United States Environmental Agency
Office of Wastewater Management (4203M)
EPA 833-R-10-004 (June/2010)
http://www.epa.gov/npdes/permitbasics

National Pollutant Discharge Elimination System Test of Significant Toxicity Technical Document

June 2010

NATIONAL POLLUTANT DISCHARGE ELIMINATION SYSTEM
TEST OF SIGNIFICANT TOXICITY
TECHNICAL DOCUMENT

An Additional Whole Effluent Toxicity Statistical Approach for Analyzing Acute and Chronic Test Data

U.S. Environmental Protection Agency
Office of Wastewater Management
Water Permits Division
1200 Pennsylvania Avenue, NW
Mail Code 4203M
EPA East Building – Room 7135
Washington, DC 20460

June 2010

NOTICE AND DISCLAIMER

This document provides the technical basis for the Test of Significant Toxicity (TST) approach under the National Pollutant Discharge Elimination System (NPDES) for permitting authorities (states and Regions) and persons interested in analyzing valid whole effluent toxicity (WET) test data using the traditional hypothesis testing approach as part of the NPDES Program under the Clean Water Act (CWA). This document describes what the U.S. Environmental Protection Agency (EPA) believes is another statistical option to analyze valid WET test data for NPDES WET reasonable potential and permit compliance determinations. The document does not, however, substitute for the CWA, an NPDES permit, or EPA or state regulations applicable to permits or WET testing; nor is this document a permit or a regulation itself. The TST approach does not result in changes to EPA's WET test methods promulgated at Title 40 of the *Code of Federal Regulations* Part 136. The document does not and cannot impose any legally binding requirements on EPA, states, NPDES permittees, or laboratories conducting or using WET testing for permittees (or for states in evaluating ambient water quality). EPA could revise this document without public notice to reflect changes in EPA policy and guidance. Finally, mention of any trade names, products, or services is not and should not be interpreted as conveying official EPA approval, endorsement, or recommendation.

CONTENTS

EXECUTIVE SUMMARY xi
ACRONYMS AND ABBREVIATIONS xix
GLOSSARY xxi

1.0 INTRODUCTION 1
1.1 Summary of Current EPA Recommended WET Analysis Approaches 1
1.2 Advantages and Disadvantages of Recommended Traditional Hypothesis Testing Approach 1
1.3 Test of Significant Toxicity 4
1.4 Regulatory Management Decisions for TST 5
1.5 Document Objectives 7

2.0 METHODS 9
2.1 Test Methods and Endpoints Evaluated 9
2.2 Data Compilation 12
2.3 Setting the Test Method-Specific α Level 14

3.0 RESULTS 19
3.1 Chronic *Ceriodaphnia dubia* Reproduction Test 19
3.2 Chronic *Pimephales promelas* Growth Test 24
3.3 Chronic *Americamysis bahia* Growth Test 28
3.4 Chronic *Haliotis rufescens* Larval Development Test 32
3.5 Chronic *Macrocystis pyrifera* Germination Test 35
3.6 Chronic *Macrocystis pyrifera* Germ-tube Length Test 39
3.7 Chronic Echinoderm Fertilization Test 42
3.8 Acute *Pimephales promelas* Survival Test 45
3.9 Chronic *Selenastrum capricornutum* Growth Test 48
3.10 Acute *Ceriodaphnia dubia* Survival Test 52

4.0 SUMMARY OF RESULTS AND IMPLEMENTING TST 57
4.1 Summary of Test Method-Specific Alpha Values 57
4.2 Calculating Statistics for Valid WET Data Using the TST Approach 58
4.3 Benefits of Increased Replication Using TST 59
4.4 Applying TST to Ambient Toxicity Programs 59
4.5 Implementing TST in WET Permitting under NPDES 60
4.6 Reasonable Potential (RP) WET Analysis 62
4.7 NPDES WET Permit Limits 62

5.0 CONCLUSIONS 65

6.0 LITERATURE CITED 67

APPENDICES
A Rationale for Using Welch's t-Test in TST Analysis of WET Data for Two-Sample Comparisons
B Step-By-Step Procedures for Analyzing Valid WET Data Using the TST Approach
C Critical *t* Values for the TST Approach

TABLES

Table 1-1. Error terminology for traditional WET hypothesis methodology 2
Table 1-2. Error terminology for TST WET hypothesis methodology.....5
Table 2-1. Summary of test condition requirements and test acceptability criteria for each EPA WET test method evaluated in TST analyses 10
Table 2-2. Summary of WET test data analyzed 13
Table 3-1. Summary of mean control reproduction and control CV derived from analyses of 792 chronic *Ceriodaphnia dubia* WET tests.....19
Table 3-2. Comparison of the percentage of chronic effluent *Ceriodaphnia* tests declared toxic using TST versus the traditional hypothesis testing approach 24
Table 3-3. Summary of mean control growth and control CV derived from analyses of 472 chronic *Pimephales promelas* WET tests.....25
Table 3-4. Comparison of the percentage of chronic effluent fathead minnow tests declared toxic using TST versus the traditional hypothesis testing approach 28
Table 3-5. Summary of mean control growth and control CV derived from analyses of 210 chronic *Americamysis bahia* WET tests.....29
Table 3-6. Comparison of percentage of chronic effluent mysid shrimp tests declared toxic using TST versus the traditional hypothesis testing approach 32
Table 3-7. Summary of mean control larval development and control CV derived from analyses of 136 chronic red abalone WET tests.....33
Table 3-8. Summary of mean control germination and control CV derived from analyses of 135 chronic giant kelp WET tests 36
Table 3-9. Summary of mean control germ-tube length and control CV derived from analyses of 135 chronic *Macrocystis pyrifera* WET tests.....39
Table 3-10. Summary of mean control fertilization and control CV derived from analyses of 177 chronic *Dendraster excentricus and Strongylocentrotus purpuratus* WET tests.....42
Table 3-11. Summary of mean control survival and control CV derived from analyses of 347 acute *Pimephales promelas* WET tests 45
Table 3-12. Percent of fathead minnow acute tests declared toxic using TST and a *b* value = 0.8 as a function of percent mean effect, number of replicates (2 or 4 replicates), and different alpha or Type I error levels.....48
Table 3-13. Summary of mean control growth, CV and standard deviation derived from the analyses of all chronic *Selenastrum capricornutum* WET test data and compared with the analysis of only the chronic *Selenastrum capricornutum* WET test in which it was assumed that EDTA was added to the controls.....49
Table 3-14. Comparison of the percentage of chronic *Selenastrum* tests declared toxic using TST versus the traditional hypothesis testing approach.....52
Table 3-15. Summary of mean control growth, CV and standard deviation derived from analyses of 239 acute *Ceriodaphnia dubia* WET tests.....52

Table 3-16. Percent of *Ceriodaphnia dubia* acute tests declared toxic using TST and a *b* value = 0.8 as a function of percent mean effect, number of replicates (4 or 6 replicates), and different alpha or Type I error levels 55

Table 4-2. Comparison of results of chronic *Ceriodaphnia* ambient toxicity tests using the TST approach and the traditional t-test analysis. $\alpha = 0.2$ and *b* value = 0.75 for the TST approach. $\alpha = 0.05$ for the traditional hypothesis testing approach 60

FIGURES

Figure 1-1. Example test performance curves for traditional WET hypothesis tests.....3
Figure 1-2. Example test performance curves for TST WET hypothesis tests. For this example, *b* is set to 0.8 (denoted by dotted line), with $\alpha = 0.05$.....6
Figure 2-1. Summary of test variability (expressed as the control 90th percentile coefficient of variation or CV) observed between 1989 and 2000 for the chronic *Ceriodaphnia dubia* EPA WET test 12
Figure 3-1. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....20
Figure 3-2. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate.....21
Figure 3-3. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and high control variability as a function of α error rate.....22
Figure 3-4. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and above average control variability and $\alpha = 0.20$, as a function of the number of test replicates 23
Figure 3-5. Percent of *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability ($\alpha = 0.20$) as a function of the number of test replicates.....24
Figure 3-6. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....26
Figure 3-7. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate 27
Figure 3-8. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate.....27
Figure 3-9. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability and an $\alpha = 0.25$, as a function of the number of test replicates 28
Figure 3-10. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....30
Figure 3-11. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate.....31
Figure 3-12. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate.....31

Figure 3-13. Percent of chronic mysid tests having a mean effluent effect of 10 percent and above average control variability declared toxic using TST and an $\alpha = 0.15$, as a function of the number of test replicates 32
Figure 3-14. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....34
Figure 3-15. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate 35
Figure 3-16. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate 35
Figure 3-17. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....37
Figure 3-18. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate 38
Figure 3-19. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate 38
Figure 3-20. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....40
Figure 3-21. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate.....41
Figure 3-22. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of the α error rate 41
Figure 3-23. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....43
Figure 3-24. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate 44
Figure 3-25. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate 44
Figure 3-26. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability.....46

Figure 3-27. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate....47
Figure 3-28. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate....47
Figure 3-29. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability....50
Figure 3-30. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate 51
Figure 3-31. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate....51
Figure 3-32. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability....53
Figure 3-33. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate 54
Figure 3-34. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate....55
Figure 4-1. Range of CV values observed in chronic *C. dubia* ambient toxicity tests for samples that were found to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*). California's SWAMP WET test data. 61
Figure 4-2. Range of CV values observed in chronic *P. promelas* ambient toxicity tests for samples that were declared to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*) 61

EXECUTIVE SUMMARY

The U.S. Environmental Protection Agency (EPA or the Agency) has developed a new statistical approach that assesses the whole effluent toxicity (WET) measurement of wastewater effects on specific test organisms' ability to survive, grow, and reproduce. This new approach is called the Test of Significant Toxicity (TST) and is a statistical method that uses hypothesis testing techniques based on research and peer-reviewed publications. The hypothesis test under the TST approach examines whether an effluent, at the critical concentration (e.g., in-stream waste concentration or IWC), as recommended in EPA's Technical Support Document (TSD; USEPA 1991) and implemented under EPA's WET National Pollutant Discharge Elimination System (NPDES) permits program, and the control within a WET test differ by an unacceptable amount (the amount that would have a measured detrimental effect on the ability of aquatic organisms to thrive and survive).

Since the inception of EPA's NPDES WET program in the mid 1980s, the Agency has striven to advance and improve its application and implementation under the NPDES WET Program. The TST approach explicitly incorporates test power, which, using the TST approach, is the ability to correctly classify the effluent as acceptable under the NPDES WET Program (i.e., non-toxic). The TST approach also provides a positive incentive to generate high quality, valid WET data to make informed decisions regarding NPDES WET reasonable potential (RP) and permit compliance determinations. Once the WET test has been conducted (using multiple effluent concentrations and other requirements as specified in the WET test methods), the TST approach can be used to analyze valid WET test results to assess whether the effluent discharge is toxic. The TST approach is designed to be used for a two concentration data analysis of the IWC or a receiving water concentration (RWC) as compared to a control concentration.

Background

In the NPDES WET Program, an effluent sample is declared toxic relative to a permitted WET limit if the no observed effect concentration (NOEC) is less than the permitted IWC using a hypothesis statistical approach. In such an approach, the question being answered is, "Is the mean response of the organisms the same or worse in the control than at the IWC?" The hypothesis testing approach has four possible outcomes: (1) the IWC is truly toxic and is declared toxic, (2) the IWC is truly non-toxic and is declared non-toxic, (3) the IWC is truly toxic but is declared non-toxic, and (4) the IWC is truly non-toxic but is declared toxic. The latter two possible outcomes represent decision errors that can occur with any hypothesis testing approach. In the NPDES WET Program, those two types of errors occur when either test control replication is poor (i.e., the within-test variability is high) so that even large differences in organism response between the IWC and control are incorrectly classified as non-toxic (outcome [3] above) or, test control replication is very good (i.e., the within-test variability is low) so that a very small difference between IWC and control is declared toxic (outcome [4] above). That former outcome stems from the fact that in the NPDES WET Program, the hypothesis approach established and controls the false positive error rate (i.e., Type I or alpha) but not the false negative error rate (i.e., Type II or beta). Establishing the beta error rate determines the power of the test (power = 1-beta), which is the probability of correctly detecting an actual toxic effect using the traditional hypothesis testing approach (i.e., declaring an effluent toxic when, in fact, it is toxic). By not establishing an appropriate beta error rate and test power in the NPDES WET

Program, the permittee has no incentive to generate more precise data within a test using the traditional hypothesis approach, and, in fact, is at a disadvantage for achieving a high level of precision.

What is the Test of Significant Toxicity Approach?

Organism responses to the effluent and control are unlikely to be exactly the same, even if no toxicity is present. They might differ by such a small amount that even if statistically significant, it would be considered negligible biologically. A more useful approach could be to rephrase the null hypothesis, "Is the mean response in the effluent less than a defined biological amount?" the Food and Drug Administration has successfully used that approach for many years to evaluate drugs, as have many researchers in other biological fields. In that approach, the null hypothesis is stated as the organism response in the effluent is less than or equal to a fixed fraction (*b*) of the control response (e.g., 0.75 of the control mean response):

$$\text{Null hypothesis: Treatment mean} \leq b \times \text{Control mean}$$

In the NPDES WET Program, to reject the null hypothesis above means the effluent is considered non-toxic. To accept the null hypothesis means the effluent is toxic. That test has been adapted for the NPDES WET Program and is referred to as the *Test of Significant Toxicity* (TST).

Before the TST null hypothesis expression could be used in the NPDES WET Program, certain decisions were needed, including what effect level in the effluent is considered unacceptably toxic and the desired frequency of declaring a truly negligible effect within a test non-toxic. Such decisions are referred to as Regulatory Management Decisions (RMDs).

What are the RMDs for TST?

In the TST approach, the *b* value in the null hypothesis represents the threshold for unacceptable toxicity. For *chronic* testing in EPA's NPDES WET Program, the *b* value in the TST analysis is set at 0.75, which means that a 25 percent effect (or more) at the IWC is considered evidence of unacceptable *chronic* toxicity. IWC responses substantially less than a 25 percent effect would be interpreted to have a lower risk potential. The RMD for *acute* WET methods is set at 0.80, which means that a 20 percent effect (or more) at the IWC is considered evidence of unacceptable *acute* toxicity. The acute RMD toxicity threshold is higher (i.e., more strict) than that for chronic WET methods because of the severe environmental implications of acute toxicity (lethality or organism death).

EPA's RMDs using the TST approach are intended to identify unacceptable toxicity in WET tests most of the time when it occurs, while also minimizing the probability that the IWC is declared toxic when in fact it is truly acceptable. This objective requires additional RMDs regarding acceptable maximum false positive (**β using a TST approach)** and false negative rates (**α using a TST approach)**. In the TST approach, the RMDs are defined as (1) declare a sample toxic between 75–95 percent of the time ($0.05 \leq \alpha \leq 0.25$) when there is unacceptable toxicity (20 percent effect for acute and 25 percent effect for chronic tests), and (2) declare an effluent non-toxic no more than 5 percent of the time ($\beta \leq 0.05$) when the effluent effect at the critical effluent concentration is 10 percent. Table ES-1 summarizes the difference in Type I and II error

expressions between the TST approach and the traditional hypothesis approach currently used in the NPDES WET Program.

Table ES-1. Definition of the Type I and Type II error under the traditional hypothesis testing approach and the TST approach.

	Traditional hypothesis approach	TST
Type I (alpha)	Set at 0.05	Set at 0.05 to 0.25 given a *b* value of 0.80 or 0.75 depending on whether the WET test method is acute or chronic, respectively
	Effluent is considered safe but declared *toxic*	Effluent is considered toxic, but declared *safe*
	Permittee concern	Regulatory concern
Type II (beta)	Not established	Set at 0.05
	Effluent is considered toxic but declared *safe*	Effluent is considered safe but declared *toxic*
	Regulatory concern	Permittee concern

How was the TST approach developed?

EPA used valid WET data from approximately 2,000 WET tests to develop and evaluate the TST approach. The TST approach was tested using nine different WET test methods comprising twelve biological endpoints (e.g., reproduction, growth, survival) and representing most of the different types of WET test designs in use. More than one million computer simulations were used to select appropriate alpha error rates for each test method that also achieved EPA's other RMDs for the TST approach.

Once the alpha error rates were established, the results of the TST approach were compared to those obtained using the traditional hypothesis testing approach for a range of test results. The alpha values identified in this project build on existing information (such as data sources and analyses examining ability to detect toxic effects) on WET published and peer reviewed by EPA, including *Understanding and Accounting for Method Variability in WET Applications Under the NPDES Program* (USEPA 2000).

This document outlines the recommended TST approach and presents the following:

- How an appropriate alpha value was identified for several common WET test methods on the basis of desired beta error rates, various effect levels, and within-test control variability.

- The degree of protectiveness of TST compared to the traditional hypothesis testing approach. In this report, *as protective as* is defined as an equal ability to declare a sample toxic at or above the regulatory management level.

Because TST is a form of hypothesis testing, analyses in this document focus on comparing results of TST to the traditional hypothesis testing approach and not to point estimate techniques such as linear interpolation (i.e., IC25). Therefore, this document does not discuss point estimate procedures.

Data analysis approach

EPA assembled a comprehensive database to analyze the utility of the TST approach with data obtained from EPA Regions, several states, and private laboratories, which represent a widespread sampling of typical laboratories and test methods for approximately 2,000 tests. Nine commonly tested WET methods were examined. For each test method, control precision (coefficient of variation [CV]) was calculated on the basis of valid WET test data compiled in the project. Cumulative frequency plots were used to identify percentiles of observed method-specific CVs (e.g., 25^{th}, 50^{th}, 75^{th} percentiles). The measures were calculated to update previous EPA analyses (USEPA 2000) using more recent valid WET test data and to characterize typical, achievable test performance in terms of within-test control variability. A similar analysis was performed for the control response for each of the nine test methods (e.g., mean offspring per female in the *Ceriodaphnia dubia* test method) to characterize typical achievable test performance in terms of control response.

Monte Carlo simulation analysis was used to estimate the percentage of WET tests that would be declared toxic using TST as a function of different α levels, within-test control variability, and mean percent effect level. The simulation analysis identified expected beta error rates (i.e., declaring an effluent toxic when in fact it is acceptable under TST) for a broad range of possible test scenarios. Using the RMDs above, an appropriate α level was then identified for a given WET test design that also yielded a β error rate ≤ 0.05 when there was a 10 percent mean effect. By simulating thousands of WET tests for a given scenario (mean percent effect and control CV), the percentage of tests declared toxic could be calculated and compared among scenarios, and between TST and the traditional hypothesis approach.

Results of the analysis

Results of all analyses indicate that TST is a suitable alternative to the traditional hypothesis approach for analyzing two-concentration WET data (i.e., IWC and control) in the NPDES WET Program. A demonstrated benefit of the TST approach is that increasing the precision and power of the test increases the chances of declaring an effluent non-toxic when there is ≤ 10 percent mean effect in the effluent. Increasing test replication (and thereby the power of the test) results in a *lower* rate of tests declared toxic using TST but a *higher* rate of tests declared toxic using the traditional hypothesis approach (see Figure ES-1). Using TST, a permittee has the ability to demonstrate that its effluent is acceptable, by improving the quality of test data (e.g., decreasing within-test variability, and/or increasing replication), if indeed the mean effect at the IWC is less than the regulatory management decision (25 percent [chronic] or 20 percent [acute]).

On the basis of EPA's analyses, the alpha levels shown in Table ES-2 are recommended for the nine EPA WET test methods examined using the TST approach. An important feature of the TST approach is that the TST's alpha is analogous to beta under the traditional hypothesis testing approach, which had not been established by EPA previously for the NPDES WET Program.

Table ES-2. Summary of alpha (α) levels or false negative rates recommended for different EPA WET test methods using the TST approach.

EPA WET test method	*b* value	Probability of declaring a toxic effluent non-toxic False negative (α) error[a]
Chronic Freshwater and East Coast Methods		
Ceriodaphnia dubia (water flea) survival and reproduction	0.75	0.20
Pimephales promelas (fathead minnow) survival and growth	0.75	0.25
Selenastrum capricornutum (green algae) growth	0.75	0.25
Americamysis bahia (mysid shrimp) survival and growth	0.75	0.15
Arbacia punctulata (Echinoderm) fertilization	0.75	0.05
Cyprinodon variegatus (Sheepshead minnow) and *Menidia beryllina* (inland silverside) survival and growth	0.75	0.25
Chronic West Coast Marine Methods		
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	0.75	0.05
Atherinops affinis (topsmelt) survival and growth	0.75	0.25
Haliotis rufescens (red abalone), *Crassostrea gigas (oyster), Dendraster excentricus, Strongylocentrotus purpuratus* (Echinoderm) and *Mytilus sp* (mussel) larval development methods	0.75	0.05
Macrocystis pyrifera (giant kelp) germination and germ-tube length	0.75	0.05
Acute Methods		
Pimephales promelas (fathead minnow), *Cyprinodon variegatus* (Sheepshead minnow), *Atherinops affinis* (topsmelt), *Menidia beryllina* (inland silverside) acute survival[b]	0.80	0.10
Ceriodaphnia dubia, *Daphnia magna, Daphnia pulex*, *Americamysis bahia* acute survival[b]	0.80	0.10

Notes:
a. α levels shown are the probability of declaring an effluent toxic when the mean effluent effect = 25% for chronic tests or 20% for acute tests and the false positive rate (β) is $\leq$ 0.05 (5%) when mean effluent effect = 10%.
b. Based on a four replicate test design

Results obtained from the TST analyses using the nine EPA WET test methods should be applicable to other EPA WET methods not examined. For example, results generated under this project for the fish *Pimephales promelas* survival and growth test is extrapolated to other EPA fish survival and growth tests (e.g., *Menidia* sp., *Cyprinus variegatus*, *Atherinops affinis*) because the test methods use a similar test design (e.g., number of replicates, number of organisms tested) and measure the same endpoints.

Figure ES-1 illustrates that conducting tests with more replicates (a priori) can assist a permittee to demonstrate that the effluent is acceptable. Conversely, increasing the number of replicates in a test does not assist a permittee using the current hypothesis testing approach.

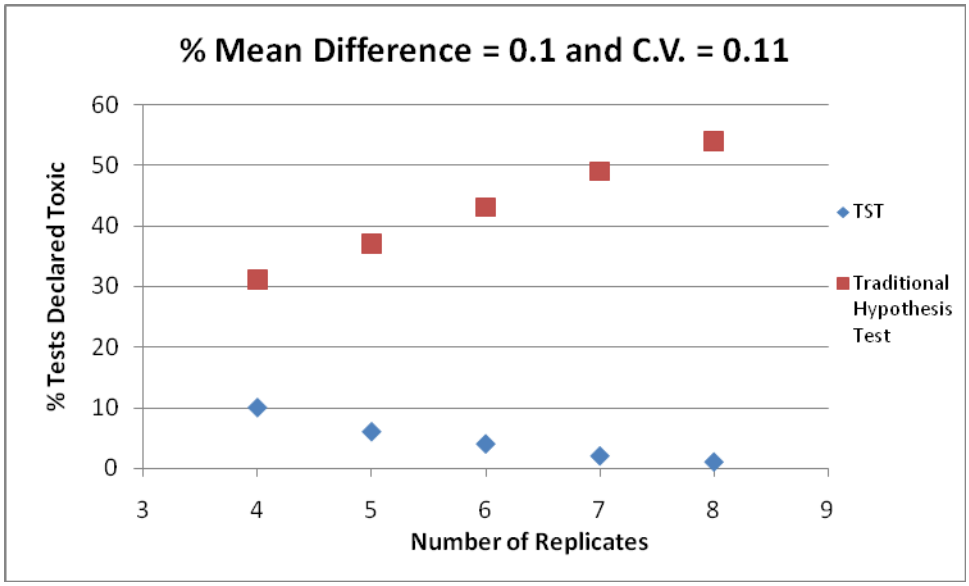

Figure ES-1. Percent of chronic fathead minnow WET tests declared toxic using TST having a mean effluent effect of 10 percent, above average control variability (CV = 0.11 or 11 percent) and an α = 0.25, as a function of the number of within-test replicates. Results using the traditional hypothesis test are shown as well.

