Psychol. **49**: 997–1003.

Cowles, M., and Davis, C. 1982. On the origins of the .05 level of statistical significance. Am. Psychol. **37**: 553–558.

Crane, M., and Newman, M.C. 1996. Scientific method in environmental toxicology. Environ. Rev. **4**: 112–122.

Cronbach, L.J., and Snow, R.E. 1977. Aptitudes and instructional methods: a handbook for research on interactions. Irvington, New York.

Culp, J.M., and Davies, R.W. 1980. Reciprocal averaging and polar ordination as techniques for analyzing lotic macroinvertebrate communities. Can. J. Fish. Aquat. Sci. **37**: 1358–1364.

Dawes, R.M. 1971. A case study of graduate admissions: Application of three principles of human decision making. Am. Psychol. **26**: 180–188.

Dawes, R.M. 1979. The robust beauty of improper linear models in decision making. Am. Psychol. **34**: 571–582.

Dawes, R.M., and Corrigan, B. 1974. Linear models in decision making. Psychol. Bull. **81**: 95–106.

Dawes, R.M., Faust, D., and Meehl, P.E. 1989. Clinical versus actuarial judgment. Science (Washington, D.C.), **243**: 1668–1674.

de Finetti, B. 1959. La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti da vista. *In* Induzione e statistica. Istituto Matematico dell'Universita, Rome.

Eberhardt, L.L., and Thomas, J.M. 1991. Designing environmental field studies. Ecol. Monogr. **61**: 53–73

Edwards, W., Lindman, H., and Savage, L.J. 1963. Bayesian statistical inference for psychological research. Psychol. Rev. **70**: 193–242.

Efron, B. 1986. Why isn't everyone a Bayesian? Am. Stat. **40**: 1–11.

Efron, B., and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. **37**: 36–48.

Einhorn, H.J. 1972. Expert measurements and mechanical combination. Organ. Behav. Hum. Perform. **7**: 86–106.

Einhorn, H.J., and Hogarth, R.M. 1981. Behavioral decision theory: processes of judgment and choice. Annu. Rev. Psychol. **32**: 53–88.

Elstein, A.S., Shulman, A.S., and Sprafka, S.A. 1978. Medical problem solving: an analysis of clinical reasoning. Harvard University Press, Cambridge, Mass.

Erman, D.C., and Helm, W.T. 1971. Comparison of some species importance values and ordination techniques used to analyse benthic invertebrate communities. Oikos, **22**: 240–247.

Faust, D. 1984. The limits of scientific judgment. University of Minnesota Press, Minneapolis.

Faust, D. 1986*a*. Learning and maintaining rules for decreasing judgment accuracy. J. Pers. Assess. **50**: 585–600.

Faust, D. 1986*b*. Research on human judgment and its application to clinical practice. Prof. Psych. Res. Pract. **17**: 420–430.

Faust, D. 1989. Data integration in legal evaluations: Can clinicians deliver on their premises? Behav. Sci. Law, **7**: 469–483.

Field, J.G. 1971. A numerical analysis of changes in the soft-bottom fauna along a transect across False Bay, South Africa. J. Mar. Biol. Ecol. **7**: 215–253.

Fischhoff, B. 1975. Hindsight $\neq$ foresight: the effect of outcome knowledge on judgment under uncertainty. J. Exp. Psychol. **1**: 288–299.

Fischhoff, B. 1982. Debiasing. *In* Judgment under uncertainty. *Edited by* D. Kahneman, P. Slovic, and A. Tversky. Cambridge University Press, New York. pp. 422–444.

Fisher, R.A. 1947. The design of experiments. 4th ed. Oliver & Boyd, Edinburgh.

Forbes, V.E., and Forbes, T.L. 1994. Ecotoxicology in theory and practice. Chapman and Hall, London, U.K.

Galen, R.S., and Gambino, S.R. 1975. Beyond normality: the predictive value and efficiency of medical diagnoses. Wiley, New York.

Germano, J.D. 1985. An evaluation of Gray's log–normal method. White paper prepared for U.S. Army Corps of Engineers, Waterways Experiment Station. EVS Environment Consultants, 200 West Mercer Street, Suite 403, Seattle, WA 98119, U.S.A.

Germano, J.D. 1991. To grab or not two grabs: infaunal benthic sampling strategies and the need for replication. A discussion of statistical power analysis. White paper submitted to EPA Region IX under Contract 68-C8–0061. EVS Environment Consultants, 200 West Mercer Street, Suite 403, Seattle, WA 98119, U.S.A.

