SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 16, No. 6, pp. 1322–1329, 1997
Printed in the USA
0730-7268/97 $6.00 + .00

*Hazard/Risk Assessment*

# REVISED APPROACH TO TOXICITY TEST ACCEPTABILITY CRITERIA USING A STATISTICAL PERFORMANCE ASSESSMENT

GLEN B. THURSBY,*† JAMES HELTSHE‡ and K. JOHN SCOTT§
†U.S. Environmental Protection Agency, Atlantic Ecology Division, 27 Tarzwell Drive, Narragansett, Rhode Island 02882
‡Department of Computer Science and Statistics, University of Rhode Island, Kingston, Rhode Island 02881, USA
§Science Applications International Corporation, Environmental Testing Center, 165 Dean Knauss Drive, Narragansett, Rhode Island 02882, USA

(*Received* 12 *August* 1996; *Accepted* 31 *October* 1996)

**Abstract**—Current acceptability requirements for toxicity tests are often more restrictive than necessary. They focus primarily on response in a control and generally ignore what a test was ''designed'' to detect as a significant difference from that control. An approach is presented that takes into account the performance of an entire test and the magnitude of the deviation from the current acceptability requirements. The procedure is based on analyzing the past statistical performance of a test method (i.e., what kind of difference from the control was the test designed to detect). It takes into account *traditional* control acceptance criteria, but adds a requirement for selecting a difference from the control desired to be detected as statistically significant (a threshold value). Choice of statistical procedure is not relevant to the approach. The proposed method allows a sliding scale of acceptance. The greater the deviation of mean control response below current requirements, the less likely a test is to be accepted. An example is presented using data from a 10-d sediment test using the marine amphipod *Ampelisca abdita.* Use of the proposed acceptability criterion will reduce the frequency of required retesting without sacrificing defensibility of data. Using the old acceptability criterion, 19% of the samples in the amphipod data set would require retesting. The proposed criterion reduces the potential percentage of retests to 9%.

**Keywords**—*Ampelisca* Amphipod *t* Test Statistical variance Test acceptability

## INTRODUCTION

Often the first question at the conclusion of a toxicity test is, ''Is the material toxic?'' followed closely by, ''How good are the data?'' Beyond looking at performance controls (and occasionally at reference toxicant data), too little time is spent pondering the answer to the second question. However, giving a more thorough answer is simple. The more thorough the answer to the second question, the more defensible the answer is to the first. This paper presents a straightforward way to evaluate the acceptability of toxicity test data using a summary of that test's past statistical performance. This approach not only incorporates past performance, but also uses all of the data from a test run (not just the control data) in the evaluation. Current acceptance criteria for toxicity tests are not eliminated. The acceptance criteria essentially are just reexpressed in terms of statistical variance. Using our approach, however, helps solve two problems with existing methods: (1) current test acceptance criteria are unnecessarily restrictive and (2) statistical significance is often substituted for scientific judgment.

A large, multiyear survival data set from a single laboratory for a 10-d sediment toxicity test using the marine amphipod *Ampelisca abdita* is presented as an example. Amphipod toxicity test data are used to help decide if sediment samples are toxic or not when compared with a performance control or a reference sample. In the past several years, the use of the marine amphipod *A. abdita* in this procedure has increased for both environmental surveys and regulatory decisions. Information on expected test performance with this species will be valuable for evaluating the acceptability of future tests. Ninety percent of the 767 sediment samples used in the data set were from large sediment toxicity surveys conducted by either the U.S. Environmental Protection Agency (1990 through 1992 Environmental Monitoring and Assessment Program in the Virginian Province, USA) or the National Oceanic and Atmospheric Administration (status and trends surveys in Tampa Bay, FL, USA; and Hudson Estuary and Long Island Sound, NY, USA). The remainder of the samples were from smaller surveys at various locations throughout the east coast of the United States. The statistical performance summaries are intended as an example of how to review existing data to derive acceptability parameters for a toxicity test method. An evaluation of how well test data meet the underlying assumptions for parametric hypothesis testing (i.e., normality of residuals and homogeneity of variance) is included. Statistical procedure, alpha and beta levels, and detection threshold selected were those actually used for almost all of the 767 samples in the data set. There are certainly other ways (perhaps even some that are better) to analyze the data. Choice of statistical assumptions and of statistical procedures, however, is not relevant to the approach. A detailed analysis is presented only to demonstrate the process of establishing criteria for test acceptability no matter what toxicity test or statistical procedure is chosen.

* To whom correspondence may be addressed.

Contribution 1732 of the U.S. Environmental Protection Agency, Narragansett, Rhode Island. Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency (EPA), it has not been subjected to EPA review. Therefore, it does not necessarily reflect the views of the EPA.

*Problem 1: Current toxicity test acceptance criteria are too restrictive*

Currently, no standard method to set acceptance criteria for toxicity tests exists. Acceptance criteria usually are based on the experiences of individuals who have participated in the development of culture or toxicity test methods for the particular species being used. This process is by its nature qualitative and subjective. Little room is allowed for flexibility based on the magnitude of the excursion from the acceptance criteria. For example, the *A. abdita* toxicity test method typically requires 90% survival in the controls [1]. Strict application of this criterion means that test results with 89% survival should be declared unacceptable no matter what the results from the samples being tested. Often a hard line is taken and retesting required when it may not be necessary.