Summary

Results of nearly 2,000 valid WET tests and thousands of simulations were conducted to develop the technical basis for the TST approach. That approach builds on the strengths of the traditional hypothesis testing approach, including use of robust statistical analyses, to determine whether an effluent sample is acceptable in WET testing. Specific benefits of using TST in WET analysis include the following:

- Provides transparent RMDs, which are incorporated into the data analysis process
- Incorporates statistical power directly into the statistical process by controlling for both alpha and beta errors, thereby, increasing the confidence in the WET test result
- Provides a positive incentive for the permittee to generate valid, high quality WET data
- Applicable to both NPDES WET permitting and 303(d) watershed assessment programs

Results of this project indicate that the TST is a viable additional statistical approach for analyzing valid acute and chronic WET test data. Using the explicit RMD and test method-specific alpha values, TST provides similar protection as the traditional hypothesis testing approach when there is unacceptable toxicity while also providing a transparent methodology for demonstrating whether an effluent is acceptable under the NPDES WET Program.

In summary, the TST approach provides another option for permitting authorities and permittees to use for analyzing WET test data. The TST approach provides a positive incentive to generate valid, high quality WET data to make informed decisions regarding NPDES WET reasonable

potential (RP) and permit compliance determinations. Using TST, permitting authorities will be better able to identify toxic or acceptable samples.

ACRONYMS AND ABBREVIATIONS

CETIS®	Comprehensive Environmental Toxicity Information System
CFR	Code of Federal Regulations
CV	coefficient of variation
WDNR	Wisconsin Department of Natural Resources
EPA	U.S. Environmental Protection Agency
IC25	25 percent inhibition concentration
IWC	in-stream waste concentration
LOEC	lowest observed effect concentration
LC50	50 percent lethal concentration
MSD	minimum significant difference
NOEC	no observed effect concentration
NPDES	National Pollutant Discharge Elimination System
QA/QC	quality assurance/quality control
RMD	regulatory management decision
RP	reasonable potential
RWC	receiving water concentration
SWAMP	Surface Water Ambient Monitoring Program (California)
TAC	Test acceptability criteria
TMDL	total maximum daily load
TSD	Technical Support Document for Water Quality-Based Toxics Control
TST	Test of Significant Toxicity
WET	whole effluent toxicity

GLOSSARY

Acute Toxicity Test is a test to determine the concentration of effluent or ambient waters that causes an adverse effect (usually mortality) on a group of test organisms during a short-term exposure (e.g., 24, 48, or 96 hours). Acute toxicity is determined using statistical procedures (e.g., point estimate techniques or a t-test).

Ambient Toxicity is measured by a toxicity test on a sample collected from a receiving waterbody.

Chronic Toxicity Test is a short-term test in which sublethal effects (e.g., reduced growth or reproduction) are usually measured in addition to lethality.

Coefficient of Variation (CV) is a standard statistical measure of the relative variation of a distribution or set of data, defined as the standard deviation divided by the mean. The CV can be used as a measure of precision within and between laboratories, or among replicates for each treatment concentration.

Effect Concentration (EC) is a point estimate of the toxicant concentration that would cause an observable adverse effect (e.g., mortality, fertilization). EC_{25} is a point estimate of the toxicant concentration that would cause observable 25% adverse effect as compared to the control test organisms.

False Negative is when the in-stream waste concentration is declared non-toxic but in fact is truly toxic. In the traditional hypothesis approach, false negative error rate is denoted by Beta (β). In the TST approach, false negative error rate is denoted as Alpha (α), which applies when the percent effect in the critical effluent concentration is $\geq$ 25% for a given test.

False Positive is when the in-stream waste concentration is declared toxic but in fact is truly non-toxic. In the traditional hypothesis approach, false positive error rate is denoted by Alpha (α). In the TST approach, false positive error rate is denoted as Beta (β), which applies when the percent effect in the critical effluent concentration is $\leq$ 10% for a given test.

Hypothesis Testing is a statistical approach (e.g., Dunnett's procedure) for determining whether a test concentration is statistically different from the control. Endpoints determined from hypothesis testing are no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC). The two hypotheses commonly tested in WET are

- **Null hypothesis (H_o)**: The effluent is non-toxic.
- **Alternative hypothesis (H_a)**: The effluent is toxic.

Inhibition Concentration (IC) is a point estimate of the toxicant concentration that would cause a given, percent reduction in a non-lethal biological measurement (e.g., reproduction or growth), calculated from a continuous model (i.e., Interpolation Method). E.g., IC_{25} is a point estimate of the toxicant concentration that would cause a 25 percent reduction in a non-lethal biological measurement.

In-stream Waste Concentration (IWC) is the concentration of a toxicant or effluent in the receiving water after mixing. The IWC is the inverse of the dilution factor. It is sometimes referred to as the receiving water concentration (RWC).

LC50 (lethal concentration, 50 percent) is the toxicant or effluent concentration that would cause death to 50 percent of the test organisms.

Lowest Observed Effect Concentration (LOEC) is the lowest concentration of an effluent or toxicant that results in statistically significant adverse effects on the test organisms (i.e., where the values for the observed endpoints are statistically different from the control).

Minimum Significant Difference (MSD) is the magnitude of difference from control where the null hypothesis is rejected in a statistical test comparing a treatment with a control. MSD is based on the number of replicates, control performance, and power of the test.

No Observed Effect Concentration (NOEC) is the highest tested concentration of an effluent or toxicant that causes no observable adverse effect on the test organisms (i.e., the highest concentration of toxicant at which the values for the observed responses are not statistically different from the control).

National Pollutant Discharge Elimination System (NPDES) is the national program for issuing, modifying, revoking and reissuing, terminating, monitoring and enforcing permits, and imposing and enforcing pretreatment requirements, under sections 307, 318, 402, and 405 of Clean Water Act.

Power is the probability of correctly rejecting the null hypothesis (i.e., declaring an effluent toxic when, in fact, it is toxic using the traditional hypothesis test approach).

Precision is a measure of reproducibility within a data set. Precision can be measured both within a laboratory (within-laboratory) and between laboratories (between-laboratory) using the same test method and toxicant.

Quality Assurance (QA) is a practice in toxicity testing that addresses all activities affecting the quality of the final effluent toxicity data. QA includes practices such as effluent sampling and handling, source and condition of test organisms, equipment condition, test conditions, instrument calibration, replication, use of reference toxicants, recordkeeping, and data evaluation.

Quality Control (QC) is the set of more focused, routine, day-to-day activities carried out as part of the overall QA program.

Reasonable Potential (RP) is where an effluent is projected or calculated to cause an excursion above a water quality standard on the basis of a number of factors including the four factors listed in Title 40 of the *Code of Federal Regulations* (CFR) 122.44(d)(1)(ii).

Reference Toxicant Test is a check of the sensitivity of the test organisms and the suitability of the test methodology. Reference toxicant data are part of a routine QA/QC program to evaluate the performance of laboratory personnel and the robustness and sensitivity of the test organisms.

Regulatory Management Decision (RMD) is the decision that represents the maximum allowable error rates and thresholds for toxicity and non-toxicity that would result in an acceptable risk to aquatic life.

Replicate is two or more independent organism exposures of the same treatment (i.e., effluent concentration) within a whole effluent toxicity test. Replicates are typically separate test chambers with organisms, each having the same effluent concentration**.**

Sample is a representative portion of a specific environmental matrix that is used in toxicity testing. For this document, environmental matrices could include effluents, surface waters, groundwater, stormwater, and sediment.

Significant Difference is a statistically significant difference (e.g., 95 percent confidence level) in the means of two distributions of sampling results.

Statistic is a computed or estimated quantity such as the mean, standard deviation, or Coefficient of Variation.

Test Acceptability Criteria (TAC) are test method-specific criteria for determining whether toxicity test results are acceptable. The effluent and reference toxicant must meet specific criteria as defined in the test method (e.g., for the *Ceriodaphnia dubia* survival and reproduction test, the criteria are as follows: the test must achieve at least 80 percent survival and an average of 15 young per surviving female in the control and at least 60% of surviving organisms must have three broods).

t-test (formally Student's t-Test) is a statistical analysis comparing two sets of replicate observations, in the case of WET, only two test concentrations (e.g., a control and IWC). The purpose of this test is to determine if the means of the two sets of observations are different (e.g., if the 100-percent effluent or ambient concentration differs from the control [i.e., the test passes or fails]).

Type I Error (alpha α) is the error of rejecting the null hypothesis (H_o) that should have been accepted.

Type II Error (beta β) is the error of accepting the null hypothesis (H_o) that should have been rejected.

Toxicity Test is a procedure to determine the toxicity of a chemical or an effluent using living organisms. A toxicity test measures the degree of effect on exposed test organisms of a specific chemical or effluent.

Welch's t-test is an adaptation of Student's t-test intended for use with two samples having unequal variances.

Whole Effluent Toxicity (WET) is the total toxic effect of an effluent measured directly with a toxicity test.

1.0 INTRODUCTION

1.1 Summary of Current EPA Recommended WET Analysis Approaches

Within the National Pollutant Discharge Elimination System (NPDES) Program, freshwater and marine acute and chronic whole effluent toxicity (WET) tests are used in conjunction with other analyses to evaluate and assess compliance of wastewater and surface waters with water quality standards of the Clean Water Act. In the NPDES WET Program, WET tests examine organism responses to effluent, typically along a dilution series (USEPA 1995, 2002a, 2002b). Acute WET test methods measure the lethal response of test organisms exposed to effluent (USEPA 2002c). The principal response endpoints for such methods are the effluent concentration that is lethal to 50 percent of the test organisms (LC50) or the effluent concentration at which survival is significantly lower than the control (e.g., t-test). Chronic WET test methods often measure both lethal and sublethal responses of test organisms. The statistical endpoints that are used in chronic WET testing in the NPDES WET Program are the no observed effect concentration (NOEC), and the 25 percent inhibition concentration (IC25). The NOEC endpoint is determined using a traditional hypothesis testing approach that identifies the maximum effluent concentration tested at which the response of test organisms is not significantly worse from the control. From a regulatory perspective, an effluent sample is declared toxic relative to a permitted WET limit if the NOEC is less than the permitted in-stream waste concentration (IWC), as recommended in EPA's Technical Support Document (TSD) (USEPA 1991) and implemented under EPA's WET NPDES permits program. The IC25, by contrast, is a point-estimation approach. It identifies the concentration at which the response of test organisms is 25 percent below that observed in the control concentration and interpolates the effluent concentration at which this magnitude of response is expected to occur. From a regulatory perspective, an effluent sample is declared toxic relative to a permitted WET limit if the IC25 is less than the permitted IWC. This document focuses on another statistical option with respect to the traditional hypothesis testing approach for analyzing and interpreting valid WET data.

1.2 Advantages and Disadvantages of Recommended Traditional Hypothesis Testing Approach

The hypotheses traditionally used in WET statistical comparisons of a biological measure (survival, growth, reproduction) in control water versus a particular effluent sample are the following:

$$\textit{Null Hypothesis}: \quad \mu_T \geq \mu_C$$
$$\textit{Alternative Hypothesis}: \quad \mu_T < \mu_C$$

where μ_C refers to the true mean for the biological measure in the control water and μ_T refers to the true mean for this measure in the effluent sample. *True mean* here refers to the mean for a theoretical statistical population of results from indefinite repetition of toxicity tests on the same control water and effluent sample. In contrast, the mean for the biological measure for a single toxicity test would be referred to as the *sample mean*, and random variation among organisms might cause a sample mean for an effluent to be less than the control even if the effluent is actually non-toxic. The traditional WET hypothesis thus assumes that the effluent sample is non-toxic. For an individual test, there must be a statistical test to determine if the null hypothesis is

rejected in favor of the alternative hypothesis; i.e., that any apparent toxicity based on the sample means is real and not simply reflective of random variation. Such a statistical test is part of current recommended practice in WET testing.

Table 1-1 summarizes the correctness of results from such statistical testing, contrasting the true condition of whether the effluent sample is toxic to the result of the statistical test. Two types of errors can occur in the statistical test result. A false positive occurs when the effluent is actually non-toxic, but the statistical test infers that it is toxic. For the statistical hypotheses here, that is a Type I error (the null hypothesis is rejected when it is true) and the probability of this error is typically designated by the variable α, so that the correct decision occurs with probability $1 - \alpha$. The other type of error, a false negative, occurs when the effluent truly is toxic, but the statistical test infers that it is non-toxic. For the statistical hypotheses here, that is a Type II error (the null hypothesis is accepted when it is false) and the probability of the error is typically designated by the variable β, so that the probability of the correct decision is $1 - \beta$, which is also referred to as the test power.

Table 1-1. Error terminology for traditional WET hypothesis methodology

Statistical test result	**True condition**	
	$\mu_T \geq \mu_C$ (sample is non-toxic)	**$\mu_T < \mu_C$ (sample is toxic)**
$\mu_T \geq \mu_C$ (Sample is non-toxic)	Correct Decision (probability=1–α)	False Negative Type II Error (probability=β)
$\mu_T < \mu_C$ (Sample is toxic)	False Positive Type I Error (probability=α)	Correct decision Test Power (1-β)

It is important to note that β does not have a single value but rather is a function of how toxic the sample actually is (i.e., there is a greater chance of incorrectly saying an effluent is non-toxic if it is only slightly toxic than if it is highly toxic). Similarly, given that the null hypothesis is an inequality, α also does not have a single value, because if effluent characteristics actually improve the biological measure, the probability with which a non-toxic effluent is called toxic will be a function of the extent of this beneficial effect. Although there is a designated single value for α in the statistical test calculations (e.g., 0.05), this error probability applies only when the true condition is exactly at $\mu_T = \mu_C$.

This variation of α and β can be better understood using Figure 1-1, which depicts the probability of declaring an effluent toxic versus the true toxicity of the effluent, expressed as the ratio of the true biological measure in the effluent to the true biological measure in the control (μ_T / μ_C). The curves on this figure are for a hypothetical statistical analysis of hypothetical toxicity tests, but exemplify performance curves that could be drawn for *any* statistical analysis of *any* toxicity test under the traditional WET hypotheses provided above. The solid line is for a toxicity test with large variability so that it is less likely that the statistical test will detect toxicity, and the dashed line is for a toxicity test with low variability. Such curves provide a useful and complete summary of the basic information desired from WET testing. How

effectively will the testing detect toxicity for different levels of true toxicity? How often will non-toxic effluents mistakenly be declared toxic? Although test performance can be appreciated from such curves without addressing specific types of statistical errors, the behavior of those errors can be illustrated using the curves. The portion of the curve with $\mu_T / \mu_C \geq 1$ gives values for α (i.e., the effluent is truly non-toxic so that calling it toxic, a false positive, is a Type I error under the traditional null hypothesis). In accordance with WET hypothesis test procedures, the example curves have $\alpha = 0.05$ when μ_T / μ_C is exactly at 1.0. The portion of the curve with $\mu_T / \mu_C < 1$ is the *power curve* for the test (i.e., 1–β, the probability of calling an effluent toxic when it truly is toxic). This illustrates how test power is very low (approaching 0.05) when the effluent is only slightly toxic, but it increases as the true toxicity increases. The two different curves illustrate how this increase in test power depends on test uncertainty—i.e., higher within-test variability in the toxicity test results in less power for the statistical analysis.

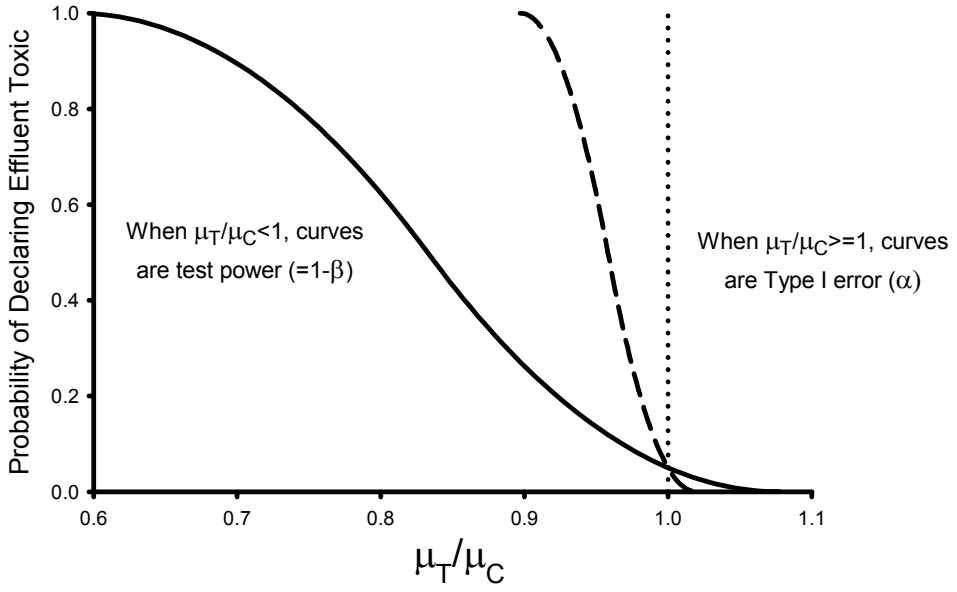

Figure 1-1. Example test performance curves for traditional WET hypothesis tests. The dotted line marks where the true mean biological measure in the effluent equals that in the control. The solid curve is for a high variability test, while the dashed curve is for a low variability test.

Various researchers have reported several advantages and disadvantages of the traditional hypothesis testing approach as currently used in the NPDES WET Program (Grothe et al. 1996). Two common limitations cited are (1) if the test control replication is very good (i.e., test is very precise), an effluent might be considered toxic when in fact its toxicity is low enough to be considered acceptable, and (2) if test control replication is poor (i.e., the test is very imprecise), a highly toxic effluent might be incorrectly classified as non-toxic. For example, the more precise test in Figure 1-1 would declare an effluent with only 5 percent toxicity to be toxic about 60 percent of the time, whereas the less precise test in Figure 1-1 would declare 20 percent toxicity to be non-toxic about 40 percent of the time. The first limitation arises because the null hypothesis is defined around $\mu_T = \mu_C$, so the goal is to call an effluent toxic if $\mu_T < \mu_C$, no matter

how small the difference. The second limitation arises from the fact that the NPDES WET Program hypothesis testing approach does not address the false negative error rate (i.e., Type II error, β) and thus does not address requirements regarding the power of the test to detect substantial levels of toxicity. By not establishing an appropriate β and test power in the NPDES WET Program, the permittee has no incentive to increase the precision of a WET test when using the traditional hypothesis approach. As illustrated in Figure 1-1, greater precision simply results in more samples being declared toxic and can lead to high rejection rates for effluents with low levels of toxicity that might be considered acceptable. Although EPA has made improvements in statistical procedures, such as including a test review step of the percent minimum significant differences (i.e., to minimize within-test variability), it is desirable to further improve the hypothesis testing approach. Such improvement is the focus of this report and a general approach for this, the Test of Significant Toxicity (TST), is discussed next.

1.3 Test of Significant Toxicity

The TST is an alternative statistical approach for analyzing and interpreting valid WET test data that also uses a hypothesis testing approach but in a different way, building on previous work conducted by EPA in the NPDES WET Program (USEPA 2000) as well as other researchers (Erickson and McDonald 1995; Shukla et al. 2000; Berger and Hsu 1996). The TST approach is based on a type of hypothesis testing referred to as *bioequivalence testing*. Bioequivalence is a statistical approach that has long been used in evaluating clinical trials of pharmaceutical products (Anderson and Hauck 1983) and by the Food and Drug Administration (Hatch 1996; Aras 2001; Streiner 2003). The approach has also been used to evaluate the attainment of soil cleanup standards for contaminated sites (USEPA 1988, 1989) and to evaluate effects of pesticides in experimental ponds (Stunkard 1990).

For the NPDES WET Program, the TST approach changes the hypotheses to the following:

$$\textit{Null Hypothesis:} \quad \mu_T \leq b \times \mu_C$$

$$\textit{Alternative Hypothesis:} \quad \mu_T > b \times \mu_C$$

The TST hypotheses thus incorporate two important differences from the traditional WET hypotheses. First, a specific value for the ratio μ_T / μ_C, designated *b*, is included to delineate unacceptable and acceptable levels of toxicity, allowing a risk management decision about what level of toxicity should be allowed if the true means were known, other than the absence of any toxicity as specified by the traditional hypothesis. Second, the inequalities are reversed so that it is assumed that the effluent sample has an unacceptable level of toxicity until demonstrated otherwise. As a result of this reversal of the inequalities, the meanings of α and β under the TST hypotheses (Table 1-2) are reversed from those under the traditional hypothesis approach (Table 1-1). Under the TST approach, α is associated with false negatives, β is associated with false positives, and statistical test power using the TST approach in the NPDES WET Program is the ability to correctly conclude that true toxicity levels are acceptable. In addition, an effluent sample would be considered acceptable under the TST approach when the null hypothesis is rejected; in contrast, a sample is considered unacceptable under the traditional hypothesis approach when the null hypothesis is rejected.

Table 1-2. Error terminology for TST WET hypothesis methodology

Statistical test result	**True condition**	
	$\mu T \leq b \times \mu C$ (Toxicity is unacceptable)	**$\mu T > b \times \mu C$ (Toxicity is acceptable)**
$\mu_T \leq b \times \mu_C$ (Toxicity is unacceptable)	Correct Decision (1–α)	False Positive Type II Error (β)
$\mu_T > b \times \mu_C$ (Toxicity is acceptable)	False Negative Type I Error (α)	Correct Decision Test Power (1–β)

Figure 1-2 provides illustrative examples of test performance under the TST approach and illustrates advantages of this approach over the traditional hypotheses. This figure shows the same basic type of performance curve as in Figure 1-1: the probability of calling an effluent unacceptably toxic versus the true toxicity in the effluent. Incorporating *b* in the hypotheses explicitly recognizes that the true mean for the organism response in an effluent can be less than that in the control by a certain amount and still be considered acceptable, and it keeps the false negative rate for this amount of toxicity constant regardless of test variability (Figure 1-2). As mentioned previously, the current NPDES WET Program does not control the false negative rate, which varies markedly at any given level of toxicity as test precision varies (Figure 1-1). By reversing the inequalities and referencing them to *b*, the TST approach also results in more precise tests having lower false positive errors (Figure 1-2); i.e., effluents with true levels of toxicity that are acceptably low are declared toxic with less frequency as precision increases, a desirable attribute for the method. That provides permittees with a clear incentive to improve the precision of test results. Thus, using the TST approach, a permittee has to demonstrate with some confidence that their effluent has toxicity in an acceptable range, but can also improve testing procedures as needed to do so (i.e., increase replicates or decrease within-test variability or both).