Germano, J.D. Reflections on statistics, ecology, and risk assessment. *In* Organism–sediment interactions. *Edited by* J.Y. Aller, S.A. Woodin, and R.C. Aller. Belle Baruch Library in Marine Science. University of South Carolina Press, Columbia. In press.

Germano, J.D., Rhoads, D.C., and Lunz, J.D. 1994. An integrated, tiered approach to monitoring and management of dredged material disposal sites in the New England regions. DAMOS Contribution 87. U.S. Army Corps of Engineers, New England Branch, 696 Virginia Road, Concord, MA 01742, U.S.A.

Goldberg, L.R. 1959. The effectiveness of clinicians' judgements: the diagnosis of organic brain damage from the Bender–Gestalt Test. J. Consult. Psychol. **23**: 25–33.

Goldberg, L.R. 1968. Simple models or simple processes? Some research on clinical judgments. Am. Psychol. **23**: 483–496.

Good, I.J. 1973. The probabilistic explication of evidence, surprise, causality, explanation, and utility. *In* Foundations of statistical inference. *Edited by* V.P. Godambe and D.A. Sprott. Holt, Rinehart, and Winston, Toronto.

Gould, S.J. 1981. The mismeasure of man. W.W. Norton & Co., New York.

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

Green, R.H. 1984. Some guidelines for the design of biological mon-

itoring programs in the marine environment. *In* Concepts in marine pollution measurements. *Edited by* H.H. White. Maryland Sea Grant College, College Park. pp. 647–655.
Hays, W.L. 1963. Statistics. Holt, Rinehart, & Winston, New York.
Hill, R.A. 1996. From science to decision-making: the applicability of Bayesian methods to risk assessment. Hum. Ecol. Risk Assess. **2**: 636–642.
Holling, C.S. 1978. Adaptive environmental assessment and management. John Wiley & Sons, Toronto.
Howson, C., and Urbach, P. 1991. Bayesian reasoning in science. Nature (Lond.), **350**: 371–374.
Hughes, R.N., and Thomas, M.L.H. 1971. The classification and ordination of shallow-water benthic samples from Prince Edward Island, Canada. J. Exp. Mar. Biol. Ecol. **7**: 1–39.
Iversen, G.R. 1984. Bayesian statistical inference. Sage University Paper series on Quantitative Applications in the Social Sciences, 07–043. Sage Publications, Beverly Hills and London.
Jefferys, W.H., and Berger, J.O. 1992. Ockham's razor and Bayesian analysis. Am. Sci. **80**: 64–72.
Kahneman, D., and Tversky, A. 1973. On the psychology of prediction. Psychol. Rev. **80**: 237–251.
Kaiser, J. 1996. Scientists angle for answers. Science (Washington D.C.), **274**: 1837–1838.
Keller, E.F. 1985. Reflections on gender and science. Yale University Press, New Haven, Conn.
Kohn, A.J., and Riggs, A.C. 1982. Sample size dependence in measures of proportional similarity. Mar. Ecol. Prog. Ser. **9**: 147–151.
Koriat, A., Lichtenstein, S., and Fischhoff, B. 1980. Reasons for confidence. J. Exp. Psychol. **6**: 107–118.
Kuhn, T.S. 1962. The structure of scientific revolutions. University of Chicago Press, Chicago.
Lee, P.M. 1989. Bayesian statistics: an introduction. Oxford University Press, New York.
Lipsey, M.W. 1990. Design sensitivity: statistical power for experimental research. Sage Publications, Newbury Park, Calif.
Loehle, C. 1987. Hypothesis testing in ecology: psychological aspects and the importance of theory maturation. Q. Rev. Biol. **62**: 397–409.
Loehle, C. 1988. Philosophical tools: potential contributions to ecology. Oikos, **51**: 97–104.
Long, E.R., and Morgan, L.G. 1990. The potential for biological effects of sediment-sorbed contaminants tested in the national Status and Trends Program. NOAA Tech. Mem. NOS-OMA-52.
Long, E.R., Field, L.J., and MacDonald, D.D. 1998. Predicting toxicity in marine sediments with numerical sediment quality guidelines. Environ. Toxicol. Chem. **17**: 714–727.