*Problem 2: Statistical significance is given precedence over judgement*

Determining significant differences should require the application of judgment. Clearly, the type of significance of concern must be defined. To some people, biological significance should be the only concern. To others any difference, if it is statistically significant, is important. Biological significance is difficult to define because of the complexities of the world in which such a standard is applied. Statistical significance, on the other hand, is simple to interpret, and is the the more familiar type of significance. Typically, an experiment is conducted with replicates, measurements are made, and a statistical analysis is done to decide if a ''statistically significant difference'' between a treatment(s) and control exists. Judgment is usually not involved, because the numbers ''are what they are.'' Either there is a statistically significant difference or not.

Perhaps statistical significance is relied upon so heavily because it seems a straightforward way for making decisions. However, relying just on statistical significance in a single experiment to decide toxicity is insufficient. A particularly good run with a particularly good batch of organisms will detect a particularly small difference as significant. On the other hand, a test that is unusually variable will require a large difference from the control before statistical significance is detected (lack of statistical significance could be a ''reward'' for sloppy procedures). In both of the above situations the decision for or against toxicity would be statistically correct. If a test is repeated, however, what would be the likelihood of drawing the same conclusion about significance? Judgment concerning the repeatability of a method must be included in the decision-making process. This paper addresses this concern. This repeatable significance can be called ''detectable significance.''

*Why run statistical tests?*

The purpose of a statistical test is not to find statistical significance. The purpose is to make a decision about the real world based on a small piece of that world. An amphipod sediment test is used to make an estimate about the existence of a real difference in toxicity between sediments from control (or reference) environments and from test environments, based on one or more subsamples. Because the real answer is never known with absolute certainty, some degree of error is always associated with a statistically based decision (Fig. 1). Two types of errors are possible (type I and type II), and although both are important, they are usually not given equal consideration in a toxicity test. Samples showing no statistical toxicity are usually ignored, with preference being given to those declared toxic. Thus, type II errors (false negatives) are often missed, although the risk of a false positive or false alarm still exists (Fig. 1). However, false positives are likely to be discovered because they are associated with environments that receive further attention, and often additional testing [2].

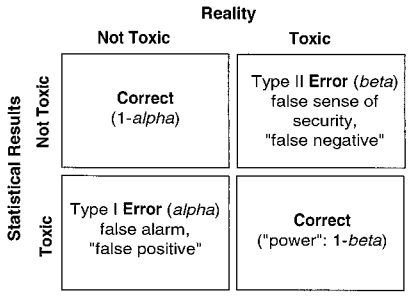


Fig. 1. Summary of the relationship between statistical results and reality.

A more thorough consideration of toxic samples would be one in which we were concerned about locations that are different in reality. This requires focusing on, rather than ignoring, type II errors [2–5]. In other words, for a given difference between sample and control means, we need to understand the probability that the test method could have detected such a difference as significant. This is important because false negatives can be more costly than false positives [2,5]. False negatives may create a false sense of security, allowing continued environmental degradation to occur. If we concern ourselves only with locations that are statistically toxic, a false negative could go on for years before being discovered.

*Detectable significance*

Detectable significance deals with the probability that a test method can statistically detect a difference (i.e., the power of the test). Differentiating detectable significance from statistical significance is easy because statistical significance is associated with a single application of a toxicity test method (a test run), whereas detectable significant is a property of the test method itself. After a single test is run, a decision will be made about the toxicity of a sample (based on statistical significance). The concern is no longer with the *probability* that the run produced a correct result. The decision, once made, is either 100% right or 100% wrong. Which situation is true can only be determined by knowing the correct answer, which of course is never known. When dealing with an evaluation of detectable difference of a test method, *before* an individual test run is conducted, the concern is with how likely the method will be to detect a meaningful difference if it exists. Type I and type II errors are a property of a method, not one run of a test. An analogous example is a diagnostic human health procedure. Home pregnancy tests are designed to decide the presence or absence pregnancy. The tests have an associated probability of false positives and false negatives. These are characteristics of the tests. Despite what a test shows when administered, the person using the test is either 100% pregnant or 100% not. After a test is administered, one does not ask, ''What is the probability I am pregnant (or not)?''. A decision for further action will be made based on the results of the test. The decision on which the action is based is either 100% right

or 100% wrong. The manufacturer designs the tests to be as accurate as possible when pregnancy exists, and when pregnancy does not exist. These are the reliability probabilities (type I and type II errors) for the home test, but they only apply to the results of the test, not to the decision based on those results.

Detectable significance incorporates historical data, allowing judgment concerning the repeatability of the test method (empirical consideration of multiple test runs). Historical data can come from repeated experiments in the same laboratory (as in the current study), or from multiple laboratories. The basic principle is to make a judgment about what level of difference between a control and treatment the test was designed to detect. Among-test variability is incorporated into the determination of significant difference. Detectable significance attempts to define what level of variability among test runs is typical or to be expected. This reduces the risk of making decisions (based on a single test run) that may be over- or underprotective due to data that are too ''tight'' or too variable, respectively. Although variability can be described with statistics, statistics cannot tell whether that amount of variability is good or bad. Statistics are tools and nothing more [6], and cannot substitute for judgment. Detectable significance requires judgment.