1.4 Regulatory Management Decisions for TST

Regulatory management decisions (RMDs) are incorporated into the TST methodology by selecting values for *b*, the dividing point between acceptable and unacceptable toxicity, and α, the false negative error rate when $\mu_T = b \times \mu_C$.

The selection of *b* should reflect what is considered acceptable if the true biological response means for the effluent and control were actually known, especially because precise tests might have performances closely approaching this ideal. For all chronic WET test methods, the RMD is to set *b* to 0.75. This *b* value (25 percent toxic effect) is consistent with EPA's use of the IC25 in point estimation methods for examining chronic WET data. Chronic effects less than 25 percent would be considered to have an acceptably low risk potential. Because of the more severe environmental implications of acute toxicity (organism death), the RMD for acute WET test methods is to set *b* higher than that for chronic WET test methods, at 0.80.

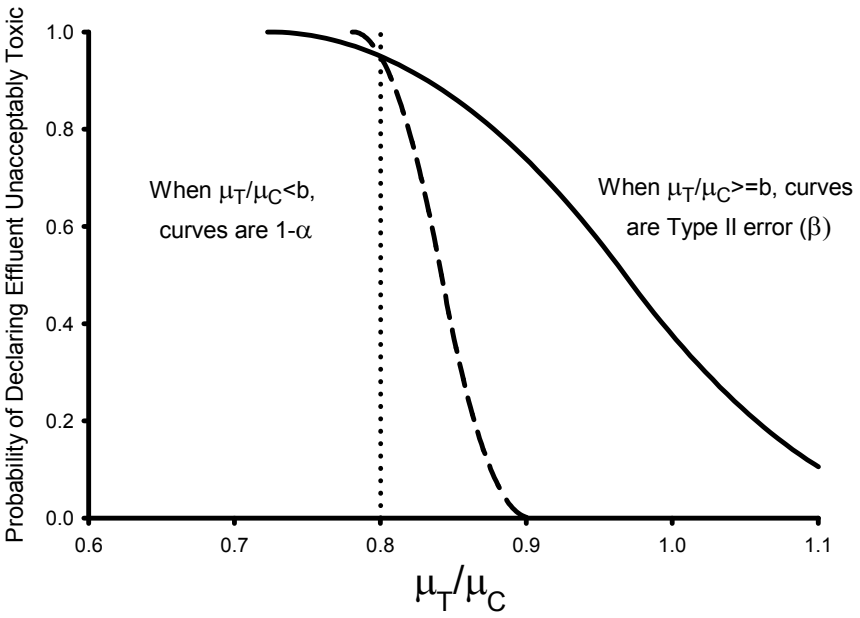

Figure 1-2. Example test performance curves for TST WET hypothesis tests. For this example, *b* is set to 0.8 (denoted by dotted line), with α = 0.05. The two curves represent test performance for tests with high (solid line) and low (dashed line) variability.

For a given test precision and value for *b*, selecting a value for α completely determines both false negative and false positive error rates at all toxicity levels, such as the curves in Figure 1-2. However, the value selected for α does not have to be based just on consideration of the desired error rate when $\mu_T = b \times \mu_C$. Rather, α can be selected on the basis of balancing goals regarding this false negative error rate with goals for false positive error rates at lower levels of toxicity. Therefore, a different α can be assigned for different types of WET toxicity tests based on test precision and on specific goals regarding false positive and false negative rates.

With regard to false negative rates, EPA's general goal is to identify unacceptable toxicity in WET tests most of the time when it occurs. It would be preferred to set α at the typical 0.05 level (i.e., if $\mu_T = b \times \mu_C$, the effluent will be declared unacceptable 95 percent of the time). However, for tests with low precision, this could result in a high rate of false positives (declaring effluents unacceptable) when toxicity is low or absent (e.g., Figure 1-2). Therefore, values of α up to 0.25 will be allowed, as needed to meet the goal regarding false positive rates discussed in the next paragraph. Thus, the false negative rate RMD is $0.05 \leq \alpha \leq 0.25$, so that there is at least a 0.75 probability that an effluent with unacceptable toxicity ($\mu_T \leq b \times \mu_C$) will be declared toxic.

With regard to false positive error probabilities, EPA's general goal is that they be low when toxicity is negligible. It is necessary to define *negligible* as a second, smaller level of effect than *acceptable* because the latter includes toxicity as high as that represented by *b*, at which point the false positive error rate always will approach $1 - \alpha$, so cannot be low. With regard to this, EPA defines negligible as 10 percent toxicity or less, and specifies that the false positive error

probability be no higher than 0.05 at 10 percent toxicity. Thus, the false positive RMD is $\beta \leq 0.05$ at $\mu_T/\mu_C=0.90$, provided this is achievable with $\alpha \leq 0.25$ (if α is at this maximum, this false positive RMD no longer applies). It should be emphasized that this RMD relates to only one point in the range of toxicity considered acceptable, and that false positives will vary widely within this range (e.g. Figure 1-2). False positive rates will be lower when toxicity is lower than 10 percent, dropping to near zero when toxicity is absent, and will be higher when toxicity values are greater than negligible but still acceptable, rising to 1-α as the toxicity approaches the unacceptable level.

Therefore, the overall RMD for α (the false negative rate when $\mu_T / \mu_C = b$) is to set it to the lowest value that results in $\beta \leq 0.05$ (the false positive rate) when the true toxicity is at $\mu_T / \mu_C = 0.90$, but that α will be no lower than 0.05 and no higher than 0.25. This selection will be primarily a function of test method within-test variability (e.g., control coefficient of variation or CV), but cannot and should not be done on an individual test basis. Rather, TST alphas are assigned for different types of WET tests on the basis of simulations that address how TST method performance is affected by the test design and types of endpoints measured, and the associated CVs.

1.5 Document Objectives

This document presents TST as a useful alternative data analysis approach for valid WET test data that may be used in addition to the approaches currently recommended in EPA's Technical Support Document (USEPA 1991) and EPA's WET test method manuals. In adapting the TST for use in evaluating WET test data, analyses were conducted to identify an appropriate Type I error rate (α) for several common EPA WET methods given certain RMDs. Once alpha error rates were established, results of the TST approach were compared to those obtained using the traditional hypothesis testing approach and a range of test results.

This document outlines the recommended TST approach and presents the following:

- How an appropriate alpha value was identified for several common EPA WET test methods on the basis of desired alpha and beta error rates using explicit RMDs (i.e., effect levels) and considering a range of within-test control variability observed in valid WET tests.
- The degree of protectiveness of TST compared to the traditional hypothesis testing approach. In this report, *as protective as* is defined as an equal ability to declare a sample toxic at or above the regulatory management decision.

In this project, emphasis was placed on comparing results of TST to traditional hypothesis testing approaches and not to point estimate techniques such as linear interpolation (i.e., IC25). Therefore, this document does not discuss linear interpolation techniques. In addition, this document discusses the TST approach only with regard to comparing individual effluent samples to a control, and does not evaluate extensions of the TST approach to simultaneous multiple comparisons such as in Erickson and McDonald (1995).

The focus of this document is on chronic WET test methods and sublethal endpoints because many different types of alternative analysis procedures have been proposed for these tests.

Applying the TST methodology to the acute fish and *Ceriodaphnia* WET test method is also included. This document provides a summary of the recommended TST method, α values for several common WET methods, and results of comprehensive analyses supporting EPA recommendations.

2.0 METHODS

Methods used to evaluate the TST approach and determine how it should be applied for WET test analysis in the NPDES WET Program proceeded using several general steps as follows:

Step 1: WET test methods and endpoints were selected for analysis in the TST evaluation. A range of the more common EPA WET test methods were identified in this step.

Step 2: WET data were compiled from several state and EPA sources to determine current WET test method performance in terms of control response and within-test control variability.

Step 3: Simulation analyses were conducted using data characteristics obtained from Step 2 to guide the types of simulated data analyzed in this project and to set test method-specific α levels.

The following sections describe in more detail each of the steps.

2.1 Test Methods and Endpoints Evaluated

Table 2-1 summarizes the nine EPA WET test methods evaluated in this project. Preference was given to valid WET data generated using the EPA 1995 WET test methods for the EPA West Coast marine species (USEPA 1995) and for all other species the 2002 EPA WET test methods (USEPA 2002a, 2002b). Examining the inter-laboratory reference toxicant data for *C. dubia* by year indicated significantly more precise data from 1996 on as compared to pre-1995 (Figure 2-1). Similar results were observed for the fathead minnow and chronic mysid test methods as well. This result is not unexpected because the EPA chronic WET test methods were substantially refined as of 1995 and laboratories had more experience with the chronic test methods by this time. Within-test control 90th percentile CVs were not significantly different among years following 1995. Therefore, only post-1995 data were used in analyses for all EPA WET test methods.

All of the WET test methods listed in Table 2-1 are commonly used by regulatory authorities in making regulatory decisions such as determining WET reasonable potential (RP) or to determine compliance with acute and chronic WET limits or monitoring triggers. These nine test methods are representative of the range of EPA WET test methods commonly required of permittees in terms of types of toxicity endpoints written into NPDES permits and test designs followed by permittee's testing laboratories. Results obtained using these nine EPA test methods should be applicable to other EPA WET test methods not examined. For example, results of this project for the fish *Pimephales promelas* survival and growth test is extrapolated to other EPA fish survival and growth tests (e.g., *Menidia sp*., *Cyprinus variegatus*, *Atherinops affinis*) because those test methods use a similar test design (e.g., number of replicates, number of organisms tested) and measure the same endpoints. Previous analyses conducted by EPA (Denton and Norberg-King 1996; Denton et al. 2003) found comparable effect sizes for a given power among similar experimental designs and test endpoints. Similarly, the acute freshwater fish WET test analyzed in this project can be extrapolated to other fish acute test methods because they use a similar test design and measure mortality or immobility. The use of both EPA saltwater and freshwater WET tests ensured that there was adequate representation of different types of discharge situations and laboratories.

Table 2-1. Summary of test condition requirements and test acceptability criteria for each EPA WET test method evaluated in TST analyses

EPA method	Organism with scientific name	Endpoint type	Test type	Minimum # per test chamber	Minimum # of rep per conc.	Minimum # effluent conc.	Test duration	Test acceptance criteria (TAC)
2000.0	Fathead minnow *(Pimephales promelas)*	Survival	Acute	10	2	5	48–96 hours	≥ 90% survival in controls
1000.0	Fathead minnow *(Pimephales promelas)*	Survival and growth (larval)	Chronic	10	4	5	7 days	≥ 80% survival in controls; average dry weight per surviving organism in control chambers equals or exceeds 0.25 mg
1002.0	Water flea *(Ceriodaphnia dubia)*	Survival and reproduction	Chronic	1	10	5	Until 60% of surviving control organisms have 3 broods (6–8 days)	≥ 80% survival and an average of 15 or more young per surviving female in the control solutions. 60% of surviving control organisms must produce three broods
1007.0	Mysid shrimp *(Americamysis bahia)*	Survival and growth	Chronic	5	8	5	7 days	≥ 80% survival; average dry weight ≥ 0.20 mg in controls
1016.0	Purple urchin *(Strongylocentro-tus purpuratus)* or Sand dollar *(Dendraster excentricus)*	Fertilization	Chronic	100	4	4	40 min (20 min plus 20 min)	≥ 70% egg fertilization in controls; %MSD < 25%; and appropriate sperm counts
1017.0	Giant kelp *(Macrocystis pyrifera*	Germination and germ-tube length	Chronic	100 for germination 10 for germ-tube length	5	4	48 hours	≥ 70% germination in controls; ≥ 10 μm germ-tube lengths in controls; %MSD of < 20% for both germination and germ-tube length NOEC must be below 35 μg/L in reference toxicant test
1014.0	Red abalone (*Haliotis rufescens*)	Larval development	Chronic	100	5	4	48 hours	≥ 80% normal larval development in controls Statistical significance @ 56 μg/L zinc % MSD < 20%

Table 2-1. continued.

EPA method	Organism with scientific name	Endpoint type	Test type	Minimum # per test chamber	Minimum # of rep per conc.	Minimum # effluent conc.	Test duration	Test acceptance criteria (TAC)
2002.0	Water flea (*Ceriodaphnia dubia)*	Survival	Acute	5	4	5	24, 48, or 96 hours	$\geq$ 90% survival in controls
1003.0	Green algae (*Selenastrum capricornutum*)	Growth (cell counts, chlorophyll fluorescence, absorbance, or biomass)	Chronic	10,000cells/ mL	4	5	96 hour	Mean cell density of at least 1 X 10^6 cells/mL in the controls; variability (CV%) among control replicates less than or equal to 20%

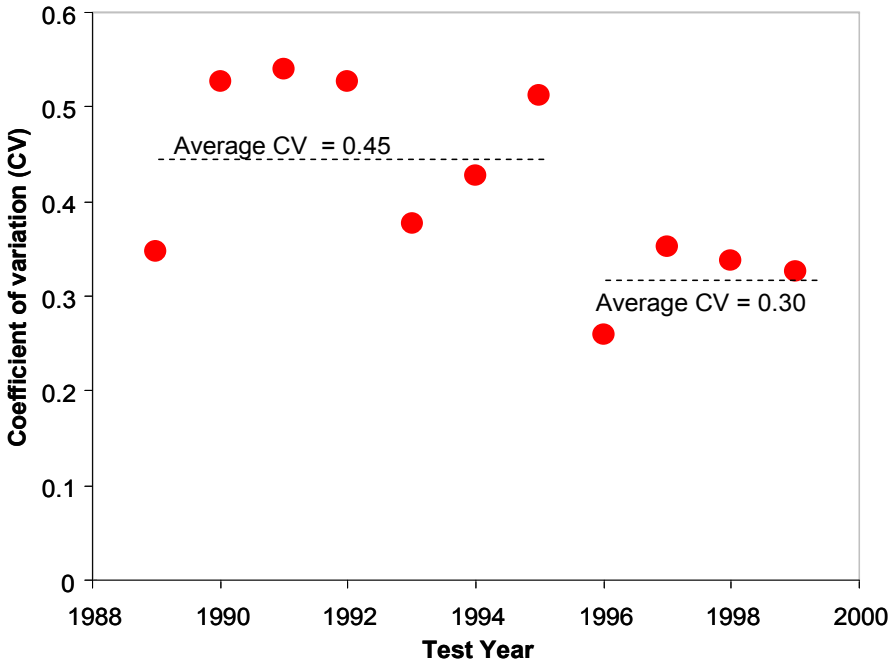

Figure 2-1. Summary of test variability (expressed as the control 90th percentile coefficient of variation or CV) observed between 1989 and 2000 for the chronic *Ceriodaphnia dubia* EPA WET test. This figure illustrates and supports the basis for using test data post 1995, as test precision improved from an average 90th percentile CV of 0.47 to 0.30.

2.2 Data Compilation

Data Sources

WET data were received from several reliable sources to identify baseline test method statistics (e.g., control CV percentiles, mean response percentiles) that were used in simulation analyses (see Section 2.4) and to help identify appropriate α values for each test method. The sources included Washington State Department of Ecology, EPA's Office of Science and Technology, North Carolina Department of the Environment and Natural Resources, California State Water Resources Control Board, and Virginia Department of Environmental Quality. Data acceptance criteria and types of WET test data desired were identified and documented in the Data Management Plan and the Quality Assurance Project Plan for this project. Nearly 2,000 valid WET tests of interest were incorporated, representing many permittees and laboratories (Table 2-2). Only data from WET tests meeting EPA's test acceptability criteria were used in the analyses.

For each set of test data received, additional metadata information was required including the following:

- Permittee name and NPDES permit number (coded for anonymity)
- Laboratory name and location (coded for anonymity)
- Design effluent concentration in the receiving water (expressed as percent effluent upon complete mix) used by the regulatory authority
- EPA test method version used (cited EPA number)
- Information indicating that all EPA test method's test acceptability criteria were met

In addition to the above effluent test data and metadata, two other sources of toxicity data were compiled in this project, which were used to help calculate the range of control organism response by endpoint for each EPA WET test method in Table 2-1. The first source of data was reference toxicant test data previously compiled for the EPA document, *Understanding and Accounting for Method Variability in Whole Effluent Toxicity Application Under the NPDES Program* (USEPA 2000). A second source of additional WET test data used in this project was data generated in ambient toxicity tests by the California State Water Resources Control Board. These data were useful in supplying information on control responses for the freshwater test methods in Table 2-1. Many states routinely conduct ambient toxicity tests as part of 305(b) monitoring; Total Maximum Daily Loads (TMDLs), and other programs (e.g., California's Surface Water Ambient Monitoring program (SWAMP), Washington Department of Ecology's ambient program, Wisconsin Department of Natural Resources' (WDNR) ambient monitoring program).

Table 2-2. Summary of WET test data analyzed

EPA WET test method	Number of tests		Number of laboratories	Number of permittees
	Effluent	Ref Tox		
Ceriodaphnia dubia (water flea) Survival and Reproduction[a]	554	238	44	68
Pimephales promelas (fathead minnow) Acute Survival[b]	347	0	15	101
Pimephales promelas (fathead minnow) Survival and Growth[b]	275	197	28	50
Americamysis bahia (mysid shrimp) Survival and Growth[c]	74	136	20	6
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) Fertilization[c]	83	94	11	10
Macrocystis pyrifera (giant kelp) Germination and Germ-tube length[d]	0	135	11	--
Haliotis rufescens (red abalone) Larval Development[c]	0	136	10	--
Ceriodaphnia dubia (water flea) Survival	7	232	27	2
Selenastrum capricomutum (green algae)	139	84	14	44

Notes:

a. Freshwater invertebrate
b. Freshwater vertebrate
c. Saltwater invertebrate
d. Saltwater algae

Representativeness of WET Data

The usefulness of the results obtained in this project depended on having valid, representative WET test data for each of the EPA WET test methods examined. Representativeness was characterized in this project as having data that met the following:

- Cover a range of NPDES permitted facility types, including both industrial and municipal permittees

- Represent many facilities for a given EPA WET test method (i.e., no one facility dominates the data for a given WET test method)
- Cover a range of target (design) effluent dilutions upon which WET RP and compliance are based, ranging from perhaps 10 percent to 100 percent effluent
- Generated by several laboratories for a given EPA WET test method
- Cover a range of observed effluent toxicity for each EPA WET test method (e.g., NOECs range from < 10 percent to 100 percent effluent)

Efforts were made to ensure that no one laboratory or permittee had > 10 percent of the test data for a given test type. The summary information presented in Table 2-2 demonstrates that WET test data were received from numerous laboratories and facilities for all EPA WET test methods analyzed under this project.

Data Processing

Processing of raw WET test data began with identifying the contents of each data package and recording the data source, test type, and related information as described in the previous section. Each valid WET test was assigned a unique code, and each laboratory was uniquely coded. A tracking system was used to help evaluate whether WET test data were needed for certain types of EPA WET test methods and to help increase representativeness of laboratories or types of facilities for a method.

Data were received in a variety of formats and compiled by test type in the database program CETIS® (Comprehensive Environmental Toxicity Information System; Tidepool Software, v. 1.0). The CETIS program is designed to analyze, store, and manage WET test data. WET test data received in either ToxCalc® or CETIS were imported directly into the CETIS database dedicated to this project. WET test data received in Excel or other spreadsheet formats were also directly imported into CETIS. In cases where the source organizations had not yet entered its WET test data electronically, they were supplied with a template so the data could be readily transferred to CETIS to minimize transcription errors. Data in CETIS were checked on 10 percent of the tests received from each source to document proper data transfer.

WET test data received as copies of bench sheets were first checked to ensure that all EPA WET method test acceptance criteria were met, as well as several other requirements discussed in the previous section. Those tests meeting all requirements were input into the CETIS database directly using the double entry mode and a comparison of entries to ensure accuracy of data input. All WET test data used in analyses originated from tests conducted with the minimum number of treatment replicates as required according to the specific EPA WET test methods (e.g., 10 replicates in chronic *Ceriodaphnia* tests). Tests using a different number of replicates per treatment were not used in analyses to generate percentiles of CV or mean response.

2.3 Setting the Test Method-Specific α Level

Monte Carlo simulation analysis was used to estimate the percentage of WET tests that would be declared toxic using TST as a function of different α levels, within-test control variability, and mean percent effect level. This analysis identified probable beta error rates (i.e., declaring an effluent toxic when in fact it is acceptable) as a function of α, mean effect at the IWC, and control CV. Using the RMDs discussed in Section 1.4, the lowest α level (with 0.05 being the

lowest α level used) was then identified for a given WET test design that also resulted in a β = 0.05 at a 10 percent mean effect in the effluent sample.

For each of the nine test methods examined, control CV was calculated on the basis of WET test data compiled as described in Section 2.2. Cumulative frequency plots were used to identify various percentiles of observed method-specific CVs (e.g., 25th, 50th, 75th percentiles). These measures were calculated to characterize typical achievable test performance in terms of control variability. A similar analysis was performed for the control endpoint responses for each of the nine test methods (e.g., mean offspring per female in the chronic *Ceriodaphnia dubia* test method) to characterize typical achievable test performance in terms of control response. The following describes the simulation analysis used to help identify appropriate alpha levels for each WET test method examined.

2.3.1 Simulation Analyses

In simulation analyses, sets of effluent and control WET test data were constructed having known properties with respect to different mean effect percentages and control CV as described below. Control CVs examined were based on CV percentiles observed in actual WET test data for a given WET test method. All simulation analyses were based on normally distributed WET test data and equal variances between the effluent and control for each scenario examined. These data were then analyzed using the one-tailed t-test published by Erickson and McDonald (1995) for bioequivalence testing (and mathematically defended in Erickson 1992 for normally distributed equal variance data) and the one-tailed traditional hypothesis t-test formulation (see Equations 1 and 2 below) to determine whether a given effluent was declared toxic using each approach at a specified α value. By simulating thousands of WET tests for a given scenario (mean percent effect and control CV and α level), the percentage of tests declared toxic could be calculated and compared among scenarios, and between the TST and the traditional hypothesis testing approach.

Equation 1: TST t-test assuming equal variances

$$t = \frac{\overline{Y_t} - b \times \overline{Y_c}}{S_p \sqrt{\frac{1}{n_t} + \frac{b^2}{n_c}}}$$

$$S_p = \sqrt{\frac{S_t{}^2 \times (n_t - 1) + S_c{}^2 \times (n_c - 1)}{(n_t + n_c - 2)}}$$

Equation 2: Traditional t-test assuming equal variances

$$t = \frac{\overline{Y_c} - \overline{Y_t}}{S_p \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

$$S_p = \sqrt{\frac{S_t{}^2 \times (n_t - 1) + S_c{}^2 \times (n_c - 1)}{(n_t + n_c - 2)}}$$

It is understood that using normally distributed data and equal variances is a simplification for some WET test methods that are prone to non-normally distributed data and heterogeneous variances (e.g., acute fathead minnow test method). Additional analyses suggested that the bioequivalence t-test of Erickson and McDonald (1995) results in a very small (< 0.01) departure of the nominal α error rate using TST with data that have even a nine-fold difference between control and effluent variances (which is greater than most variance ratios observed in nearly 2,000 WET tests) and with data that were non-normally distributed (Appendix A). Thus, results of simulation analyses should be applicable to the types of non-normality and variance heterogeneity encountered in WET tests. This was further supported by additional research showing that WET test data distributions are typically not highly skewed or long-tailed because of the way in which the tests are designed and because there are boundaries on test acceptability criteria that truncate the data within a test and the difference in variance one observes between control and an effluent treatment. A review of the statistical literature as well as additional analyses in developing the TST approach confirmed that Welch's t-test is appropriate for the types of non-normal data distributions encountered in actual effluent WET tests as well as for normally distributed data (see Appendix A).