Lord, C.G., Ross, L., and Lepper, M.R. 1979. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. J. Pers. Soc. Psychol. **37**: 2098–2109.
Lusted, L.B. 1968. Introduction to medical decision making. Charles Thomas, Springfield, Ill.
Lykken, D.T. 1968. Statistical significance in psychological research. Psychol. Bull. **70**: 151–159.
Meehl, P.E. 1954. Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis.
Meehl, P.E. 1965. Seer over sign: the first good example. J. Exp. Res. Pers. **1**: 27–32.
Meehl, P.E. 1967. Theory testing in psychology and physics: a methodological paradox. Philos. Sci. **34**: 103–115.
Meehl, P.E. 1973. Psychodiagnosis: selected papers. University of Minnesota Press, Minneapolis.
Meehl, P.E. 1986. Causes and effects of my disturbing little book. J. Pers. Assess. **50**: 370–375.
M'Gonigle, R.M., Jamieson, T.L., McAllister, M.K., and Peterman, R.M. 1994. Taking uncertainty seriously: from permissive regulation to preventative design in environmental decision making. Osgoode Hall Law J. **32**: 99–169.
Mohr, L.B. 1990. Understanding significance testing. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–073. Sage Publications, Newbury Park, Calif.
Morrison, D.E., and Henkel, R.E. 1970. The significance test controversy. Aldine, Chicago.
Neyman, J., and Pearson, E. 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Tran. R. Soc. Lond. **231**: 289–337.
Nisbett, R.E., and Ross, L. 1980. Human inferences: strategies and shortcomings of social judgment. Prentice-Hall, Englewood Cliffs, N.J.
Nisbett, R.E., and Wilson, T.D. 1977. Telling more than we can know: verbal reports on mental processes. Psychol. Rev. **84**: 231–259.
Nisbett, R.E., Borgida, E., Crandall, R., and Reed, H. 1976. Popular induction: information is not necessarily informative. *In* Cognition and social behavior. *Edited by* J.S. Carroll and J.W. Payne. Erlbaum, Hillsdale, N.J. pp. 227–236.
Nunnally, J. 1960. The place of statistics in psychology. Educ. Psychol. Meas. **20**: 641–650.
Oskamp, S. 1965. Overconfidence in case-study judgments. J. Consult. Psychol. **29**: 261–265.
Oskamp, S. 1967. Clinical judgment from the MMPI: simple or complex. J. Clin. Psychol. **23**: 411–415.
Pielou, E.C. 1984. The interpretation of ecological data. John Wiley & Sons, New York.
Pitz, G.F., and Sachs, N.J. 1984. Judgment and decision: theory and application. Annu. Rev. Psychol. **35**: 139–163.
Platt, J.R. 1964. Strong inference. Science (Washington D.C.), **146**: 347–353.
Pollard, P., and Richardson, J.T.E. 1987. On the probability of making Type I errors. Psychol. Bull. **102**: 159–163.
Poore, G.C.B., and Mobley, M.C. 1980. Canonical correlation analysis of marine macrobenthos survey data. J. Mar. Biol. Ecol. **45**: 37–50.
Press, S.J. 1989. Bayesian statistics: principles, models, and applications. John Wiley & Sons, New York.
Preston, T. 1981. The clay pedestal. Madrona Publishers, Seattle, Wash.
PTI. 1988. The apparent effect threshold. Briefing report to the EPA Science Advisory Board. PTI Environmental Services, 15375 SE 30th Place, Suite 250, Bellevue, WA 98007, U.S.A.
Punt, A.E., and Hilborn, R. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish Biol. Fish. **7**: 35–63.
Raiffa, H. 1968. Decision analysis. Random House, New York.
Rand, R.R. 1996. Statistics for the social sciences. Academic Press, San Diego, Calif.
Rand, R.R. 1997. Introduction to robust estimation and hypothesis testing. Academic Press, San Diego, Calif.
Rand, R.R. 1998. How many discoveries have been lost by ignoring modern statistical methods? Am. Psychol. **53**: 300–314.
Reckhow, K.H. 1990. Bayesian inference in non-replicated ecological studies. Ecology, **71**: 2053–2059.
Reckhow, K.H. 1994. Importance of scientific uncertainty in decision-