### *Recommended change*

Our recommended change to acceptance criteria is based, in part, on an empirical database of minimum significant differences (MSDs). The inclusion of MSD limits as additional acceptance criteria for toxicity tests has begun recently [7,8]. This additional criterion is an improvement over currently used criteria because it uses all of the data (not just that of the controls) in making a final decision concerning acceptability of test data. However, the addition of an MSD limit still maintains the absolute requirements placed on control responses. The approach we propose takes the use of MSDs a step further, and justifies a more flexible attitude toward control data. This approach also takes into account the overall performance of the entire test, not just that of the performance control. The acceptance criterion is expressed in terms of statistical variance rather than survival. In addition, the criterion requires a judgment on the difference from the control that is to be considered important. Survival is used as an example for an endpoint, and a *t* test as an example of the statistical procedure. However, the approach can be applied to most measurements (including chemistry data) and many other statistical procedures. Briefly, the proposed changes in the acceptability criterion take two simple steps. The first step involves the selection of a threshold value that represents the magnitude of response it is important to be able to detect as different from the control. The second step just compares the MSD to the threshold value (more precisely the mean control value minus the threshold value). If the test results allow the threshold value to be declared significantly different from the control, then the test should be considered acceptable. The example presented in the present paper demonstrates how a threshold was selected for *A. abdita* and the consequences of applying that threshold as a criterion for test acceptability. Other threshold values could be used, but the methodology for determining an acceptable test run would remain the same.

## MATERIALS AND METHODS

### *Toxicity test*

Amphipods in each test series were exposed to sediments for 10 d under static conditions using standardized procedures [1,9]. All tests were conducted in the same laboratory. Each test series consisted of a performance control (sediment from central Long Island Sound, CT, USA) and a series of environmental samples (generally 10–15). Approximately 200 ml of control or test sediment was placed into each 1-quart (0.9-L) exposure container and covered with approximately 600 ml of seawater. All seawater (28–32‰) came from lower Narragansett Bay, Rhode Island, USA. *Ampelisca abdita* were used for all tests and were collected from tidal flats in the Pettaquamscutt (Narrow) River, a small estuary flowing into Narragansett Bay. Animals were fed the laboratory-cultured diatom *Phaeodactylum tricornutum* before testing, and were not fed during tests. Twenty subadult animals were added per replicate. The animals were randomly assigned to exposure chambers and the chambers were randomly assigned positions within a 20°C water bath. At the end of a test series, animals were sieved from each replicate using a 0.5-mm mesh stainless-steel screen. All recovered amphipods were scored as living or dead. Missing test organisms were presumed to have died and decomposed during the test. Results from each replicate were expressed as percentage survival. Data from all tests for which control survival was at least 80% were included in the analyses (approx. 90% of the test runs during the time covered).

### *Data analysis*

Data from 767 sediment samples and 63 controls were used to test normality of residuals and homogeneity of variance. Normality and homogeneity of variance were tested using both untransformed (percentage survival) and transformed (arcsine of the square root) data. Minimum significant differences were calculated with only the 637 sediment samples for which five replicates were used (the standard for the method). Survival data (untransformed only) from each of these sediments were compared with the appropriate performance control using a one-way unpaired *t* test, assuming unequal variance. One-way analysis was used because we were only concerned with whether or not a sample had amphipod survival less than that of the performance control. The assumption of unequal variances is recommended by Moser and Stevens [10]. The statistical error associated with this assumption when variances are in reality equal is much less than that associated with assuming equal variances when in reality the variances are unequal. Analysis of variance (ANOVA) was not used because samples were considered independent. Under these conditions multiple comparison procedures (e.g., *t* tests) are appropriate [11].

Rank plots for both control and sample variances were created for all of the sediment sample and control results. Variances were ranked from lowest to highest and then numbered from one through the highest number. Variances were then plotted as their rank expressed in percent to facilitate marking off various percentiles.

The equation for calculating a *t* value assuming unequal variances was taken from Moser and Stevens [10]. The calculated *t* value was compared with an interpolated critical *t* value from a standard statistical table ($\alpha = 0.05$) at the appropriate degrees of freedom. Empirical MSDs were calculated by solving the original equation for MSD and substituting the

table critical value for $t$ at the appropriate degrees of freedom ($n = 5$ and $\alpha = 0.05$). The resulting equation is

$$\text{MSD} = (\overline{x_1} - \overline{x_2}) = t_{\text{critical}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

$\overline{x_1} - \overline{x_2}$ = difference between mean values for treatments 1 and 2 (e.g., sample and control)

$t_{\text{critical}}$ = $t$ value from standard statistical table, for $\alpha = 0.05$ and the appropriate degrees of freedom

$s_1^2$, $s_2^2$ = variances for treatments 1 and 2

$n_1$, $n_2$ = numbers of replicates for treatments 1 and 2

The calculated MSDs were expressed as a percentage of the control response. An empirical curve showing the probability of rejection of the null hypothesis (e.g., probability of declaring sample and control statistically different) was constructed as a cumulative frequency curve of the MSDs. Theoretical power curves ($\alpha = 0.05$ and $n = 5$ for both control and sample) were calculated to compare with the empirically derived curve. These curves used the equation presented by Alldredge [12] for optimal sample size. Instead of solving for optimal sample size, a fixed sample sized was substituted ($n = 5$), $\alpha$ set at 0.05, and the equation solved for $\beta$ using a range of MSDs. Power was calculated as $1 - \beta$. All that was needed was an estimate of the variance for the mean survivals. Theoretical curves were calculated using 10th, 25th, 50th, 75th, and 90th percentiles of the variances of all samples tested.