Probabilities of accepting the null hypothesis for the traditional and TST approaches will differ according to different settings for a number of parameters, including population variances, test sample size, and effect size (i.e., fraction of the control response). Each of these factors was varied in simulation analysis as follows:

Population Variances: Population variances were defined by test method (control CVs in a large number of actual WET tests for a given method). The population mean was set to the median value of observed control mean values from actual effluent tests, and the CV value ranged from approximately the 10^{th} to 90^{th} percentile of the observed control CV range. N samples (representing the minimum number of replicates required in the test method) from the control population were selected for each simulation.

Effect Size: Population mean for the treatment group was defined by a specified effect size. Five different effect sizes (from 10 percent to 30 percent of the control mean) were evaluated for each treatment group. For example, when the control mean = 25 and the effect size = 10 percent, N samples (corresponding to the minimum number of replicates required in the test method) were picked at random from a population with mean = 25 × ([100 – 10] percent).

Sample Size (N): For certain WET test methods, sample size for each test method was increased up to double the minimum number of replicates required for a given test method. For example, number of replicates for the chronic *C. dubia* test ranged from 10 to 20 in simulation analyses. This analysis provided useful information indicating potential benefits to a permittee if they conducted a WET test method with additional replicates, given a specified mean percent effect level and control CV observed, and a specified α level.

Alpha Error: The maximum allowable Type I error (α) in TST was specified at different levels ranging from 0.05 to 0.30 (6 values). Results of these analyses indicated potential β error rates (probability of declaring a sample toxic when it is acceptable) given a specified mean percent effect in the effluent and control CV. These results were also compared with results using the traditional hypothesis testing approach and an $\alpha = 0.05$ (the EPA-recommended α level using the

traditional hypothesis testing approach) to compare β error rates using both approaches. While comparison of results between TST and the traditional approach were not used to set test method α levels, this analysis was useful in documenting whether the TST approach was as protective as the traditional approach using a given α level.

After N samples of control and effluent were randomly selected from specified populations, the traditional hypothesis testing approach and TST were conducted as specified in equations 1 and 2 above. The one tail probabilities of declaring the test toxic using the traditional hypothesis testing approach and the TST approach were calculated and saved. This simulation was repeated 10,000 times for each combination of effect levels, CV, and alpha level. The percent of tests declared toxic was then calculated for each simulation setting.

Once β error rates were identified for a WET method given different α levels, control CVs, and percent mean effect levels, bivariate plots were used to compare the percentage of tests declared toxic as a function of α and the ratio of effluent mean: control mean at various within-test variability percentiles (e.g., 25^{th}, 50^{th}, 75^{th}) and the RMD effect thresholds identified as either toxic (25 percent effect for chronic and 20 percent for acute) or negligible (10 percent mean effect). The results were then used to identify an appropriate α error rate for a test method given the RMDs noted in Section 1.4.

Finally, where there was sufficient effluent test data available, an analysis of actual effluent data was conducted using TST and the α level identified for the test method, and using the traditional hypothesis testing approach. Results of that analysis were used to estimate potential results if TST was used in the NPDES WET Program and to compare those results with those using the traditional hypothesis testing approach.

3.0 RESULTS

3.1 Chronic *Ceriodaphnia dubia* Reproduction Test

On the basis of actual WET data (N = 792 tests), the mean control reproduction ranged from 15.0 to 51.7, with a median mean value of 25.5 (Table 3-1). Control CVs ranged from 0.04 to 1.22 with a median value of 0.15 (Table 3-1). Using these data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in reproduction between the control and effluent concentration.

Table 3-1. Summary of mean control reproduction and control CV derived from analyses of 792 chronic *Ceriodaphnia dubia* WET tests

Percentile	Mean control reproduction	Control CV	Control SD
10th	17.7	0.08	2.07
25th	21.2	0.10	2.64
50th	25.5	0.15	3.79
70th	28.4	0.22	5.27
75th	29.4	0.24	5.82
85th	31.6	0.31	7.24
90th	33.3	0.35	8.41
95th	35.6	0.40	10.25

Identifying Test Method-Specific α

A summary of the simulation results is shown graphically in Figure 3-1. An alpha level of 0.20 satisfies both RMDs of (1) ensuring at least a 75 percent probability of declaring a 25 percent mean effect as toxic regardless of within-test control variability (denoted as effluent mean: control mean value of 0.75 on the x-axis of each graph in Figure 3-1), and (2) ensuring that a negligible effect (10 percent mean effect denoted as effluent mean: control mean value of 0.90) is declared toxic $\leq$ 5 percent of the time. Lower α levels (e.g., $\alpha = 0.10$) resulted in > 5 percent tests declared toxic when there was a 10 percent effect under average within-test CV values (i.e., $\beta > 0.05$). Note that using an $\alpha = 0.20$, a *Ceriodaphnia* test having a 20 percent mean effect at the IWC (effluent mean:control mean = 0.8) and median control variability (control CV = 0.15) will be declared toxic approximately 50 percent of the time using TST (Figure 3-1). Thus, as discussed in Section 1.3 and shown in Figure 1.2, some percentage of tests having an effluent mean effect less than the RMD threshold of 25 percent will be declared toxic using TST, even when the test control responds acceptably. Likewise, at an $\alpha = 0.20$, a *Ceriodaphnia* test exhibiting a 10 percent mean effect in the effluent (0.9 on the x-axis in Figure 3-1) and relatively high control variability (control CV = 0.25, 75th percentile for this WET test method) will have approximately a 25 percent probability of being declared toxic (Figure 3-1), even though a 10 percent mean effect is considered acceptable using TST.

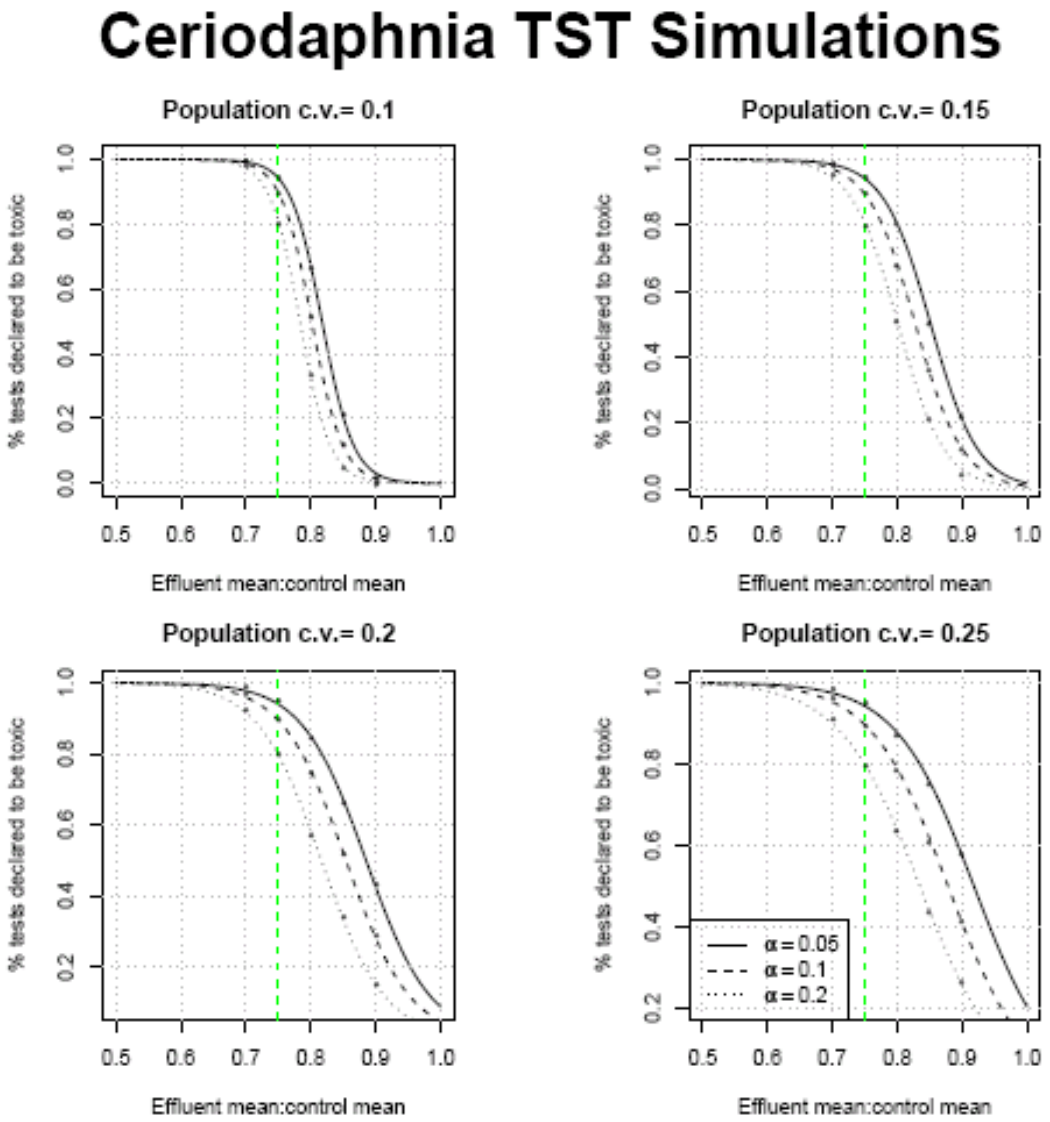

Figure 3-1. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs of 0.1, 0.15, 0.2, and 0.25 correspond to the approximate 25th, 50th, 70th, and 75th percentiles for the chronic *Ceriodaphnia* WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

The above results illustrate two features of the TST approach that should be understood: (1) At mean effect levels < the RMD toxicity threshold, there are differing probabilities of an effluent being declared toxic (i.e., different actual α error rates) depending on within-test variability and the difference in mean responses observed between control and IWC (see Figure 1-2). An effluent with a mean effect substantially lower than the RMD threshold of 25 percent will have some probability of being declared toxic. (2) For this WET test method and some others examined in this project, there is some probability of declaring a test non-toxic when the mean effect in the effluent exceeds the RMD threshold of 25 percent; e.g., at an $\alpha = 0.20$ and relatively

high within-test variability, a 30 percent mean effect in the effluent might not be declared toxic as much as 10 percent of the time.

The following examples give representative results of the simulation analysis, illustrating the effect of different alpha levels in terms of meeting RMDs for TST.

In the first example, there is a 10 percent mean effect in the effluent and a median level of within-test control precision (50th percentile CV of 0.15). Use of alpha levels ranging from 0.05 to 0.30 resulted in failure to reject the null hypothesis in ~20 percent to ~5 percent of tests, respectively, with α levels $\geq$ 0.20 meeting the RMD of $\beta \leq 0.05$ at a 10 percent mean effect level (Figure 3-2).

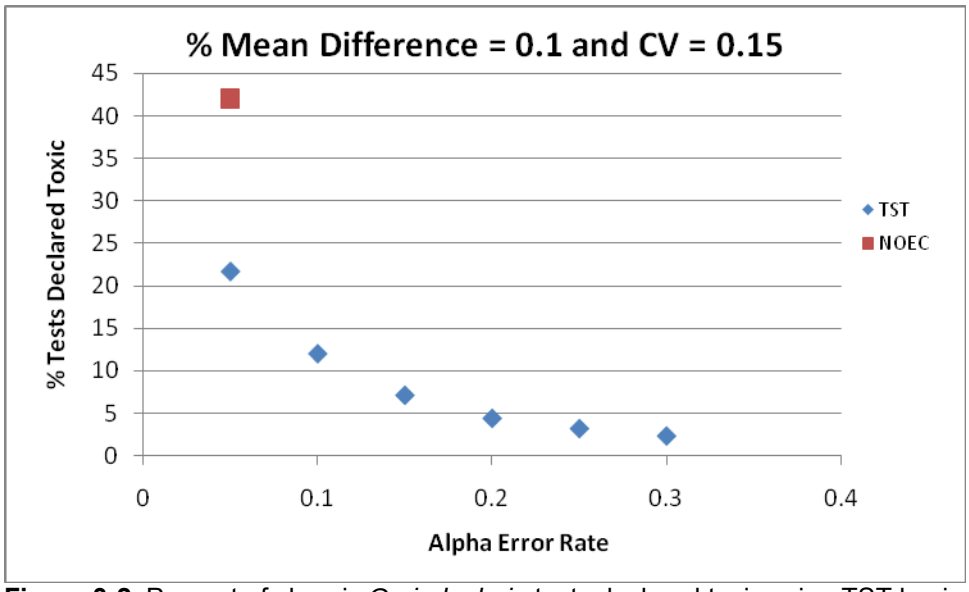

Figure 3-2. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

In a second example, the effluent has a mean effect of 25 percent and above average control CV (75th percentile). At α levels < 0.25, the percentage of tests declared toxic is $\geq$ 75 percent, meeting the RMD for false negative rate (α).

The rate at which tests were declared toxic was evaluated using both the traditional hypothesis testing approach with an alpha error rate of 0.05 (as recommended in the EPA WET test methods) and the TST approach with different alpha error rates. At a 50th percentile CV (0.15) and a mean effect of 10 percent, use of the TST approach results in fewer declared toxic tests relative to the traditional hypothesis approach at all alpha error rates examined (Figure 3-2). For tests with the same mean effect (10 percent) but higher control variability (CV = 0.25), TST yields a higher rate of tests declared toxic at alpha error rates of 0.05, 0.10, and 0.15 and approximately equivalent percent toxic tests at alpha error rates of 0.20 and 0.25 (Figure 3-2). Those results are in keeping with the RMD that tests with negligible (10 percent) mean effect in

the effluent are declared non-toxic most of the time but are declared to be toxic more frequently as test precision is poorer.

Tests with a mean effect of 25 percent and above average precision (CV = 0.25) result in a higher rate of tests declared toxic using TST than using the traditional hypothesis approach (Figure 3-3). This result is a direct consequence of the RMDs defined for TST but illustrate disincentives to collect more precise data using the traditional hypothesis approach currently used.

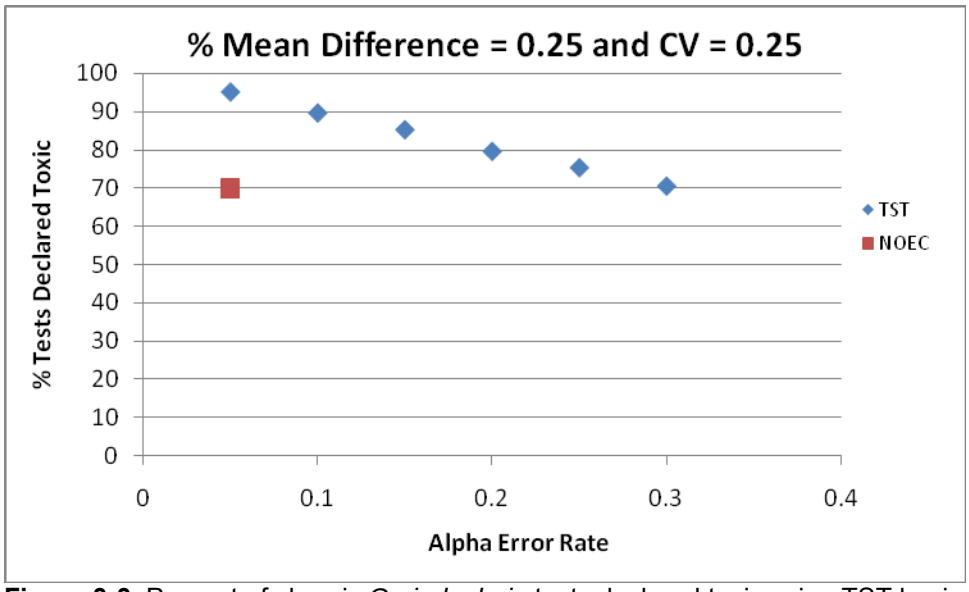

Figure 3-3. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and high control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Effect of Increased Number of Within-Test Replicates

One of the intended benefits of the TST approach is that increasing the precision and power of the test increases the chances of rejecting the null hypothesis and declaring a sample non-toxic when it meets the RMD for acceptability. This increases the ability of the permittee to *prove the negative* that a sample is acceptable. To demonstrate this benefit, the effect of increasing test replication on the TST β error rate (declaring a sample toxic when it is not) was explored using simulated data.

Increasing test replication with this method (and thereby the power of the test) results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach and a *lower* rate of tests declared toxic using the TST approach (e.g., Figure 3-4). For tests with a mean effect of 10 percent and a control CV of 0.25 (approximately 75th percentile for this method), slightly more tests will be declared toxic using the TST approach as compared to the traditional hypothesis testing approach when the minimum test design of 10 replicates is used for this WET method. If the number of within-test replicates is increased, the TST approach demonstrates an improved ability to declare such a test as acceptable. As the mean effect at the effluent approaches 25 percent, the percentage of tests declared toxic is less affected by increased replication using TST because the *b* value and α value were selected to identify a 25 percent mean effect in the IWC as

toxic ≥ 75 percent of the time. However, the percentage of tests declared toxic continues to increase using the traditional hypothesis approach even when there is a negligible effect (10 percent effect) of the effluent as defined by TST (Figure 3-5). Thus, increasing test replication increases TST's ability to confirm that an effluent is acceptable in tests with mean effect less than 25 percent.

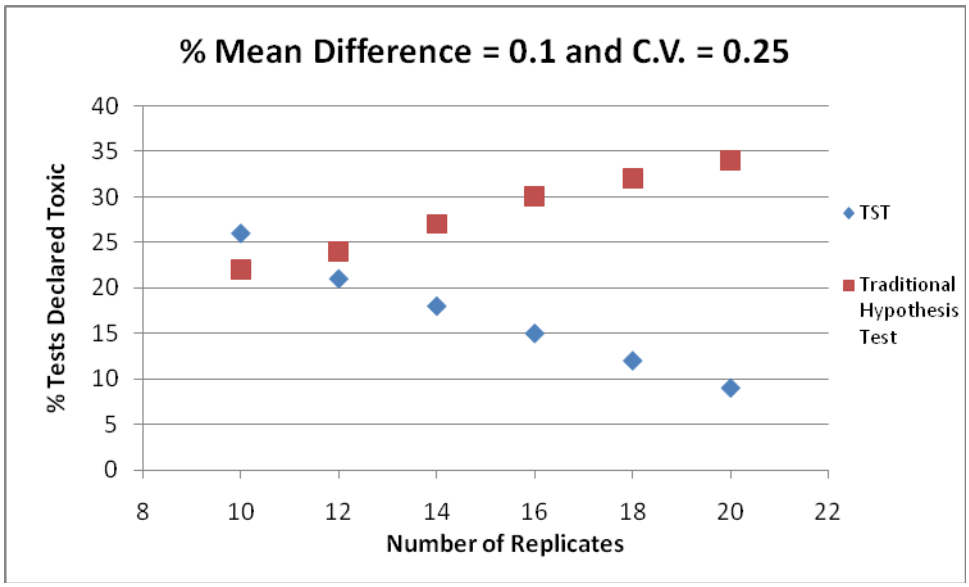

Figure 3-4. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and above average control variability and α = 0.20, as a function of the number of test replicates. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Effluent Data Results

Results from actual effluent tests were compared between TST and the traditional hypothesis testing approach for those tests having control CV between 0.15–0.24 (Table 3-2). At a mean effect of 10–15 percent at the IWC (N = 48), TST declared a lower percentage of tests toxic than the traditional hypothesis testing approach. This result is consistent with the RMD that a 10 percent mean effect should be declared acceptable much (95 percent) of the time. However, when the mean effect was greater than 25 percent (N = 303), TST declared 100 percent of the tests toxic while the traditional hypothesis testing approach did not. This result is also consistent with the TST goal that as the mean effect approaches 25 percent at least 75 percent of the tests should be declared toxic. This result also indicates that given the effluent data available, TST is at least as protective as the traditional hypothesis approach currently used.

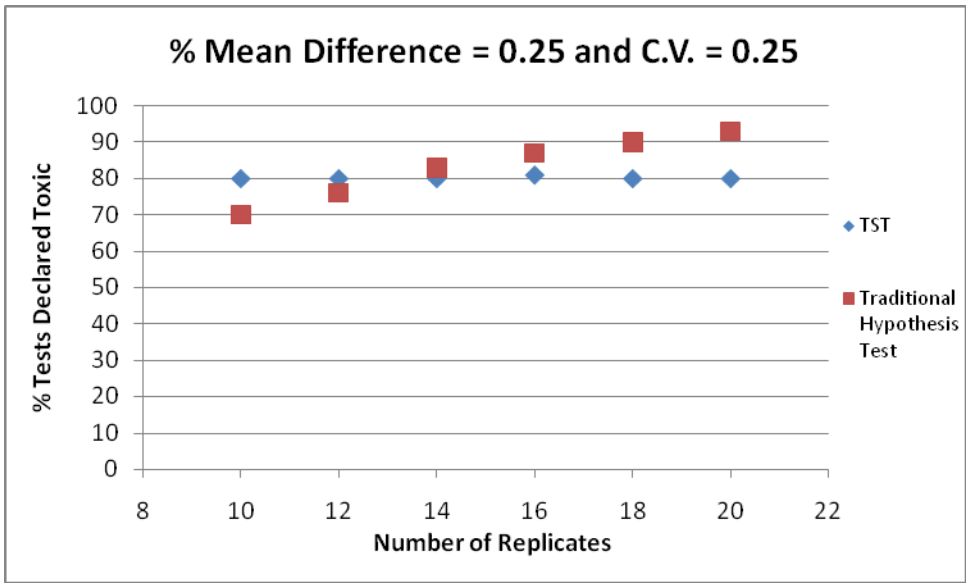

Figure 3-5. Percent of *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability (α = 0.20) as a function of the number of test replicates. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Table 3-2. Comparison of the percentage of chronic effluent *Ceriodaphnia* tests declared toxic using TST versus the traditional hypothesis testing approach

% Mean effect	N	% tests toxic using TST	% Tests toxic using traditional hypothesis testing approach
10–15	48	6.2	18.7
20–30	48	100	87.5
> 25	303	100	95.2

3.2 Chronic *Pimephales promelas* Growth Test

On the basis of actual WET data (N = 472 tests), the mean control growth ranged from 0.31 to 1.30, with a median mean value of 0.62 (Table 3-3). Control CVs ranged from 0.03 to 0.50 with a median value of 0.09 (Table 3-3). Using these data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in growth between the control and effluent concentration.

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-6), an alpha error rate of 0.25 is appropriate for use in applying the TST approach to analysis of two concentration chronic *P. promelas* data because using that alpha error rate satisfies both RMDs of (1) ensuring at least an 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time.