making. Environ. Manage. **18**: 161–166.
Reckhow, K.H., and Chapra, S.C. 1983. Confirmation of water quality models. Ecol. Modell. **20**: 113–133.
Rhoads, D.C., and Germano, J.D. 1982. Characterization of organism–sediment relations using sediment profile imaging: an efficient method of remote ecological monitoring of the seafloor [REMOTS^{TM} System]. Mar. Ecol. Prog. Ser. **8**: 115–128.
Rhoads, D.C., and Germano, J.D. 1986. Interpreting long-term changes in benthic community structure: a new protocol. Hydrobiologia, **142**: 291–308.
Rohlf, F.J., and Sokal, R.R. 1969. Statistical tables. W.H. Freeman & Co., San Francisco.
Rose, K.A., and Smith, E.P. 1992. Experimental design: the neglected aspect of environmental monitoring. Environ. Manage. **16**: 691–700.
Ross, L., Lepper, M.R., Strack, F., and Steinmetz, J. 1977. Social explanation and social expectation: effects of real and hypothetical explanations on subjective likelihood. J. Pers. Soc. Psychol. **35**: 817–829.
Rousseau, D.L. 1992. Case studies in pathological science. Am. Sci. **80**: 54–63.
Rozeboom, W.W. 1960. The fallacy of the null hypothesis significance test. Psychol. Bull. **57**: 416–428.
Sawyer, J. 1966. Measure and prediction, clinical and statistical. Psychol. Bull. **66**: 178–200.
Schlaifer, R. 1961. Introduction to statistics for business decisions. McGraw–Hill, New York.
Schmitt, S.A. 1969. Measuring uncertainty: an elementary introduction to Bayesian statistics. Addison–Wesley, Reading, Mass.
Schwartz, W.B., Gorry, G.A., Kassirer, J.P., and Essig, A. 1973. Decision analysis and clinical judgment. Am. J. Med. **55**: 459–472.
Shweder, R.A. 1977. Likeness and likelihood in everyday thought: magical thinking in judgments about personality. Curr. Anthropol. **18**: 637–648.
Slovic, P., and Fischhoff, B. 1977. On the psychology of experimental surprises. J. Exp. Psychol. **3**: 544–551.
Slovic, P., and MacPhillamy, D. 1974. Dimensional commensurability and cue utilization in comparative judgment. Organ. Behav. Hum. Perf. **11**: 172–194.
Smedslund, J. 1963. The concept of correlation in adults. Scand. J. Psychol. **4**: 165–173.
Spies, R.B. 1989. Sediment bioassays, chemical contaminants and benthic ecology: New insights or just muddy water? Mar. Environ. Res. **27**: 73–75.
Stearns, S.C. 1976. Life-history tactics: a review of the ideas. Q. Rev. Biol. **51**: 3–47.
Stevens, S.S. 1971. Issues in psychophysical measurement. Psychol. Rev. **78**: 426–450.
Summers, D.A., Taliaferro, D.J., and Fletcher, D.J. 1969. Subjective vs. objective description of judgment policy. Psychon. Sci. **18**: 249–250.
Toft, C.A., and Shea, P.J. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am. Nat. **122**: 618–625.
Tversky, A., and Kahneman, D. 1971. Belief in the law of small numbers. Psychol. Bull. **76**: 105–110.
Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: heuristics and biases. Science (Washington D.C.), **183**: 1124–1131.
Tversky, A., and Kahneman, D. 1978. Causal schemata in judgments under uncertainty. *In* Progress in social psychology. *Edited by* M. Fishbein. Erlbaum, Hillsdale, N.J. pp. 49–72.
Tyler, R.W. 1931. What is statistical significance? Educ. Res. Bull. **10**: 115–118.
USEPA/USACE. 1991. Evaluation of dredged material proposed for ocean disposal (testing manual). U.S. Environmental Protection Agency, Office of Water (WH-556F) and Department of the Army, U.S. Army Corps of Engineers. EPA-503/8-91/001.
USEPA/USACE. 1998. Evaluation of dredged material proposed for discharge in waters of the U.S. Testing manual — inland testing manual. U.S. Environmental Protection Agency/U.S. Army Corps of Engineers. EPA-823-B-94-002. U.S. Environmental Protection Agency, Office of Water (4305), Washington, D.C.
Ward, W.C., and Jenkins, H.M. 1965. The display of information and the judgment of contingency. Can. J. Psychol. **19**: 231–241.
Wickelgren, W.A. 1974. How to solve problems. Elements of a theory of problems and problem solving. Freeman, San Francisco.
Wiggins, J.S. 1973. Personality and prediction: principles of personality assessment. Addison–Wesley, Reading, Mass.
Wiggins, J.S. 1981. Clinical and statistical prediction: Where are we and where do we go from here? Clin. Psychol. Rev. **1**: 3–18.
Williams, W.T., Clay, H.J., and Bunt, J.S. 1982. The analysis, in marine ecology, of three-dimensional data matrices with one dimension of variable length. J. Exp. Mar. Biol. Ecol. **60**: 189–196.
Winkler, R.L. 1967. The assessment of prior distributions in Bayesian analysis. J. Am. Stat. Assoc. **62**: 776–800.