### *Selection of threshold value for detectable significance*

The approach for test acceptability recommended in this paper requires the selection of a cutoff, or threshold value that represents a meaningful difference from the control that the test should be able to detect as significant. For the *A. abdita* test method, selection was based on two existing pieces of information: the desired minimum acceptable mean control response, and a difference from the control that is meaningful as significant. For the 10-d sediment test using *A. abdita* the minimum mean control survival used was 90% [1], and 80% of the control was the meaningful difference [13]. A threshold of 72% survival was arrived at by taking 80% of the 90% minimum control survival.

## RESULTS AND DISCUSSION

### *Statistical assumptions*

Data should meet the underlying assumptions for the particular statistical test being used (e.g., normality and homogeneity of variance) in order to minimize errors associated with statistical procedures. The main assumption for the $t$ test (assuming unequal variance) used with the *A. abdita* data is that the residuals (replicate values standardized by subtracting the mean value with which they are associated) are normally distributed. A potential problem with typical toxicity test data is that the total number of replicates ($n$) is often too small to give much confidence in statistical conclusions concerning issues of meeting underlying statistical assumptions. In this study this problem was addressed by combining results from a large number of test series. Data are presented in Figure 2 as frequency histograms of the residuals for all 767 test sediments and 63 control samples. The residuals are presented for data expressed as percentage survival (top) and the arcsine square-root transformed data (bottom). The histogram of the residuals from the transformed data is similar to that of the untransformed data. Based on the shape of the histograms, normality was a reasonable assumption with or without the arcsine square-root transformation. Formal statistical tests of normality of residuals are not necessary, as small departures from normality have little effect on the $t$ test, and with 767 sets of residuals, we are very likely to detect small departures from normality.

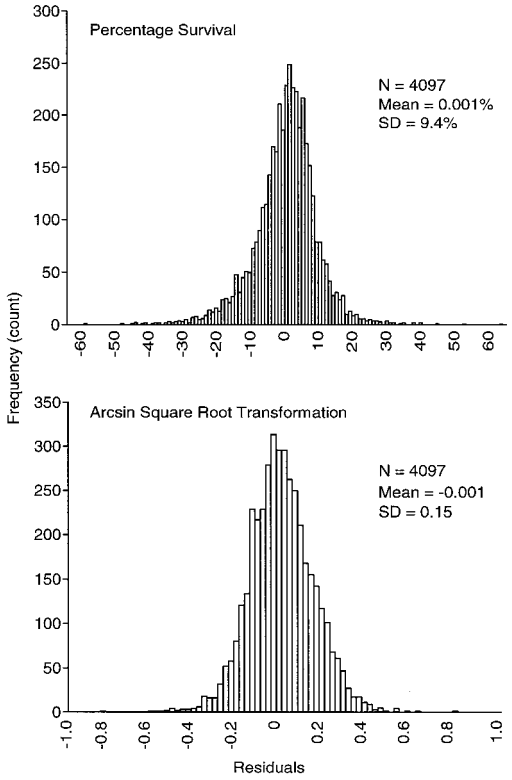


Fig. 2. Frequency histograms for residuals from 10-d solid-phase toxicity tests using the marine amphipod *Ampelisca abdita.* Sediments are included from 767 test samples and 63 controls. The top chart represents the survival data expressed as a percentage of the control. The bottom chart represents the same data, but after they were transformed by taking the arcsine of the square root.

Unlike the normality plots, variance plots were different between untransformed and transformed survival data. Box and whisker plots of means (includes treatments and controls) versus statistical variance are shown in Figure 3. Again, both percentage survival (top) and arcsine square-root transformed data (bottom) are shown. Variance tended to be lower at mean survivals below 20% and above 80%. Transformation helps even out the extreme values (those less than 20% and those greater than 80%). Although a wide range of values existed in the middle of both the untransformed and transformed plots, the highest and lowest values increased with the transformation (as expected). If a statistical procedure is used that assumes homogeneity of variance (e.g., ANOVA), then transforming the data is recommended. However, because the $t$ test for the *A. abdita* data in this paper assumed unequal variance, all subsequent analyses are presented with untransformed data only.

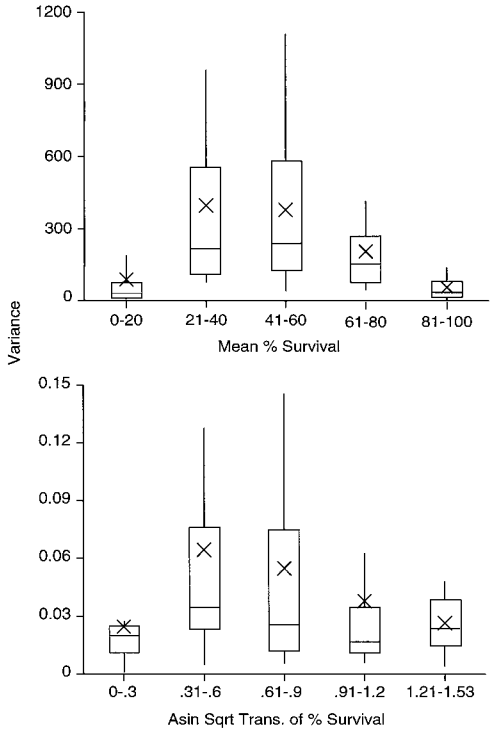


Fig. 3. Box and whisker diagrams showing the summary of the distributions of variances relative to mean values. The top chart represents the survival data expressed as a percentage of the control and the bottom presents the transformed data. The ranges of the two sets of means were divided into five equal parts and each part was summarized with a separate box and whisker plot. The mean variance for each data subset is represented by an ×. The horizontal line within each box is the 50th percentile; the lower and upper limits of the boxes are the 25th and 75th percentiles, respectively; and the limits of the vertical lines are the 10th and 90th percentiles.