Table 3-3. Summary of mean control growth and control CV derived from analyses of 472 chronic *Pimephales promelas* WET tests

Percentile	Mean control growth	Control CV	Control SD
10th	0.34	0.04	0.02
25th	0.43	0.06	0.03
50th	0.62	0.09	0.05
70th	0.76	0.12	0.07
75th	0.79	0.13	0.08
85th	0.86	0.16	0.10
90th	0.89	0.17	0.11
95th	0.94	0.21	0.13

As noted for the *Ceriodaphnia* chronic test in Section 3.1, the Type I error rate will vary from the RMD Type I error rate of 0.25 depending on the level of toxicity observed in the effluent and control variability within a test. When toxicity is > 25 percent mean effect in the effluent, the Type I error rate is lower. However, as noted in Section 1.3, there is some probability (< 10 percent) that a mean effect > 25 percent in the IWC will be declared non-toxic depending on within-test variability. Likewise, a reasonable percentage (as much as 50 percent) of tests having a mean effect = 15 percent in the effluent will be declared toxic using the TST approach, again depending on within-test variability: the greater the within-test variability the greater the probability of declaring toxicity at mean effect levels below the toxicity decision threshold of 25 percent.

For example, at a 10 percent mean effect in the effluent and above average within-test control variability (between the 50th and 75th percentile, CV of 0.11), use of an alpha level of 0.25 results in failure to reject the null hypothesis ~5 percent of the time (Figure 3-7). Lower alpha levels resulted in a higher percentage of tests declared toxic at that mean effect level and CV range (Figure 3-6). That indicates that using an alpha = 0.25 for this test method, TST achieves the RMD of correctly identifying an acceptable sample (based on the RMD that a 10 percent mean effect is negligible). However, less precise tests (but still well within normal test method performance) result in less ability to reject the null hypothesis that the sample is toxic and the rate of tests declared toxic increases even at a percent mean effect of 10 percent (Figure 3-6). For tests with a mean effect of 25 percent (the RMD toxicity threshold) and alpha error rate of 0.25, 75 percent of the tests are declared toxic as expected (Figure 3-8).

Effect of Increased Number of Within-Test Replicates

As expected, increasing test replication (and thereby the power of the test) results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach and a *lower* rate of tests declared toxic using the TST approach and chronic *P. promelas* test data (e.g., Figure 3-9). For tests with a mean effect of 10 percent in the effluent and a control CV of 0.15 (slightly greater than the 75th percentile for this method), slightly more tests are declared toxic using the TST approach as compared to the traditional hypothesis testing approach when the minimum test design of four replicates is used for this WET endpoint. If replicates are added to the test design, the TST approach demonstrates an increased ability to declare the results acceptable. As the mean effect approaches 25 percent, the percentage of tests declared toxic is less affected by

increased replication using TST because a 25 percent effect is the RMD used to define *b* and the null hypothesis. However, the percentage of tests declared toxic continues to increase using the traditional hypothesis testing approach even when there is a 10 percent effect of the effluent. Thus, increasing test replication increases TST's ability to confirm an acceptable effluent when the mean effect is less than 25 percent in the effluent.

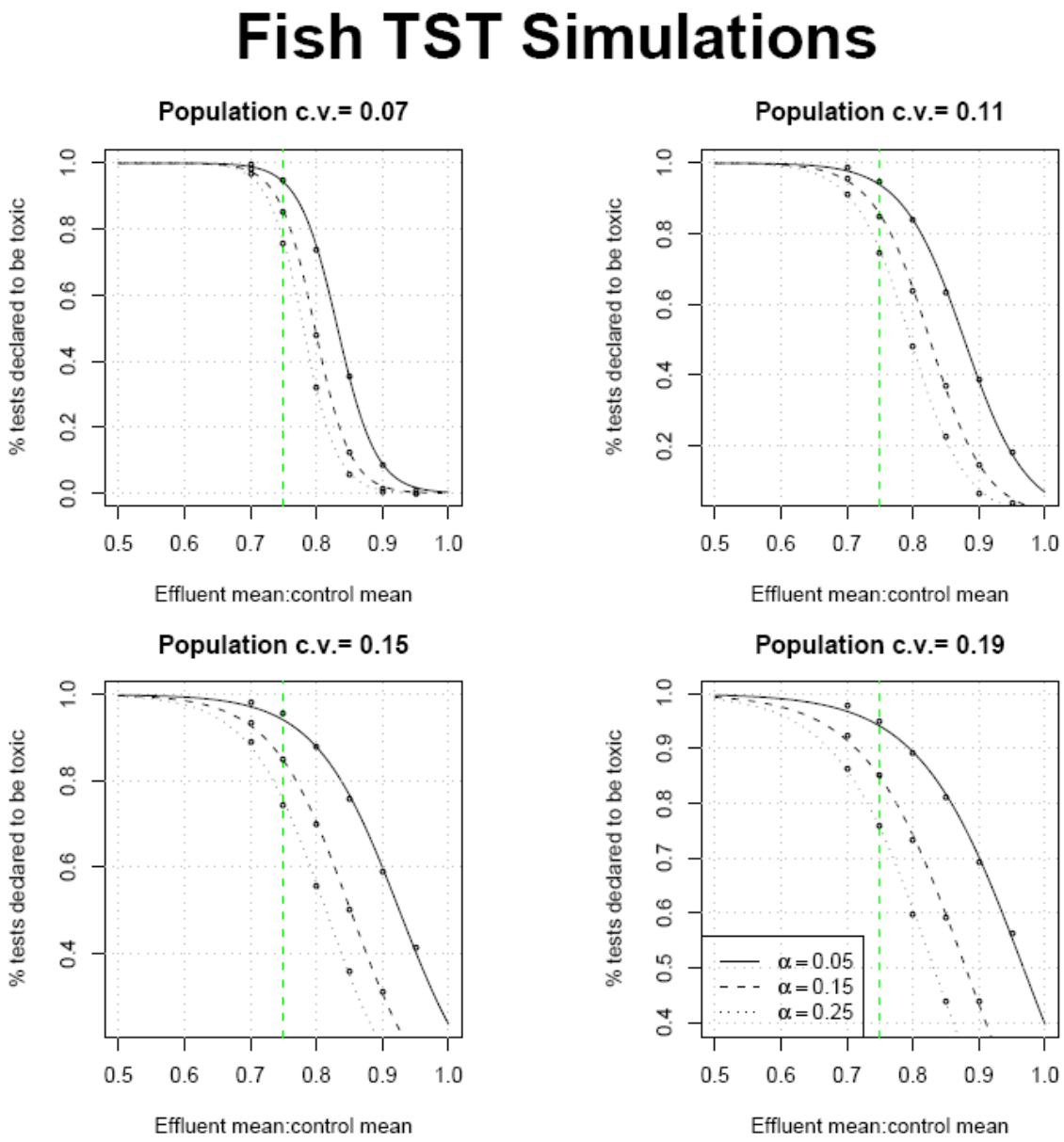

Figure 3-6. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic fathead minnow WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

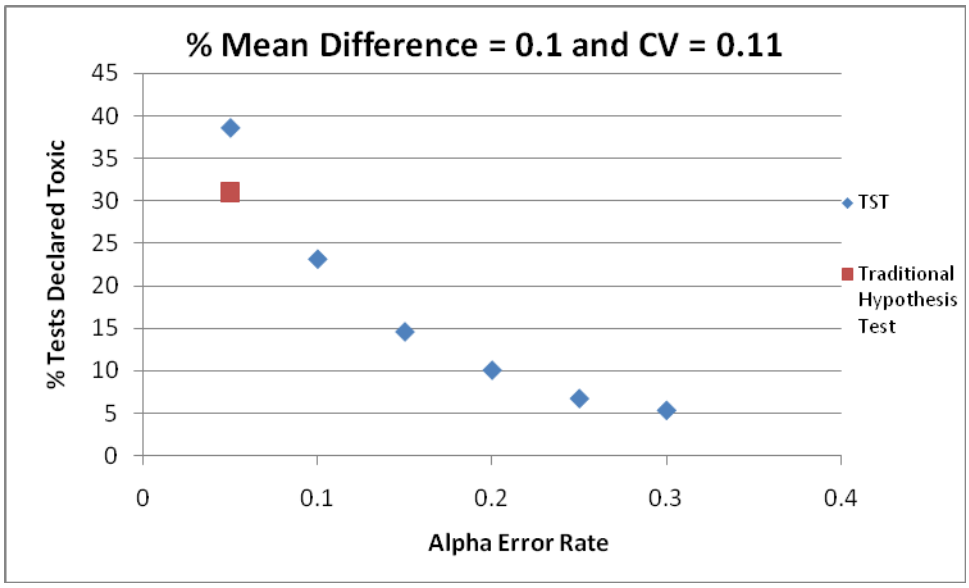

Figure 3-7. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional approach (α = 0.05) is shown as well.

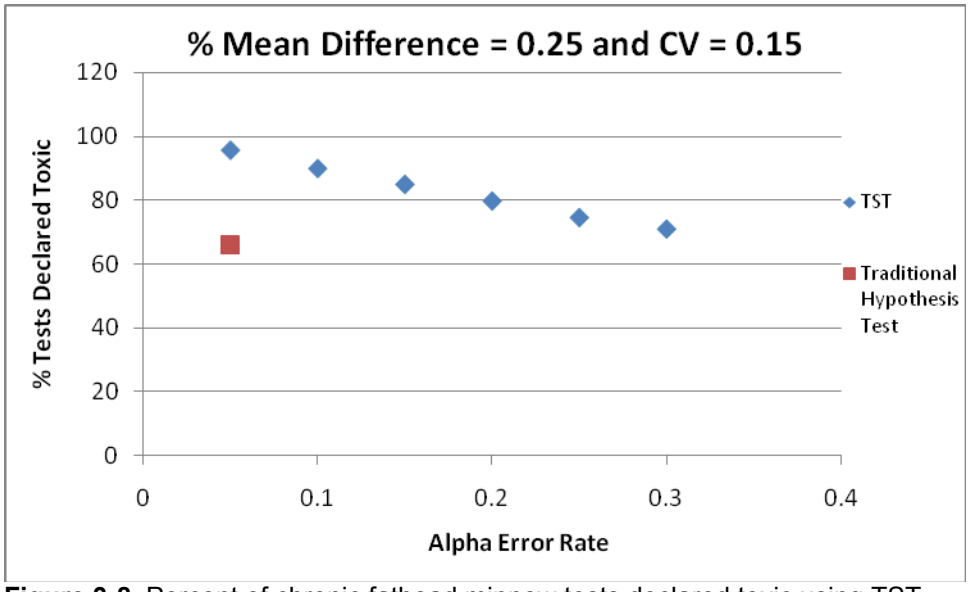

Figure 3-8. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional approach (α = 0.05) is shown as well.

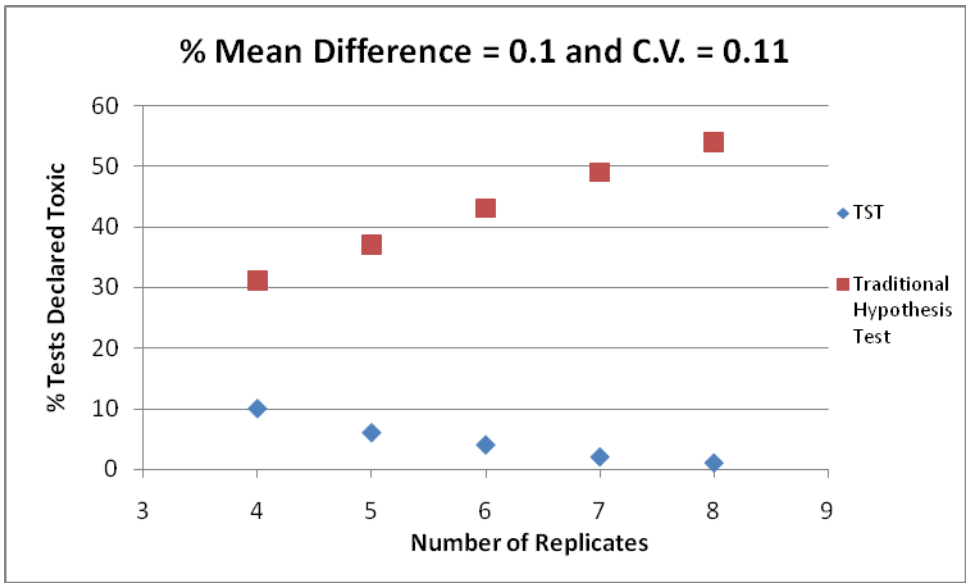

Figure 3-9. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability and an α = 0.25, as a function of the number of test replicates. Result using the traditional approach (α = 0.05) is shown as well.

Effluent Data Results

Results from actual effluent tests were compared between TST and the traditional hypothesis testing approach for those tests having control CV between 0.09–0.13 (Table 3-4). At a mean effect of 10–15 percent (N = 58), TST declared none of the tests toxic while the traditional hypothesis testing approach declared nearly all of the tests toxic. However, if the mean effect is greater than 25 percent (N = 136), both approaches declared 100 percent of the tests toxic. Those results indicate that TST is as protective as the current hypothesis testing approach for those tests when the TST RMD threshold for toxicity is exceeded.

Table 3-4. Comparison of the percentage of chronic effluent fathead minnow tests declared toxic using TST versus the traditional hypothesis testing approach

% Mean effect	N	% tests toxic using TST	% tests toxic using traditional hypothesis testing approach
10–15	58	0	98
> 25	136	100	100

3.3 Chronic *Americamysis bahia* Growth Test

On the basis of actual WET data (N = 210 tests), the mean control growth ranged from 0.20 to 0.66, with a median value of 0.30 (Table 3-5). Control CVs ranged from 0.07 to 0.87 with a median value of 0.14 (Table 3-5). Using those data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in growth between the control and effluent concentration.

Table 3-5. Summary of mean control growth and control CV derived from analyses of 210 chronic *Americamysis bahia* WET tests

Percentile	Mean control growth	Control CV	Control SD
10th	0.22	0.08	0.02
25th	0.25	0.10	0.03
50th	0.30	0.14	0.04
70th	0.36	0.17	0.06
75th	0.38	0.18	0.06
85th	0.41	0.22	0.07
90th	0.43	0.27	0.08
95th	0.47	0.35	0.11

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-10), an alpha error rate of 0.15 is appropriate for use in applying the TST approach to analysis of chronic mysid data because using this alpha error rate satisfies both RMDs of (1) ensuring at least an 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average or better than average test performance.

For example, at a 10 percent mean effect in effluent and an approximate median level of precision (50^{th} percentile CV of 0.14), an alpha level of 0.15 or greater resulted in failure to reject the null hypothesis in ≤ 5 percent of tests (Figure 3-11). For tests with a mean effect of 25 percent, the rate of tests declared toxic > 75 percent is achieved for alpha values ≤ 0.25 (Figure 3-12).

At a ~50^{th} percentile CV (0.13) and a mean effect of 10 percent, use of the TST approach results in significantly fewer toxic tests relative to the traditional hypothesis testing approach at all alpha error rates (Figure 3-11). For tests with the same mean effect (10 percent) but lower control precision (CV = 0.18), TST yields a higher rate of tests declared toxic at an alpha error rate of 0.05 and approximately equivalent percent toxic tests at a alpha error rate of 0.10.

Tests with a mean effect of 25 percent and above average precision (CV = 0.18) result in a high rate of tests declared toxic (Figure 3-12). The results are in agreement with the RMDs of the TST: As the mean effect approaches 25 percent, a greater proportion of the tests are determined to be toxic. Further, the less precise the test control data, the greater the rate of tests declared toxic (i.e., fail to reject the null hypothesis).

Effect of Increased Number of Within-Test Replicates

As expected, increasing test replication (and thereby the power of the test) results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach and a *lower* rate of tests declared toxic using the TST approach at a negligible effect of 10 percent, as shown in the example using chronic *A. bahia* test data (e.g., Figure 3-13). If replicates are added to the test design, the TST approach demonstrates an increased ability to declare such a test as non-toxic. As the mean effect approaches 25 percent, the percentage of tests declared toxic is less affected by increased replication using TST because a 25 percent effect is the RMD toxicity threshold identified in TST. However, the percentage of tests declared toxic continues to increase using the

traditional hypothesis testing approach even when there is a negligible effect (10 percent effect as defined by TST) of the effluent. Thus, increasing test replication increases TST's ability to confirm an acceptable level of toxicity in tests with mean effect less than 25 percent.

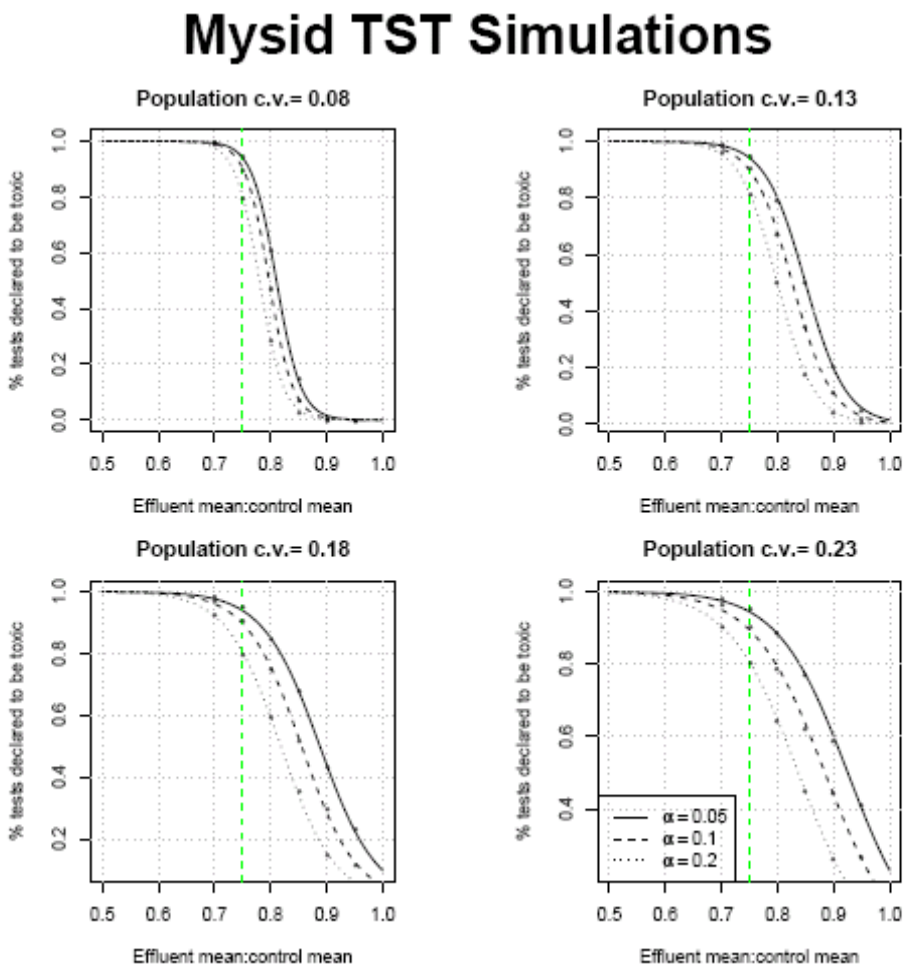

Figure 3-10. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 70th, and 90th percentiles for the chronic mysid WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

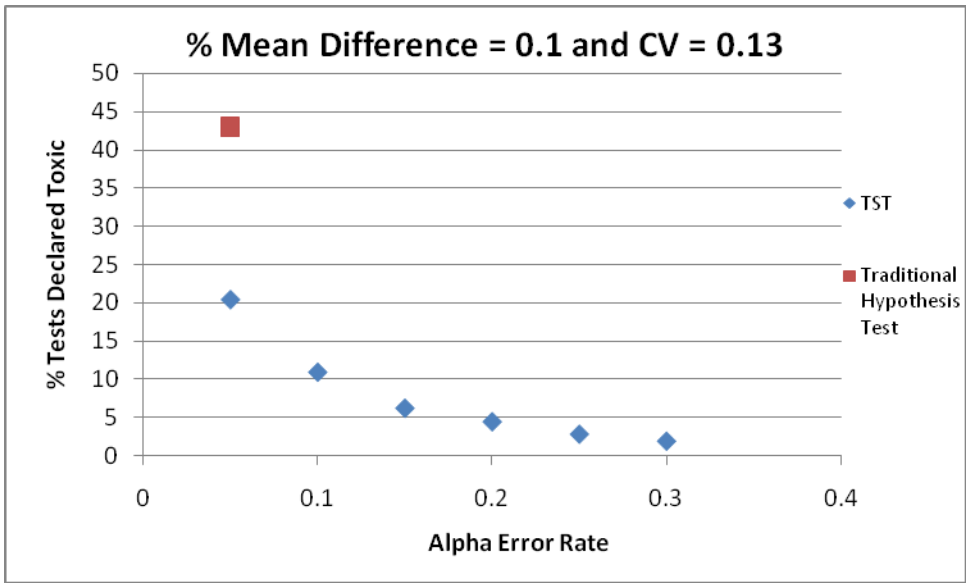

Figure 3-11. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

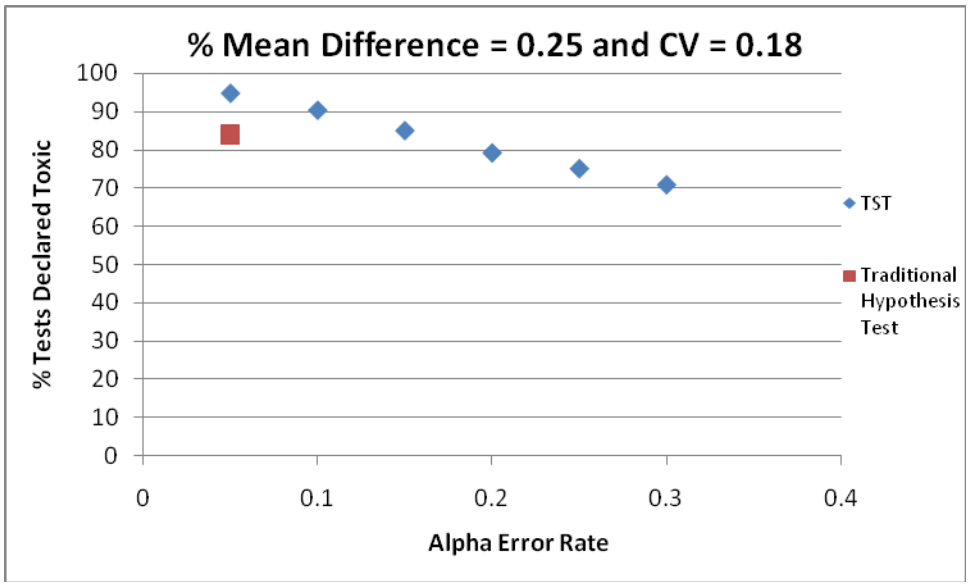

Figure 3-12. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

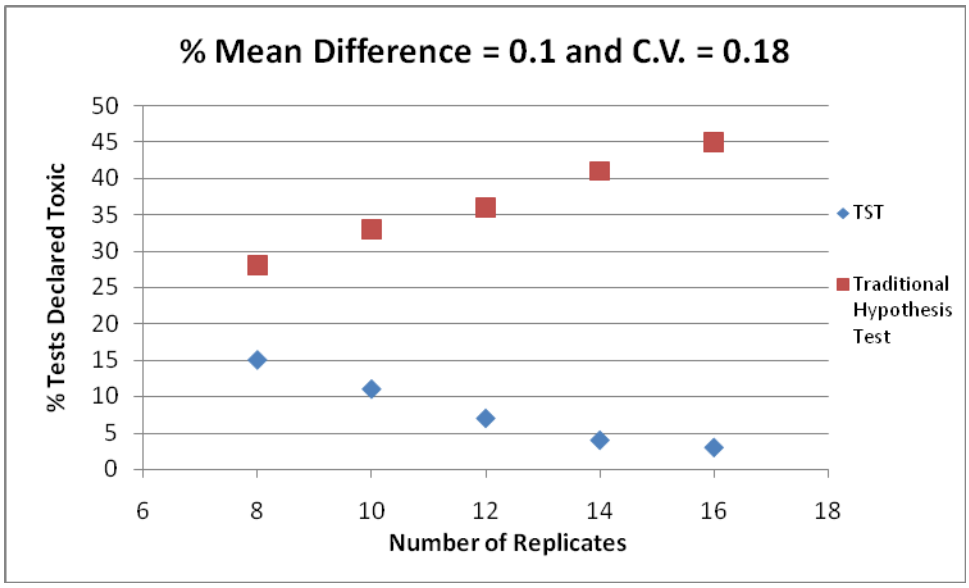

Figure 3-13. Percent of chronic mysid tests having a mean effluent effect of 10 percent and above average control variability declared toxic using TST and an α = 0.15, as a function of the number of test replicates. Results using the traditional hypothesis approach (α = 0.05) are shown as well.

Effluent Data Results

Results from actual effluent tests were compared between TST and the traditional hypothesis testing approach for those tests having control CV between 0.14–0.26 (75^{th} – 90^{th} percentile; Table 3-6). At a mean effect of 5–15 percent (N = 52), TST declared a lower percentage of tests toxic than the traditional hypothesis approach. That is expected because 10 percent mean effect in the effluent is considered negligible. However, when the mean effect in the effluent is greater than 25 percent (N = 95), both approaches declared 100 percent of the tests toxic.

Table 3-6. Comparison of percentage of chronic effluent mysid shrimp tests declared toxic using TST versus the traditional hypothesis testing approach

% Mean effect	N	% tests toxic using TST	% tests toxic using traditional hypothesis testing approach
5-15	52	1.9	11.5
> 25	95	100	100

3.4 Chronic *Haliotis rufescens* Larval Development Test

From actual WET data (N = 136 reference toxicant tests), mean control larval development ranged from 0.800 to 1.000, with a median mean value of 0.938 (Table 3-7). Control CVs ranged from 0.000 to 0.333 with a median value of 0.03 (Table 3-7). Using those data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in larval development between the control and effluent concentration.

Identifying Test Method-Specific α

On the basis of simulation results and power analyses (Figure 3-14), an alpha error rate of 0.05 is appropriate for use in applying the TST approach to analysis of chronic *H. rufescens* data because using this alpha error rate satisfies both RMDs of (1) ensuring at least an 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect (≤ 10 percent mean effect) is declared toxic ≤ 5 percent of the time (Figure 3-14). Note that higher alpha levels would also satisfy the above RMDs; however, as noted in Section 1.4, the Type I error rate is set as close to 0.05 as practicable given routine control performance.

Table 3-7. Summary of mean control larval development and control CV derived from analyses of 136 chronic red abalone WET tests

Percentile	Mean control larval development	Control CV	Control SD
10th	0.839	0.02	0.01
25th	0.900	0.02	0.02
50th	0.938	0.03	0.03
70th	0.961	0.04	0.04
75th	0.968	0.05	0.04
85th	0.977	0.06	0.05
90th	0.982	0.06	0.06
95th	0.988	0.07	0.07

At a 10 percent mean effect in the effluent, for example, and ~80th percentile CV of 0.05, alpha levels ranging from 0.05 to 0.30 result in failure to reject the null hypothesis in none of the tests (Figure 3-15). The rate of rejection of the null hypothesis using TST decreases only slightly with increasing CV. This result is indicative of the low within-test control variability routinely achieved using this WET test method.

For tests with a mean effect of 25 percent, the rate of tests declared toxic ranges from ~95 to ~70 percent, at approximately the 80th percentile CV value for alpha levels ranging from 0.05 to 0.30, respectively (Figure 3-16). Thus, at an alpha = 0.05, the rate of tests declared toxic at a 25 percent mean effect in the effluent meets the RMD.

At ~80th percentile CV (0.05) and a mean effect of 10 percent, use of the TST approach results in significantly fewer toxic tests relative to the traditional hypothesis approach at all alpha error rates (Figure 3-15). Those results are in keeping with the RMD of the TST approach; tests with a negligible (10 percent) mean effect of the effluent are declared non-toxic 95 percent of the time when test control data have average precision.

Tests with a mean effect of 25 percent and above average precision (CV = 0.05) resulted in an equivalent rate of tests declared toxic as the traditional hypothesis approach when the TST α = 0.05 (Figure 3-16). The results further support the selection of TST α = 0.05 for this test method.

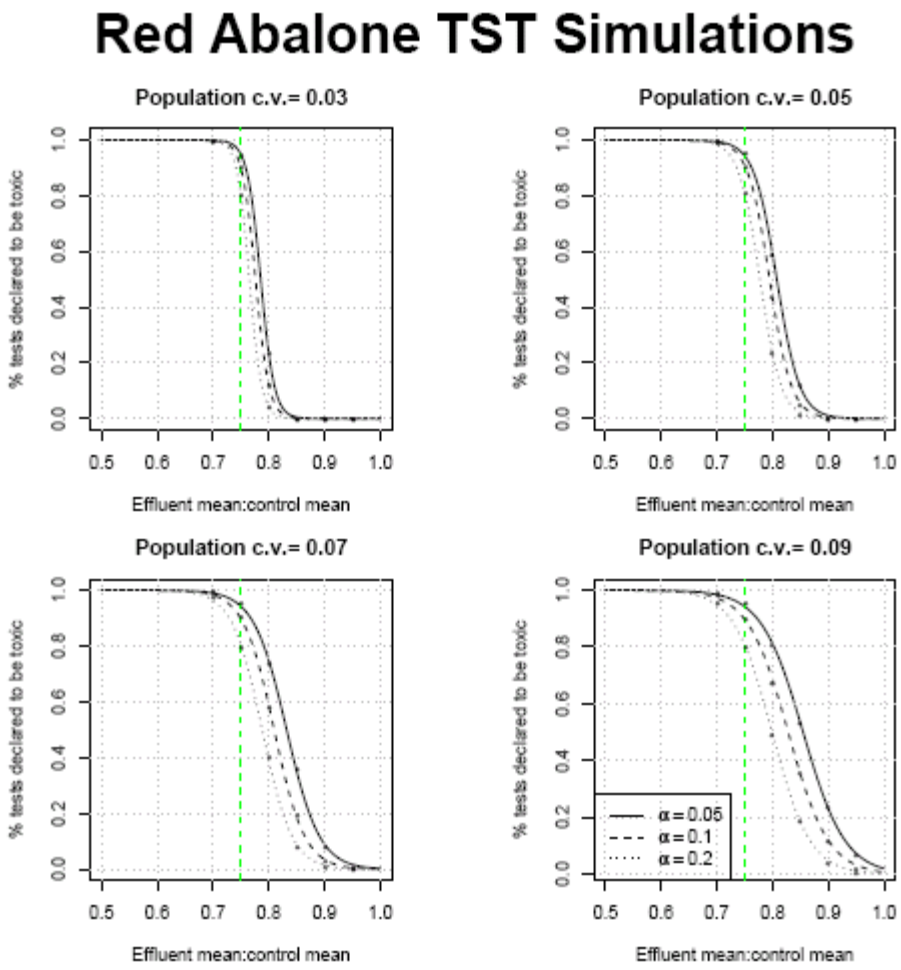

Figure 3-14. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 98th percentiles for the chronic red abalone WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

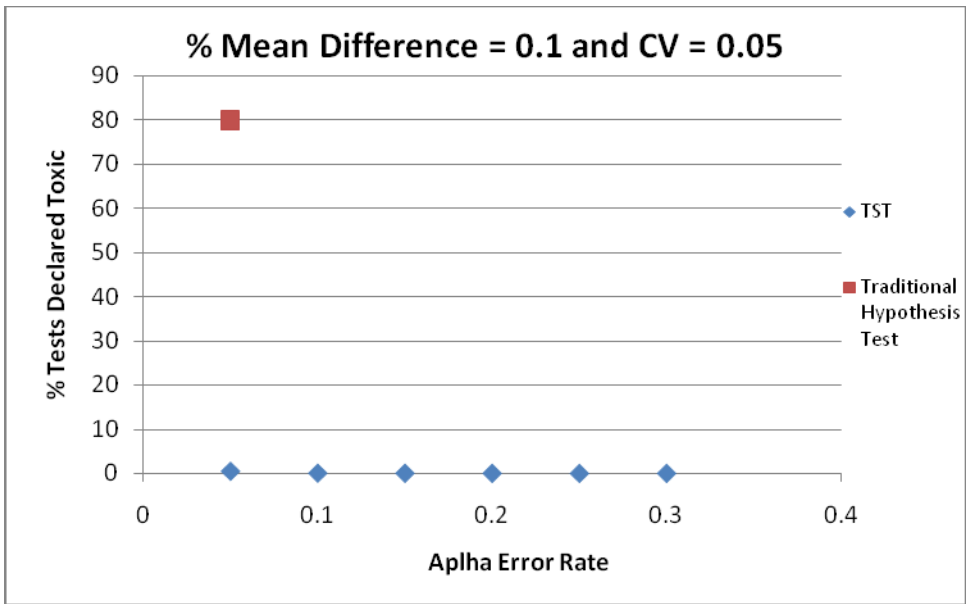

Figure 3-15. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

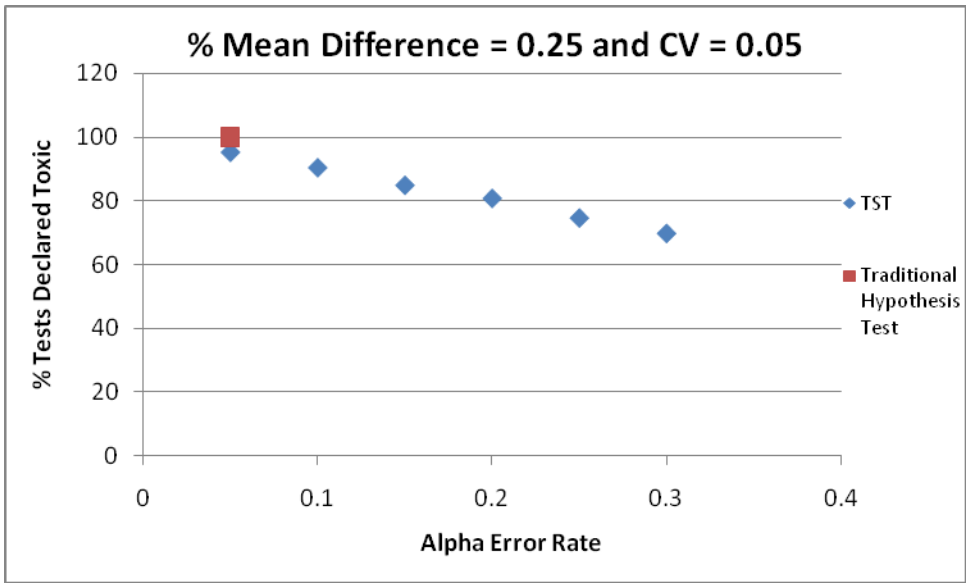

Figure 3-16. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

3.5 Chronic *Macrocystis pyrifera* Germination Test

On the basis of actual WET data (N = 135 reference toxicant tests), mean control germination ranged from 0.700 to 0.985, with a median mean value of 0.908 (Table 3-8). Control CVs ranged

from 0.006 to 0.560 with a median value of 0.04 (Table 3-8). Using that data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in germination between the control and effluent concentrations.

Table 3-8. Summary of mean control germination and control CV derived from analyses of 135 chronic giant kelp WET tests

Percentile	Mean control germination	Control CV	Control SD
10th	0.783	0.02	0.02
25th	0.859	0.03	0.02
50th	0.908	0.04	0.03
70th	0.936	0.05	0.04
75th	0.940	0.05	0.05
85th	0.958	0.07	0.06
90th	0.965	0.07	0.06
95th	0.973	0.10	0.09

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-17), an alpha error rate of 0.05 is appropriate for use in applying the TST approach to analysis of chronic *M. pyrifera* germination data because using this alpha error rate satisfies both RMDs of (1) ensuring at least an 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average test performance. As noted above for the Abalone test method, higher alpha levels also satisfy the above RMDs; however, an alpha level of 0.05 is selected because it is more protective at effect levels > 25 percent.

At a 10 percent mean effect in the effluent for example, and routine, achievable control precision (~75th percentile CV of 0.05), alpha levels ranging from 0.05 to 0.30 resulted in failure to reject the null hypothesis in none of tests (Figure 3-18). Thus, for this test endpoint, low within-test control variability is routinely achieved.

For tests with a mean effect of 25 percent, the rate of tests declared toxic ranges from ~95 percent to ~70 percent, at alpha levels ranging from 0.05 to 0.30, respectively, and approximately the 75th percentile CV level (Figure 3-19). All alpha levels < 0.25 achieved the RMD that a 25 percent mean effect is declared toxic at least 75 percent of the time.

At ~75th percentile CV (0.05) and a mean effect of 10 percent, use of the TST approach results in significantly fewer tests declared toxic relative to the traditional hypothesis approach at all alpha error rates (Figure 3-18). Those results are because the RMD for effluent acceptability (10 percent mean effect) is designed to be met ≥ 95 percent of the time.

Tests with a mean effect of 25 percent and above average precision (CV = 0.05) result in a similar rate of tests declared toxic (Figure 3-19) as the traditional hypothesis approach when the TST $\alpha = 0.05$. The results further support the selection of TST $\alpha = 0.05$ for this test endpoint.

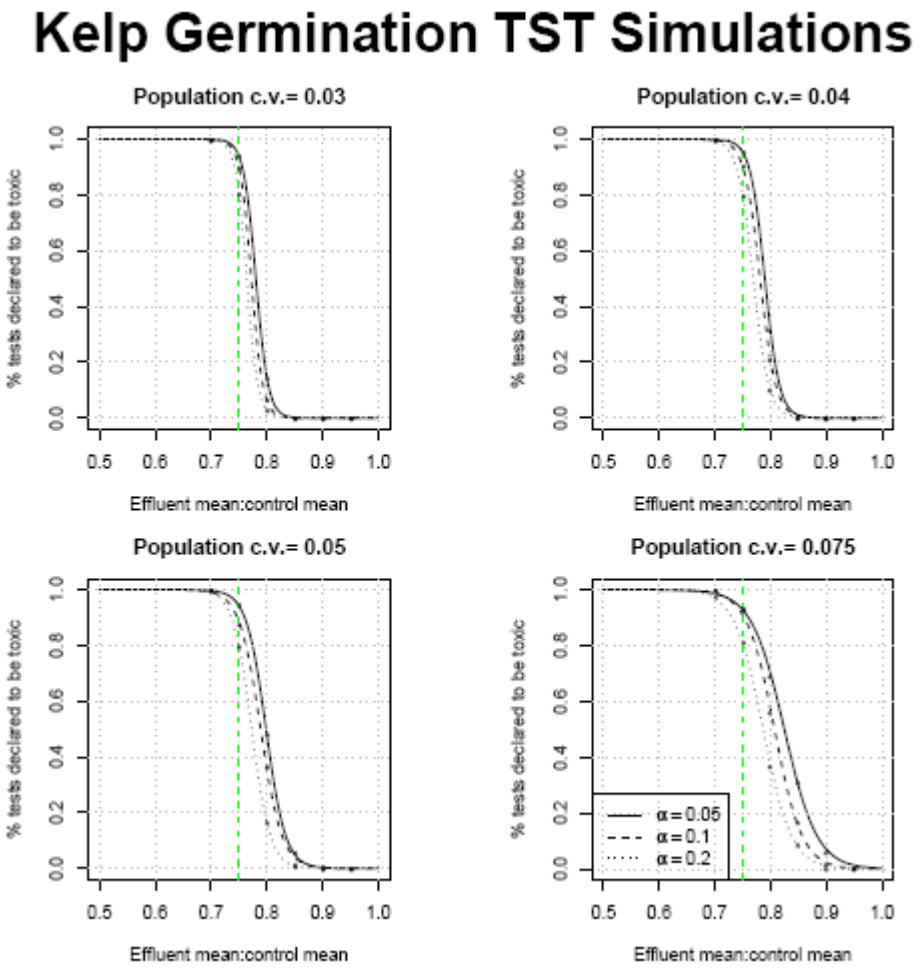

Figure 3-17. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 95th percentiles for the chronic giant kelp germination WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

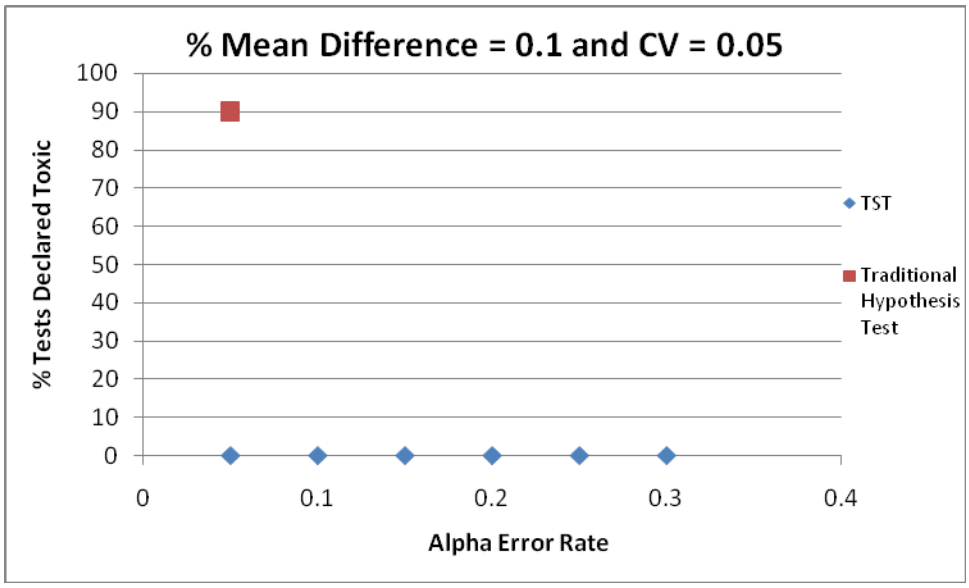

Figure 3-18. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

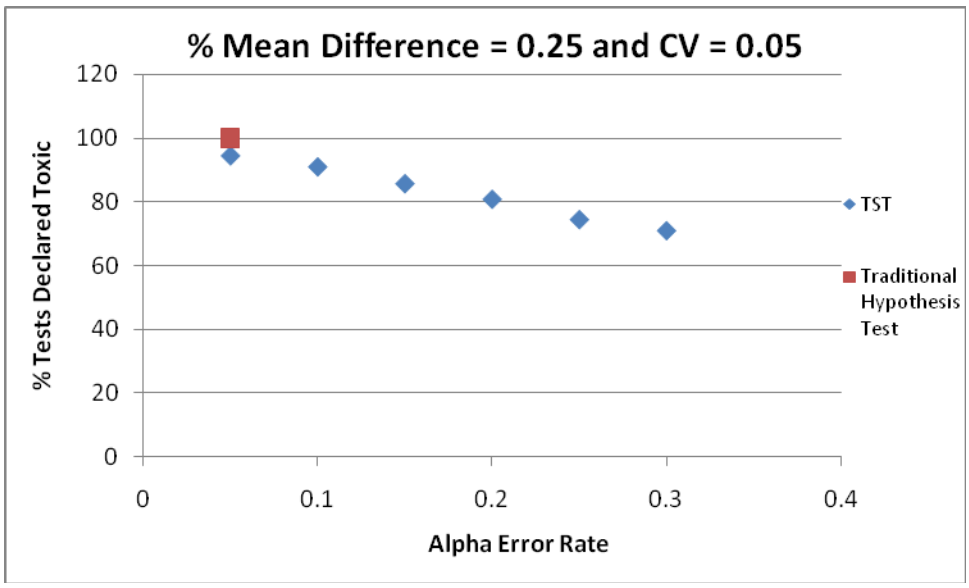

Figure 3-19. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

3.6 Chronic *Macrocystis pyrifera* Germ-tube Length Test

On the basis of actual WET data (N = 135 reference toxicant tests), the mean control germ-tube length ranged from 10.200 to 20.778, with a median mean value of 14.014 (Table 3-9). Control CVs ranged from 0.009 to 0.189 with a median value of 0.073 (Table 3-9). Using that data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), CVs, and percent mean effect in germ-tube length between the control and effluent concentration.

Table 3-9. Summary of mean control germ-tube length and control CV derived from analyses of 135 chronic *Macrocystis pyrifera* WET tests

Percentile	Mean control germ-tube length	Control CV	Control SD
10th	11.965	0.03	0.46
25th	12.704	0.05	0.71
50th	14.014	0.07	1.04
70th	15.210	0.09	1.22
75th	15.554	0.09	1.29
85th	16.848	0.11	1.54
90th	17.568	0.12	1.74
95th	18.694	0.14	1.89

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-20), an alpha error rate of 0.05 is appropriate for use in applying the TST approach to analysis of chronic *M. pyrifera* tube-length data because using that alpha error rate satisfies both RMDs of (1) ensuring at least a 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average test performance. As noted for the germination endpoint of this species above, higher alpha levels would also satisfy these RMDs; however, in such cases, the lowest alpha $\geq$ 0.05 is selected.

At a 10 percent mean effect in the effluent for example and ~50th percentile CV of 0.07, alpha levels ranging from 0.05 to 0.30 resulted in failure to reject the null hypothesis in almost none of the tests (Figure 3-21). For tests with a mean effect of 25 percent, the rate of tests declared toxic ranged from ~95 to ~70 percent, at alpha error rates ranging from 0.05 to 0.30, respectively, and the 75th percentile CV value (Figure 3-22). Thus, alpha levels < 0.25 achieved the RMD that a 25 percent mean effect is declared toxic at least 75 percent of the time.

At ~50th percentile CV (0.07) and a mean effect of 10 percent, use of the TST approach results in significantly fewer tests declared toxic relative to the traditional hypothesis approach at all alpha error rates examined (Figure 3-21). These results are because of the RMDs of the TST approach; tests with a small (10 percent) mean effect of the effluent are declared non-toxic most of the time when test control data are average or better.