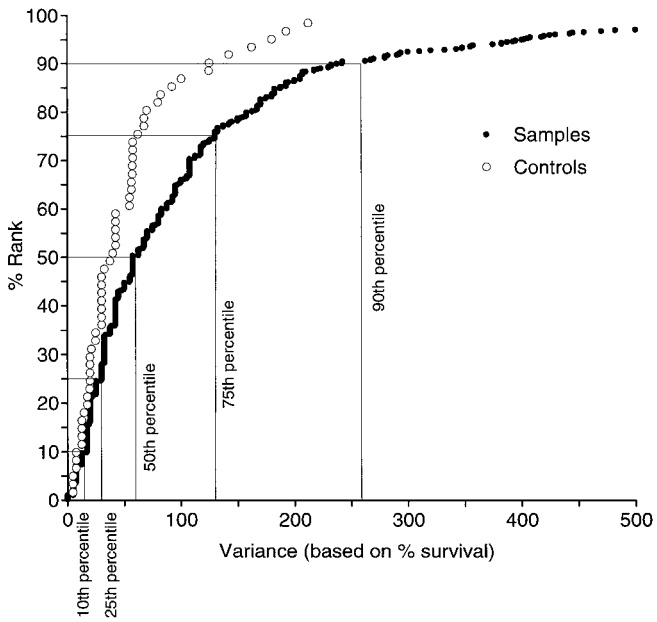


Fig. 4. A percent rank plot of the variances for the controls and the samples. The variance data were truncated at 500. One control variance and 19 sample variances were greater than 500. To create the plot, the variances were ordered from lowest to highest and each was given a rank (expressed as a percentage of the highest rank). For samples, only values from tests with five replicates were used (a total of 637 for the samples, and 63 for the controls). The 10th, 25th, 50th, 75th, and 90th percentiles are marked for the sample variances.

### *Power curves*

Traditional criteria of acceptability (i.e., those that only emphasize control response) do not eliminate all of the potential problems that might arise in interpreting test data. Situations can exist in which either small differences occur that are statistically different from the control or large differences occur that are not statistically different from the control. Either of these situations brings the reliability of a conclusion concerning the toxicity of a sample into question. The variability among control and sample replicates (i.e., statistical variance) will have the greatest influence on whether either of the above two situations will occur. Thus, one way to examine if a difference is too small or too large is to look at the historical variances. Figure 4 is a rank plot using the variances for both the controls and samples from this study. The 10th, 25th, 50th, 75th, and 90th percentiles for sample variances are marked. Control variances were generally less than sample variances. This would be expected because the controls all averaged survivals greater than 80%. Plots such as those in Figure 4 could be used to make judgments concerning possible outliers. For example, care should be taken when interpreting data whose variance lies outside the 10th and 90th percentiles. The greater the variance in the data the greater the probability of making a type II error (declaring a sample nontoxic when in reality toxicity exists). The smaller the variance in the data, the greater the possibility of a type I error (declaring a sample toxic when in reality it is not).

Power is inversely related to variance. As we stated earlier, the purpose of a toxicity test is not to find statistical differences; it is to decide (with an acceptable degree of uncertainty) whether a sample is toxic or not. An ''acceptable'' degree of uncertainty is based on judgment. By convention and years of statistical teaching, 5% has become the standard as the type I error rate ($\alpha$). However, because type II error rates ($\beta$) are often ignored, a clear consensus does not exist on a standard $\beta$, although 20% has been used by some [12]. Although there may not be agreement on an acceptable $\beta$ error, it can be described for a given test method. One easy way to show $\beta$ error is through a graphic representation of statistical power ($1 - \beta$). Figure 5 shows calculated theoretical power curves using the 10th, 25th, 50th, 75th, and 90th percentile variances. Superimposed on Figure 5 is an empirical curve showing the probability of rejection of the null hypothesis (e.g., probability of a statistical difference) for the *A. abdita* 10-d sediment toxicity test created from the cumulative frequency of calculated MSDs ($\alpha = 5\%$ and $n = 5$). The empirical curve falls between the 50th and 75th percentile curves.

Power curves can be used in several ways. First, if a fixed error rate is needed (e.g., $\beta = 20\%$), then the power curve shows that a 15% difference from the control mean response is the minimum difference that should be considered as truly significant from that control (empirical curve, Fig. 5). This does not mean that the statistical test cannot detect a difference less than 15%, it means that any detected difference less than 15% is not as repeatable. Alternately, a judgment can be made concerning what difference from the control is important to be determined as significant (e.g., 20%). From the power curve the $\beta$ error associated with a 20% difference from the control is less than 10% (i.e., power $> 90\%$). The *A. abdita* test has a high power to detect this size of difference from the control as significant.

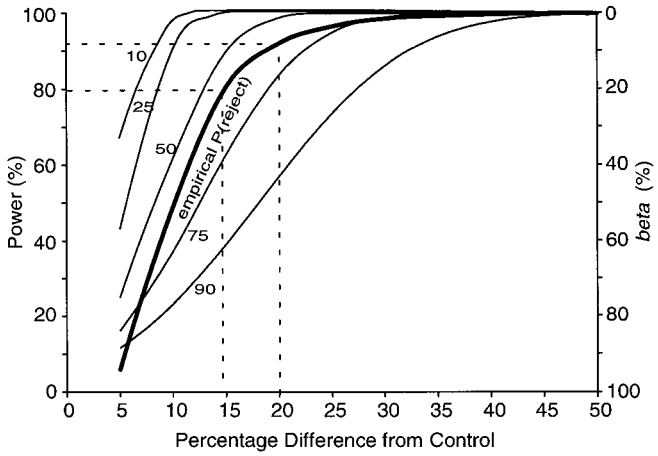


Fig. 5. Empirical probablity of rejection of the null hypothesis and theoretical power curves. The empirical curve was created as a cumulative frequency curve for calculated minimum significant differences ($\alpha = 0.05$) for all of the tests for which there were five replicates for both the control and sample (a total of 637). The empirical curve has the same *y*-axis as power. The other curves are theoretical power curves ($n = 5$ and $\alpha = 0.05$) using the 10th, 25th, 50th, 75th, and 90th percentile sample variances from Figure 4. The inner dashed line shows the minimum percentage difference from a control that one can expect to be detectable for a desired type II error of 20% (80% power). The outer dashed line shows the power of the method to detect an amphipod survival 20% less than the control response.

Power curves are very useful in experimental design and show the ''error'' consequences associated with judgments concerning various differences from the control (i.e., various thresholds). Returning to the two situations concerning the reliability of either small differences that are significantly different from the control or large differences that are not, a power curve can be used to help decide if those conclusions are reliable. For example, the power to detect differences drops quickly below 15%, therefore care should be taken when declaring samples less than 15% different from the control as toxic. Likewise, there are occasions in which large differences (e.g., >40%) from the control are not statistically significant. This happened about 1% of the time in the *A. abdita* data set, because of an occasional high sample variance. However, the power curve shows that the test procedure should declare this size of a difference significant with great confidence (power approx. 99%). Therefore, care should be taken when declaring samples to be not toxic when differences from the control exceed 20 to 25%.

### *Proposed changes to acceptance criteria*

One way to deal with potential problems of too small or too large differences is to incorporate power curves and other uses of statistical variance into criteria for acceptable tests. As stated above, the biological criteria for accepting results from many types of toxicity tests use primarily responses associated with the control. These criteria are often absolute, leaving little room for judgment. The approach described below takes into account the performance of the entire test and would represent a more flexible scale of acceptance, although this flexibility is based on the original absolute criterion. The criterion is not changing as much as are the units in which it is expressed. Instead of percentage survival (as is the case now with *A. abdita*) the acceptance criterion is based on statistical variance, or, more precisely, on MSDs.

The approach requires two judgments before test initiation. The first is a decision as to the desired minimum acceptable control response. With a standardized test this is usually stated in the test protocol. For the 10-d solid-phase test using *A. abdita* this minimum acceptable survival is 90% overall with a no less than 80% survival in any one replicate [9]. The second judgment is a decision on what difference from the control is desired to be detected as significant. Empirical or theoretical power curves can be used as an aid in making this decision. Eighty percent of the control has been used with the amphipod test [13].

In our example, these two judgments can be combined and restated as: acceptable results are any that can declare 72% survival as significantly less than the control. This number, or threshold, is 80% of the minimum mean control survival of the original acceptance criteria (90%). The exact method of determining a threshold, however, is not as important as having such a threshold. Once a threshold for significance is determined, a means exists for including all of a test's data in the acceptability decision for that test. Thus, any test results that allow 72% survival in a sample to have been declared as significantly less than the control should be acceptable. In other words, the MSD must be less than or equal to the control mean survival minus the threshold. Acceptability will be dependent on the magnitude of the deviation from the control, and the statistical variance associated with the control and the sample results. This technique will allow some tests with controls less than the desired overall minimum to be acceptable. It also minimizes specific requirements for individual control replicate response, relying instead on control and sample variances. In addition, as will be shown with an example below, the greater the control deviates from (is less than) the originally desired minimum, the less likely it is that a test will be acceptable. Thus, there is a sliding scale of acceptability that is a function, in part, of the magnitude of the control response.

Examples are presented in Figures 6 and 7. It is not necessary to create such plots to determine acceptability of tests. The plots are presented only to demonstrate the interdependence of sample and control variance in determining test acceptability. Selecting a threshold value is the only requirement for using the proposed method of determining acceptable tests. In Figure 6 the threshold value is held constant at 72% survival. Plots are shown for mean percentage control survivals of 80, 85, 90, and 95%. For a given control mean survival and control variance the plot shows the maximum sample variance that allows 72% survival to be declared statistically less than that control. The figure clearly shows the sliding scale for accepting tests. As the control mean for survival decreases, the constraints on the variances get tighter. The likelihood of accepting a test with an 80% control survival is much less than one at 90%. If control variances for 80 and 90% survival were both zero, the maximum sample variances ''allowed'' would be approximately 75 and 375, respectively. Sample variances of 375 or less occurred greater than 90% of the time (Fig. 4), whereas variances 75 or less only approximately 55% of the time. The restrictions at 80% control survival become even greater when considering the fact that zero control variance is not a likely event. Figure 7 shows what would happen to the constraints on accepting a test if the threshold value was changed. Under certain scenarios a more conservative threshold value may be desired. Other uses may be concerned only with large differences relative to the control. Each plot in Figure 7 is for an average control mean percentage survival