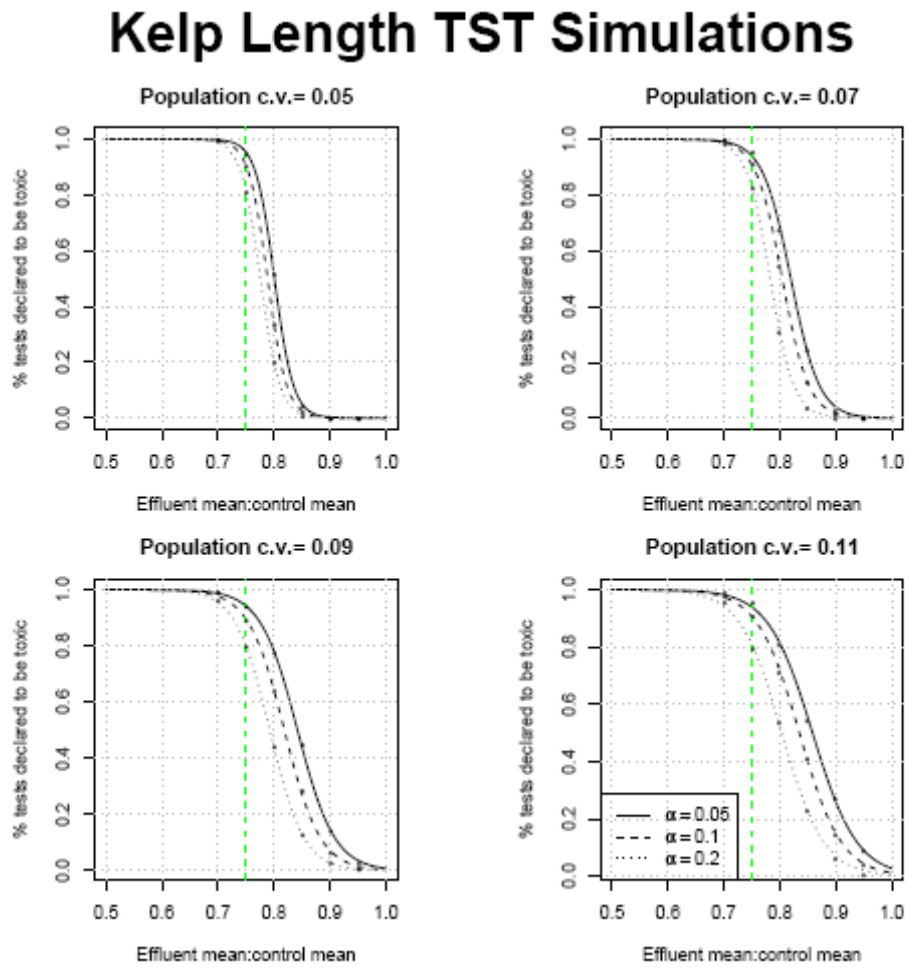

Figure 3-20. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic giant kelp germ-tube length WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

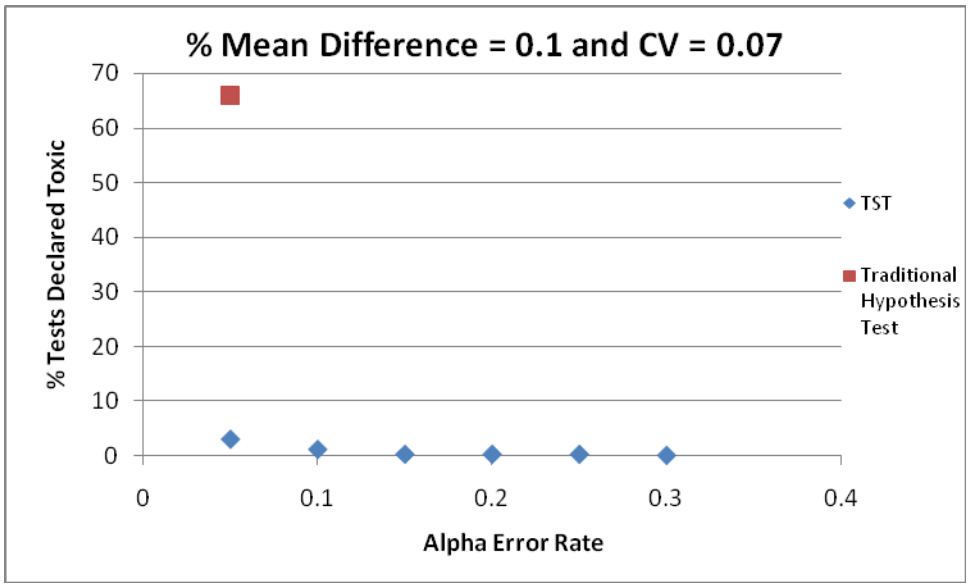

Figure 3-21. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

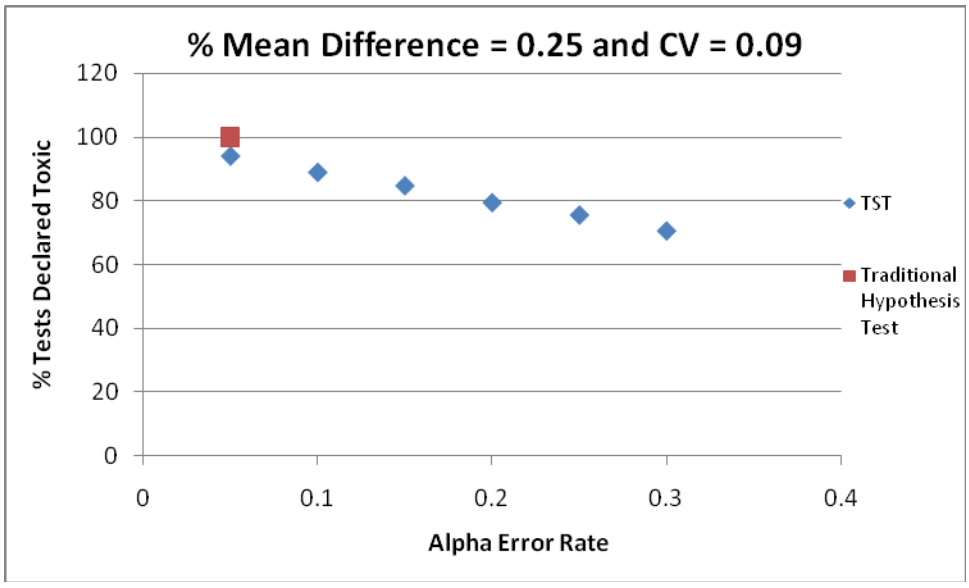

Figure 3-22. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Tests with a mean effect of 25 percent and above average precision (CV = 0.09) result in a similar rate of tests declared toxic as the traditional approach when alpha = 0.05 (Figure 3-22). These results further support the selection of 0.05 as the alpha value under TST for this WET endpoint.

3.7 Chronic Echinoderm Fertilization Test

On the basis of actual WET data (N = 177 tests), mean control fertilization ranged from 0.538 to 1.000, with a median mean value of 0.953 (Table 3-10). Control CVs ranged from 0.000 to 0.667 with a median value of approximately 0.03 (Table 3-10). Using that data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.3), CVs, and percent mean effect in reproduction between the control and effluent concentration of concern.

Table 3-10. Summary of mean control fertilization and control CV derived from analyses of 177 chronic *Dendraster excentricus and Strongylocentrotus purpuratus* WET tests

Percentile	Mean control fertilization	Control CV	Control SD
10th	0.826	0.01	0.58
25th	0.875	0.01	1.16
50th	0.953	0.03	2.45
70th	0.975	0.05	4.32
75th	0.978	0.07	5.97
85th	0.990	0.09	7.44
90th	0.993	0.11	9.32
95th	0.996	0.14	11.00

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-23), an alpha error rate of 0.05 is appropriate for use in applying the TST approach to analysis of chronic *D. excentricus* and *S. purpuratus* data because using this alpha error rate satisfies both RMDs of (1) ensuring at least an 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average test performance. As with the other West Coast chronic WET test methods, higher alpha values also satisfy the above RMDs. In these cases, the alpha value $\geq$ 0.05 that satisfies the RMDs is used.

At a 10 percent mean effect in the effluent for example, and ~50th percentile CV of 0.03, alpha levels ranging from 0.05 to 0.30 result in failure to reject the null hypothesis in none of the tests (Figure 3-24). For tests with a mean effect of 25 percent, the rate of tests declared toxic ranged from ~95 to ~70 percent, at alpha error rates ranging from 0.05 to 0.30, respectively, and approximately the 80th percentile CV value (Figure 3-25). Thus, alpha levels < 0.25 achieved the RMD that a 25 percent mean effect in the effluent is declared toxic at least 75 percent of the time regardless of within-test variability.

At ~50th percentile CV for this test endpoint (0.03) and a mean effect of 10 percent in the effluent, TST resulted in significantly fewer tests declared toxic relative to the traditional hypothesis approach at all alpha error rates (Figure 3-24). This results from the fact that the RMD is that tests with a negligible (10 percent) mean effect in the effluent are declared non-toxic most of the time when test control data are average or better.

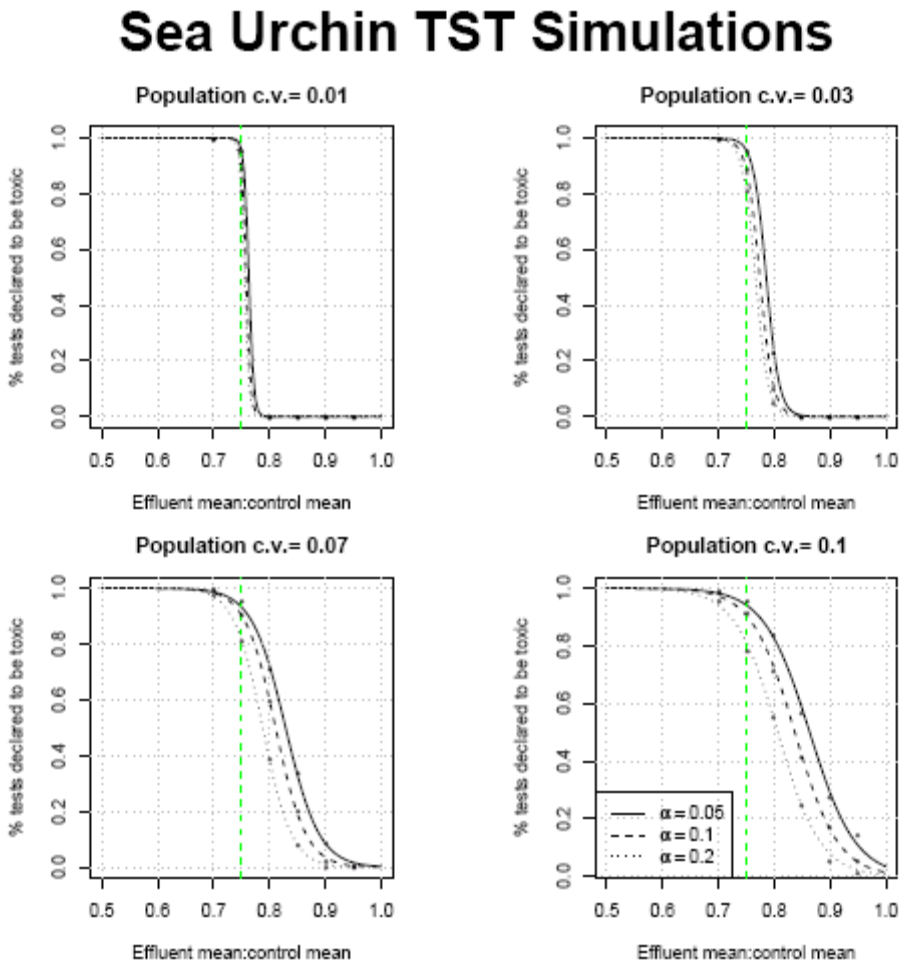

Figure 3-23. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic echinoderm fertilization WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

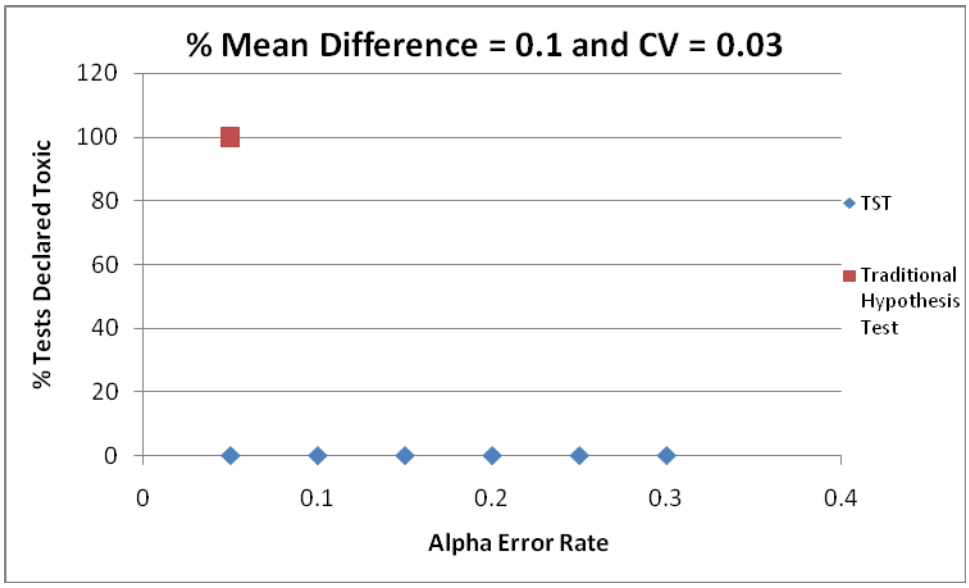

Figure 3-24. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

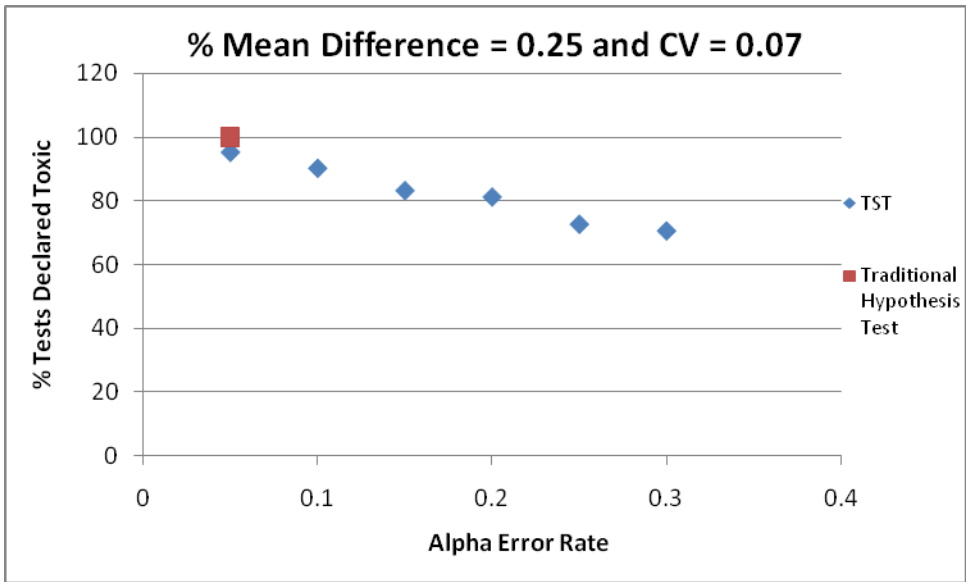

Figure 3-25. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Tests with a mean effect of 25 percent and above average precision (CV = 0.07) result in a similar rate of tests declared toxic as the traditional hypothesis approach when alpha = 0.05 (Figure 3-25). The results further support the selection of alpha = 0.05 for this WET test endpoint.

3.8 Acute *Pimephales promelas* Survival Test

As noted in the RMD discussion in Section 2.1, acute toxicity (i.e., mortality or immobility of organisms) needs to be tightly controlled because of the potential environmental implications of acute toxicity. Therefore, the RMD toxicity threshold for acute WET methods is set higher than that for the chronic WET methods, with the acute WET method *b* value = 0.80, rather than 0.75 as in the chronic methods. Consequently, the following analyses and results incorporated a *b* value of 0.80.

On the basis of actual WET data (N = 347 tests), mean control survival ranged from 0.900 to 1.000, with a median mean value of 1.000 (Table 3-11). Control CVs ranged from 0.000 to 0.185 with a median value of 0.00 (Table 3-11). The very low control variability observed is expected because of the strength and repeatability of the test endpoint (survival) and the fact that test acceptability criteria for acute WET methods require no less than 90 percent survival in controls. Using that data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.20), a range of CVs corresponding to between the 75th to the 90th percentiles, and percent mean effect in reproduction between the control and effluent concentration.

Table 3-11. Summary of mean control survival and control CV derived from analyses of 347 acute *Pimephales promelas* WET tests

Percentile	Mean control survival	Control CV	Control SD
10th	0.95	0.00	0.00
25th	1.00	0.00	0.00
50th	1.00	0.00	0.00
70th	1.00	0.00	0.00
75th	1.00	0.00	0.00
85th	1.00	0.09	0.15
90th	1.00	0.12	0.18
95th	1.00	0.19	0.23

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-26), an alpha error rate of 0.10 is appropriate for use in applying the TST approach to analysis of acute *P. promelas* data because using this alpha error rate satisfies both RMDs of (1) ensuring at least a 75 percent probability of declaring a 20 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average control performance.

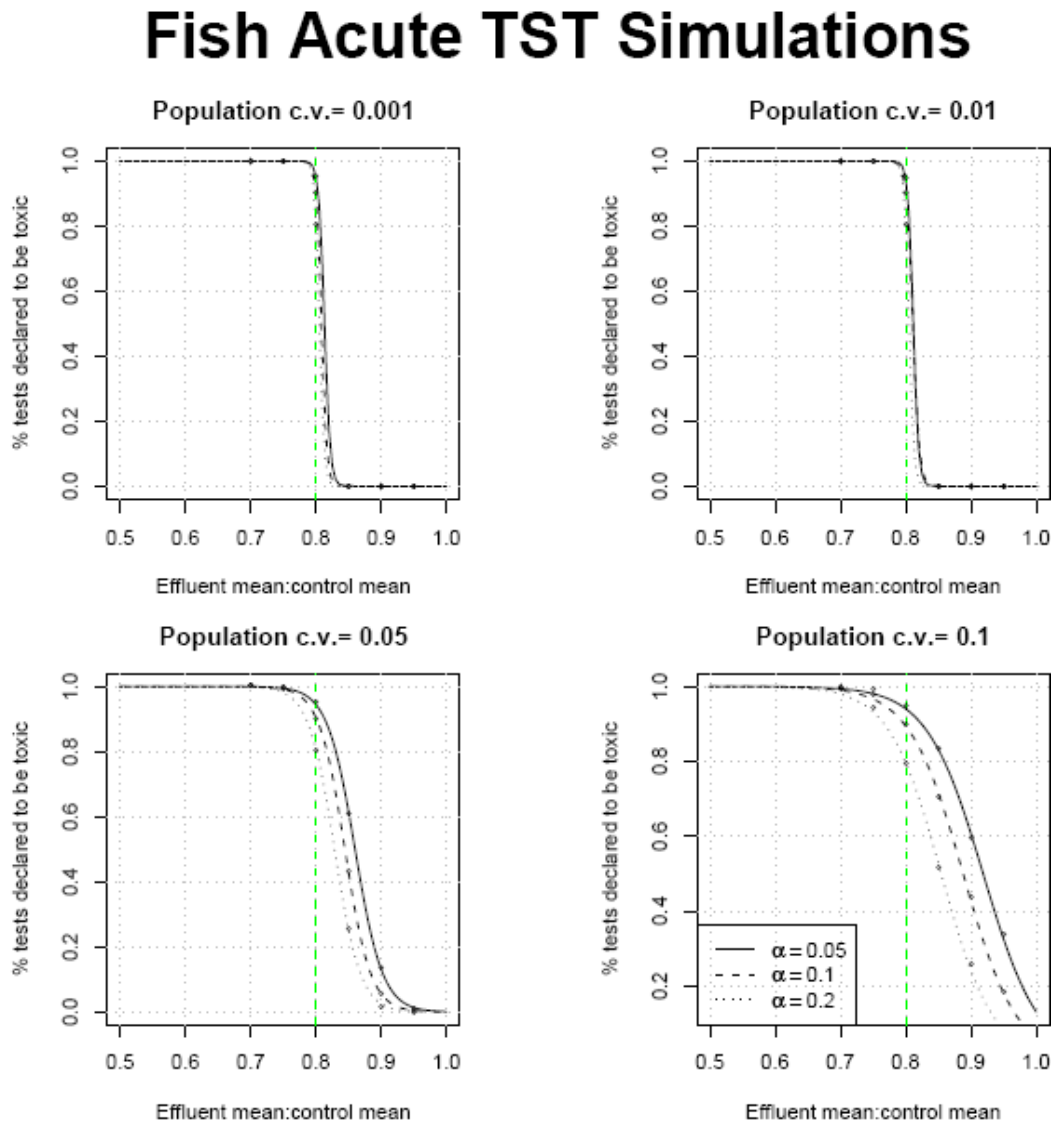

Figure 3-26. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 75th, 80th, 85th, and 88th percentiles for the acute fathead minnow WET method. The dashed line indicates the 80 percent mean effect level, which is the decision threshold for acute tests.

At a 10 percent mean effect in the effluent and a CV of 0.001 (slightly higher than the 75th percentile), alpha levels ranging from 0.05 to 0.20 resulted in failure to reject the null hypothesis in none of the tests (Figure 3-27). At the 88th percentile CV of 0.10 and a mean effect of 10 percent, alpha levels ranging from 0.05 to 0.20 resulted in declaring between 60 and 25 percent of the tests toxic, respectively. At more moderate CVs (85th percentile), an alpha of 0.10 results in 5 percent of the tests declared toxic. A lower alpha has a higher percentage of tests declared toxic.

For tests with a mean effect of 20 percent, the rate of tests declared toxic ranged from ~100 percent to ~80 percent, at alpha levels ranging from 0.05 to 0.20, respectively, and above average

CV values (Figure 3-28). The rates of tests declared toxic are consistent with the RMD that a 20 percent mean effect in the effluent is declared toxic at least 75 percent of the time. With more routine test performance, an alpha = 0.10 results in 95 percent of the tests declared toxic at a mean effect of 20 percent.