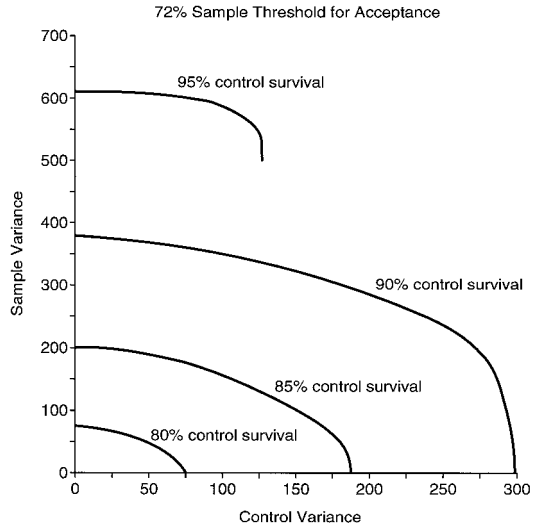


Fig. 6. Demonstration of the sliding scale for test acceptance based on control mean survival. Plots are based on an unpaired *t* test that assumes unequal variances ($\alpha$ = 5%, $n$ = 5). The threshold value (see text) is the same for all curves: 72% sample survival. Plots are shown for mean percentage control survivals of 80, 85, 90, and 95%. For a given control mean survival and control variance a plot shows the maximum sample variance that allows 72% survival to be declared statistically less than that control. The shoulder on the right side of each plot occurs partly because the two axes have different scales, and partly because there are more restrictions on the possible control variances (because of fixed mean control survivals).

of 90%. The 72% threshold plot is the same as the ''90%'' curve in Figure 6. Additional plots are shown for an 80% and a 60% cutoff. Clearly, the higher the threshold, the narrower the ranges of variances that are acceptable.

Figure 8 shows an example of the potential consequence of using the new criterion versus the current acceptance cri-

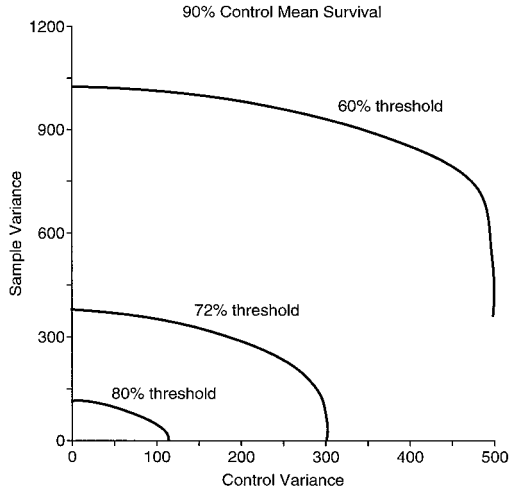


Fig. 7. Demonstration of the effect of a change in threshold value on acceptable variances. Plots are based on an unpaired *t* test that assumes unequal variances ($\alpha$ = 5%, $n$ = 5). The mean control percentage survival is the same for all curves: 90%. Plots are shown for sample thresholds (see text for explanation) of 60, 72, and 80%. For a given control mean survival and control variance a curve shows the maximum sample variance that allows the threshold value to be declared statistically less than that control. The shoulder on the right side of each plot occurs partly because the two axes have different scales, and partly because there are more restrictions on the possible control variances (because of fixed mean control survivals).

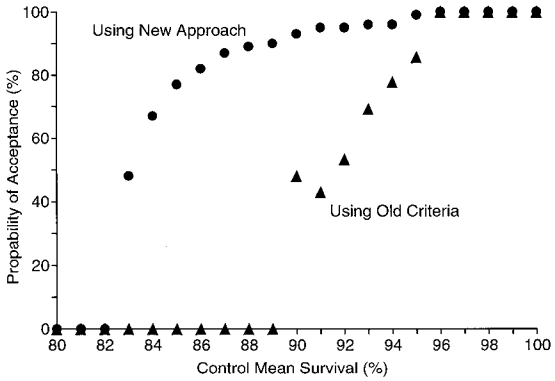


Fig. 8. An example of the potential effect of the proposed approach on the probability of accepting tests. The triangles show the probability of acceptable tests for various mean control survivals using the old control survival acceptance criteria (see text). The circles show one example of an estimate of the probability of accepting a test using the new approach with a threshold value of 72% and a control variance of 120 (the maximum for percentage control survival meeting the current criteria). Note that under these conditions, control survivals less than or equal to 82% would not be considered acceptable, even with ''zero'' sample variance.

terion based on control performance alone. Figure 8 shows two plots. One is the probability of acceptable tests for various mortalities that meet the current control survival acceptance criteria (90% average with no less than 80% in any one replicate). By examining all possibilities to distribute dead organisms among five replicates (five is the standard for the test) for a given mean control mortality, the theoretical probability of having an acceptable test can be calculated. The other plot in Figure 8 uses the newly proposed criteria with a threshold value of 72% and a control variance of 120 (the maximum for percentage control survival meeting the old acceptability limits) for 80 to 95% mean survival or the maximum possible variance for mean survivals from 96 to 100% (a variance of 120 is not possible for these means). Using the above constraints, the latter plot was constructed by first calculating the maximum possible sample variance for each control mean survival from 80 to 100%. Next, Figure 4 was used to estimate the proportion of the sample variances in the data set that were less than or equal to that maximum. This assumes that all variances in the data set have an equal probability of occurring. Because a sample variance is a function of mean survival, this assumption is not necessarily true. However, it does give some idea of the potential value of the new approach compared to the current one.

As with any acceptance criterion, judgment still must be an option in the final decision to accept or reject test results. For instance, a situation could occur in which the control mean was 85%, the MSD calculation did not allow the cut off to be significantly different from the control, but all of the sample means were high (e.g., greater than 90%). In this case these samples probably would not be declared toxic no matter what the control was, and retesting could be a waste of time and resources. Likewise, the same control situation as above could exist, with the test being technically unacceptable, but the survivals within the samples all being very low (e.g., less than 40%). It is unlikely that a conclusion about the samples being toxic would change with a retest. Judgment should never be eliminated from the process. Careful judgment in the selection of a threshold value will give an additional advantage to that value beyond determining test acceptability. If an observed

response is greater than the threshold then the sample can be assumed to be nontoxic, period. Only if the observed response is less than the threshold and not statistically significant is the sample a candidate for a retest.

If we examine the set of *A. abdita* data used in this study, we can see the value of the proposed criterion with real world data. Using the current criterion, 146 (19%) of the 767 samples tested would require retesting, even if we applied judgment and accepted samples with mean survivals greater than 90% or less than 40% survival. Switching to the proposed criterion, in our example we accept all samples from tests that would have allowed 72% survival to have been declared statistically less than the mean control response for that test. The number of samples that would be candidates for retesting is reduced to 72, approximately half that using the old criterion. However, 21 of these 72 samples would have been accepted using the old criterion. Thus, some samples that could pass under the old criterion may not pass under the proposed criterion. This is possible because the proposed criterion takes into account the results for both the control and the sample, not just the former.

Because resources will always be a limiting factor in any study, whenever possible those resources should be conserved and used toward testing new samples rather than on needless retesting of old ones. The changes recommended in this paper will reduce the need for retesting, and should not compromise the defensibility of the data. The focus of the new approach is to evaluate acceptability of data from a single test run by revising the old approach to determine acceptability of that test run. The approach is test methodology-specific and independent of the statistical procedure selected. The approach also is aimed at solving some of the problems associated with judging test acceptability. It does not address judgment that is necessary to determine if observed toxicity is toxic enough to be of concern.

*Acknowledgement*—All of the *Ampelisca abdita* survival data were provided by Science Applications International Corporation's Environmental Testing Center in Narragansett, Rhode Island, USA. We thank Walter Berry, Ann Kuhn-Hines, and Charlie Strobel for valuable reviews of earlier versions of the manuscript.

## REFERENCES

1. **U.S. Environmental Protection Agency.** 1994.Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA/600/R-94/025. Technical Report. Washington, DC.
2. **Fairweather, P.G.** 1991. Statistical power and design requirements for environmental monitoring. *Aust. J. Mar. Freshwater Res.* **42**:555–567.
3. **Toft, C.A.** and **P.J. Shea.** 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. *Am. Nat.* **122**:618–625.
4. **Andrew, N.L.** and **B.D. Mapstone.** 1987. Sampling and the description of spatial pattern in marine ecology. *Oceanogr. Mar. Biol. Annu. Rev.* **25**:39–90.
5. **Peterman, R.M.** 1990. Statistical power analysis can improve fisheries research and management. *Can. J. Fish. Aquat. Sci.* **47**: 2–15.
6. **Campbell, R.C.** 1967. *Statistics for Biologists.* Cambridge University Press, New York, NY, USA.
7. **Oris, J.T.** and **A.J. Bailer.** 1993. Statistical analysis of the *Ceriodaphnia* toxicity test: Sample size determination for reproductive effects. *Environ. Toxicol. Chem.* **12**:85–90.
8. **U.S. Environmental Protection Agency.** 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. EPA/600/R-95/136. Technical Report. Washington, DC.
9. **American Society for Testing and Materials.** 1994. Standard guide for conducting 10-day static sediment toxicity tests with marine and estuarine amphipods. E 1367-92. In *Annual Book of ASTM Standards, Volume 11.04.* Philadelphia, PA, pp. 1161–1186.
10. **Moser, B.K.** and **G.R. Stevens.** 1992. Homogeneity of variance in the two-sample means test. *American Statistician* **46**:19–21.
11. **O'Brien, P.C.** 1983. The appropriateness of analysis of variance and multiple-comparison procedures. *Biometrics* **39**:787–794.
12. **Alldredge, J.R.** 1987. Sample size for monitoring of toxic chemical sites. *Environ. Monit. Assess.* **9**:143–154.
13. **Strobel, C.J., H.W. Buffum, S.J. Benyi, E.A. Petrocelli, D.R. Refsteck** and **D.J. Keith.** 1995. Statistical summary: EMAP—Estuaries Virginian Province—1990 to 1993. EPA/620/R-94/026. Technical Report. U.S. Environmental Protection Agency, Narragansett, RI.