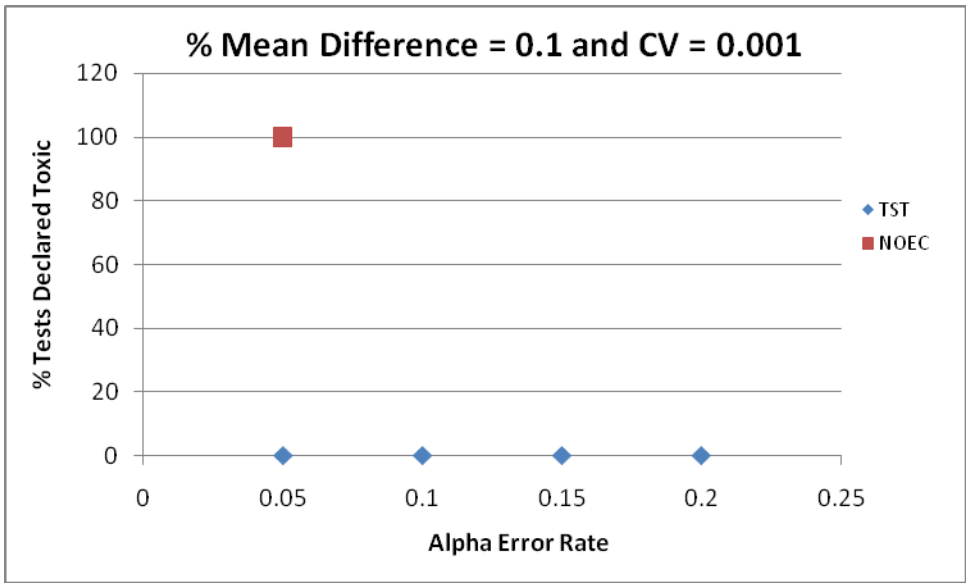

Figure 3-27. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

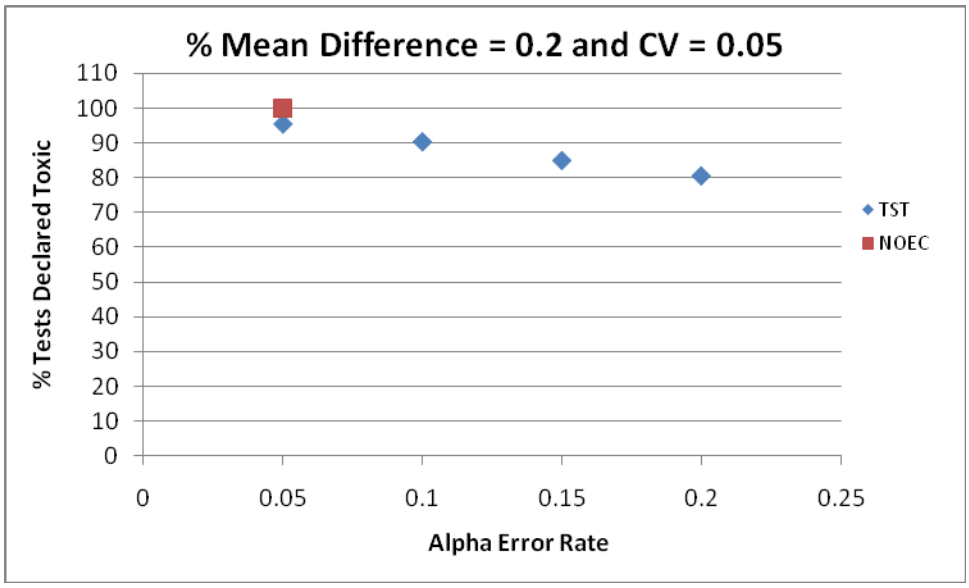

Figure 3-28. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

At a CV of 0.001 and a mean effect of 10 percent, use of the TST approach results in significantly fewer toxic tests relative to the traditional hypothesis approach at all alpha levels (Figure 3-27). These results are due to the RMD that tests with a 10 percent mean effect at the IWC are declared non-toxic most of the time.

Tests with a mean effect of 20 percent and a CV of 0.05 (85th percentile) result in a similar rate of tests declared toxic at alpha = 0.05 and 10 percent fewer tests declared toxic (90 percent of tests) at alpha = 0.10 (Figure 3-28). Because all the results noted above, an alpha = 0.10 is considered appropriately protective for this WET test method.

Effect of Increased Number of Within-Test Replicates

As expected, increasing test replication from two (the minimum allowed in the EPA WET test methods for acute fish tests) to four replicates results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach and a *lower* rate of tests declared toxic using the TST approach at a 10 percent effect using *P. promelas* acute test data. For tests with a mean effect of 10 percent and a control CV of 0.05 (corresponding to between the 75th and 90th percentile), if replicates are added to the test design, the TST approach demonstrates an increased ability to declare such a test as non-toxic (Table 3-12). As the mean effect approaches 20 percent, the percentage of tests declared toxic is less affected by increased replication using TST because a 20 percent effect in the effluent is the toxicity threshold using TST. However, the percentage of tests declared toxic continues to increase with increased replication using the traditional hypothesis approach, even when there is a negligible effect (10 percent effect as defined by TST) of the effluent. Thus, increasing test replication increases TST's ability to confirm an acceptable effluent test with mean effect less than 20 percent.

Table 3-12. Percent of fathead minnow acute tests declared toxic using TST and a *b* value = 0.8 as a function of percent mean effect, number of replicates (2 or 4 replicates), and different alpha or Type I error levels

				Alpha			
***B* value**	**CV**	**% effect**	**# reps**	**0.05**	**0.1**	**0.15**	**0.2**
0.8	0.05	0.10	2	57	33	21	13
0.8	0.05	0.20	2	95	91	85	80
0.8	0.05	0.10	4	14	5	3	1
0.8	0.05	0.20	4	95	90	85	80

3.9 Chronic *Selenastrum capricornutum* Growth Test

On the basis of actual WET data (N = 223 tests), the mean control growth ranged from 1,019,250 cells to 14,109,450 cells, with a median value of 3,331,250 cells (Table 3-13). Control CVs ranged from 0.00 to 0.20 with a median value of 0.06 (Table 3-13). Using those data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.25), CVs, and percent mean effect in growth between the control and effluent concentration. In addition, WET

test data (N = 173), in which EDTA was added to the controls, as required in the 2002 *Selenastrum* method, were evaluated independently and compared to the simulation results. For those tests the mean control growth ranged from 1,019,250 cells to 14,109,450 cells, with a median value of 3,430,000 cells (Table 3-13). Control CVs from those tests ranged from 0.00 to 0.20 with a median value of 0.06, similar to the results observed for all 223 tests (Table 3-13).

Table 3-13. Summary of mean control growth, CV and standard deviation derived from the analyses of all chronic *Selenastrum capricornutum* WET test data and compared with the analysis of only the chronic *Selenastrum capricornutum* WET test in which it was assumed that EDTA was added to the controls.

All Tests (N = 223)				Only Tests With EDTA Addition (N = 173)			
Percentile	Mean Cell Density	Control CV	Control SD	Percentile	Mean Cell Density	Control CV	Control SD
10th	1233050.0	0.02	44928.62	10th	1554500.0	0.02	43664.06
25th	2245833.5	0.04	108449.85	25th	2502500.0	0.03	135154.20
50th	3331250.0	0.06	277653.90	50th	3430000.0	0.06	309232.90
70th	4869000.0	0.10	407505.12	70th	5581650.0	0.10	417361.66
75th	6179667.0	0.11	444887.25	75th	8220000.0	0.11	447446.50
85th	9265500.0	0.13	545764.05	85th	9785000.0	0.14	543717.8
90th	9888000.0	0.16	599644.32	90th	10048000.0	0.16	583299.40
95th	10149500.0	0.18	751884.62	95th	10279000.0	0.18	669780.04

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-29), an alpha error rate of 0.25 is appropriate, for both tests with EDTA addition and tests with no EDTA addition, for use in applying the TST approach to analysis of chronic *Selenastrum* data. Using this alpha error rate addresses both RMDs of (1) ensuring at least a 75 percent probability of declaring a 25 percent mean effect as toxic and (2) ensuring that a negligible effect ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 5 percent of the time under average or better than average test performance.

For example, at a 10 percent mean effect and a low level of precision (~70th percentile for all tests, CV of 0.10), an alpha level of 0.25 resulted in failure to reject the null hypothesis in $\leq$ 5 percent of tests with or without EDTA addition (Figure 3-29). For all tests with a mean effect of 25 percent, and a similar precision, the rate of tests declared toxic is 75 percent at an alpha value of 0.25, consistent with RMDs (Figure 3-29).

At ~70th percentile CV (0.10) and a mean effect of 10 percent, for both tests with and without EDTA addition, use of the TST approach results in fewer toxic tests relative to the traditional hypothesis testing approach at all alpha error rates, including the alpha error rate of 0.25 which declared less than 5 percent of the tests toxic (Figure 3-30).

Tests with a mean effect of 25 percent, regardless of precision (CV = 0.10 or 0.15), result in a 75 percent or greater rate of tests declared toxic, which is significantly more than that using the traditional hypothesis testing approach using any alpha value between 0.05 and 0.25 (Figure 3-31). The percent of tests found to be toxic using the TST approach with a mean effect of 25 percent was not significantly affected by the change in CV values.

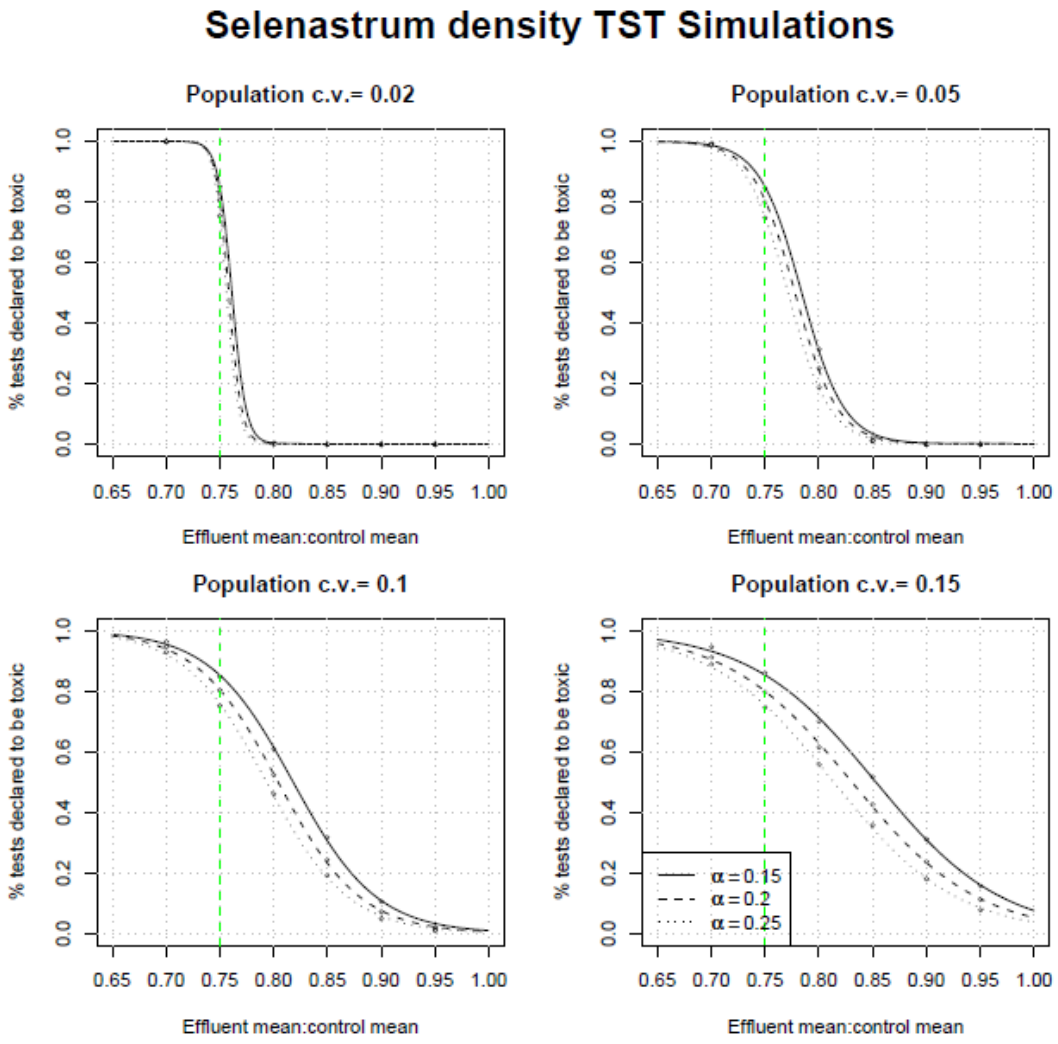

Figure 3-29. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 10th, 40th, 70th, and 85th percentiles for the chronic *Selenastrum* WET method. The dashed line indicates the 75 percent mean effect level, which is the decision threshold for chronic tests.

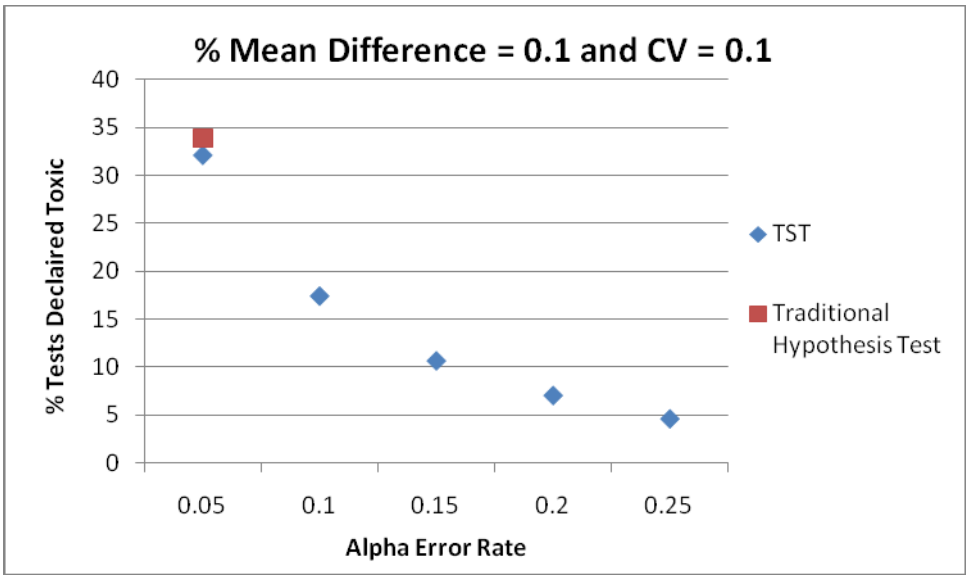

Figure 3-30. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

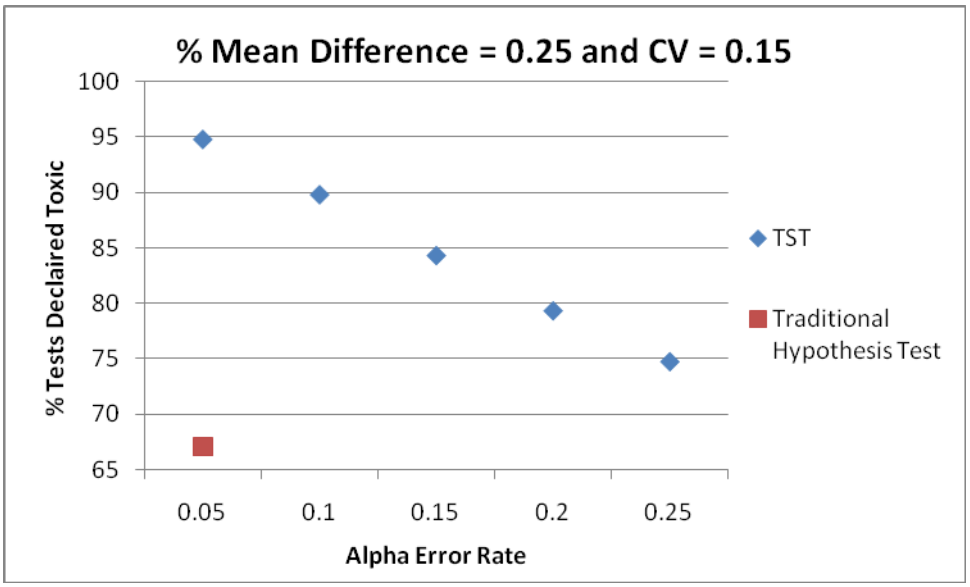

Figure 3-31. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Effluent Data Results

Results from actual effluent tests were compared between TST and the traditional hypothesis testing approach for all control CV's (Table 3-14). At a mean effect of 10–15 percent (N = 25),

TST declared none of the tests toxic while the traditional hypothesis testing approach declared 67 percent of the tests toxic. However, if the mean effect is greater than 25 percent (N = 97), TST declared 100 percent of the tests toxic, while the traditional hypothesis testing approach declared 98 percent of the tests toxic. These results indicate that TST is as protective as the current hypothesis testing approach for those tests when the TST RMD threshold for toxicity is exceeded.

Table 3-14. Comparison of the percentage of chronic *Selenastrum* tests declared toxic using TST versus the traditional hypothesis testing approach

% Mean effect	N	% tests toxic using TST	% tests toxic using traditional hypothesis testing approach
10–15	25	0	67
> 25	97	100	98

3.10 Acute *Ceriodaphnia dubia* Survival Test

Acute toxicity (i.e., mortality or immobility of organisms) needs to be tightly controlled because of the potential environmental implications of acute toxicity. Therefore, the RMD toxicity threshold for acute WET methods is set higher than that for the chronic WET methods, with the acute WET method *b* value = 0.80, rather than 0.75 as in the chronic methods. Consequently, the following analyses and results incorporated a *b* value of 0.80.

On the basis of actual WET data (N = 239 tests), mean control survival ranged from 0.900 to 1.000, with a median mean value of 1.000 (Table 3-15). Control CVs ranged from 0.00 to 0.22 (the minimum and maximum levels obtainable using the test acceptability criteria) with a median value of 0.00 (Table 3-15). The very low control variability observed is expected because of the strength and repeatability of the test endpoint (survival) and the fact that test acceptability criteria for acute WET methods stipulate no less than 90 percent survival in the controls. Using that data, simulation analyses were conducted to evaluate the percentages of tests declared toxic (i.e., failure to reject the null hypothesis) by TST at various alpha error rates (between 0.05 and 0.30), a range of CVs, and percent mean effect in survival between the control and effluent concentration.

Table 3-15. Summary of mean control growth, CV and standard deviation derived from analyses of 239 acute *Ceriodaphnia dubia* WET tests.

Percentile	Mean Survival (%)	Control CV	Control SD
10th	0.95	0.00	0.00
25th	1.00	0.00	0.00
50th	1.00	0.00	0.00
70th	1.00	0.00	0.00
75th	1.00	0.00	0.00
85th	1.00	0.00	0.00
90th	1.00	0.11	0.10
95th	1.00	0.11	0.10

Identifying Test Method-Specific α

On the basis of all simulation results (Figure 3-32), an alpha error rate of 0.10 is appropriate for use in applying the TST approach to analysis of acute *Ceriodaphnia dubia* data because using this alpha error rate best satisfies both RMDs of (1) ensuring at least a 75 percent probability of declaring a 20 percent mean effect as toxic and (2) ensuring that a negligible effect (≤10 percent mean effect) is declared toxic ≤ 5 percent of the time under average control performance.

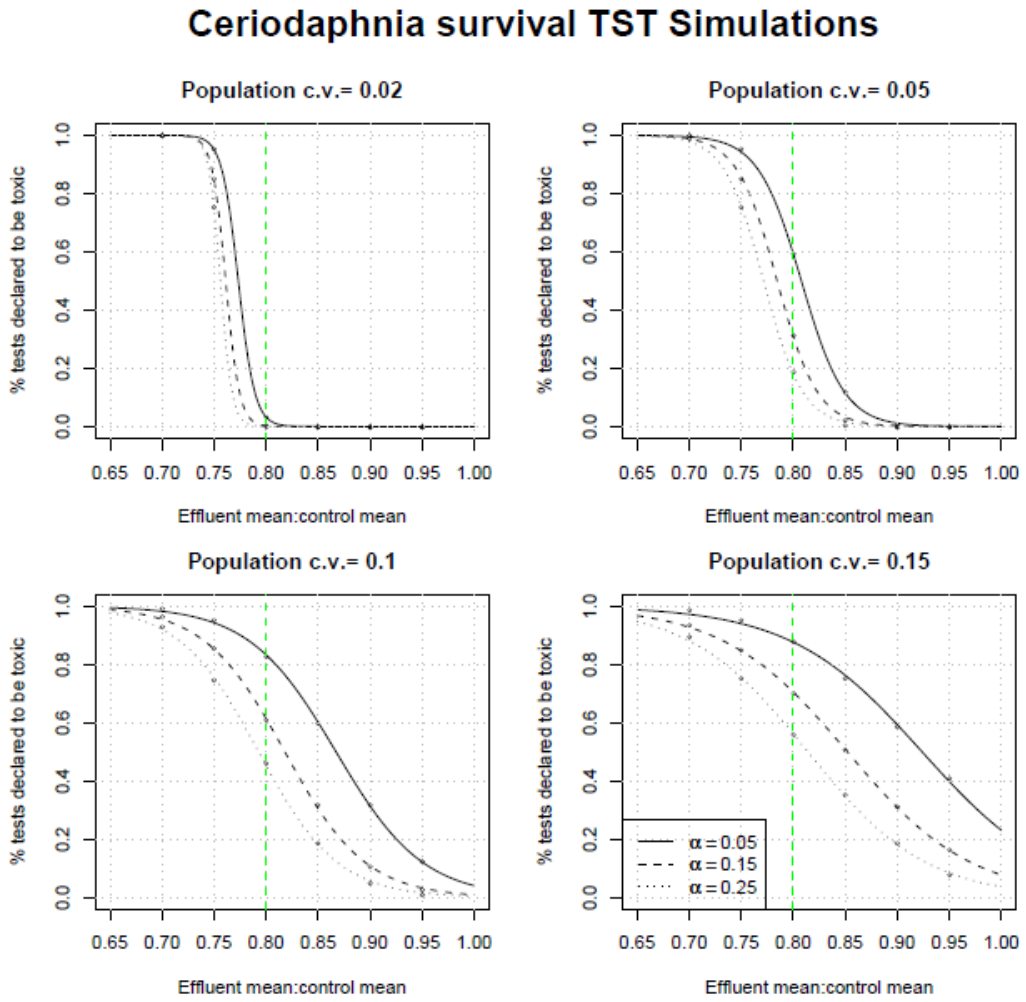

Figure 3-32. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. The first two CVs correspond to the 85th percentile, and the following two correspond to the 95th and ~98th, respectively for the acute *Ceriodaphnia dubia* WET method. The dashed line indicates the 80 percent mean effect level, which is the decision threshold for acute tests.

For example, at a 10 percent mean effect in the effluent and a CV of 0.02 (slightly higher than the 85th percentile), alpha levels ranging from 0.05 to 0.25 resulted in failure to reject the null hypothesis in ≤ 5 percent of the tests (Figure 3-32). However, at the 90th and 95th percentile CVs of 0.10 and a mean effect of 10 percent, the alpha level of 0.25 resulted in 19 percent of the tests

found toxic. For tests with a mean effect of 20 percent, and ~85th percentile precision (CV of 0.02), 75 percent of the tests are declared toxic, achieving the RMD using an alpha value of 0.25 (Figure 3-32).

For tests with a mean effect of 20 percent, the rate of tests declared toxic ranged from ~95 percent to ~75 percent, at alpha levels ranging from 0.05 to 0.25, respectively, using all CV values that correspond to $\leq$ 95th percentile. (Figure 3-32). The rates of tests declared toxic are consistent with the RMD that a 20 percent mean effect in the effluent is declared toxic at least 75 percent of the time. With more routine test performance, an alpha of 0.10 results in 90 percent of the tests declared toxic at a mean effect of 20 percent.

At a CV of 0.02 (~85th percentile) and a mean effect of 10 percent, use of the TST approach results in no toxic tests, while the traditional hypothesis approach results in 100 percent toxic tests at all alpha levels (Figure 3-33).

Tests with a mean effect of 20 percent and a range of within-test control precision values (CV of 0.02 to 0.15) result in at least 75 percent of the tests declared toxic using an alpha = 0.10 (Figure 3-34). In contrast fewer tests are declared toxic at a 20% effect when using the traditional hypothesis testing approach and any alpha value between 0.05 and 0.25 (Figure 3-34). Thus, the percent of tests found to be toxic using the TST approach with a mean effect of 20 percent was not significantly affected by the change in CV values.

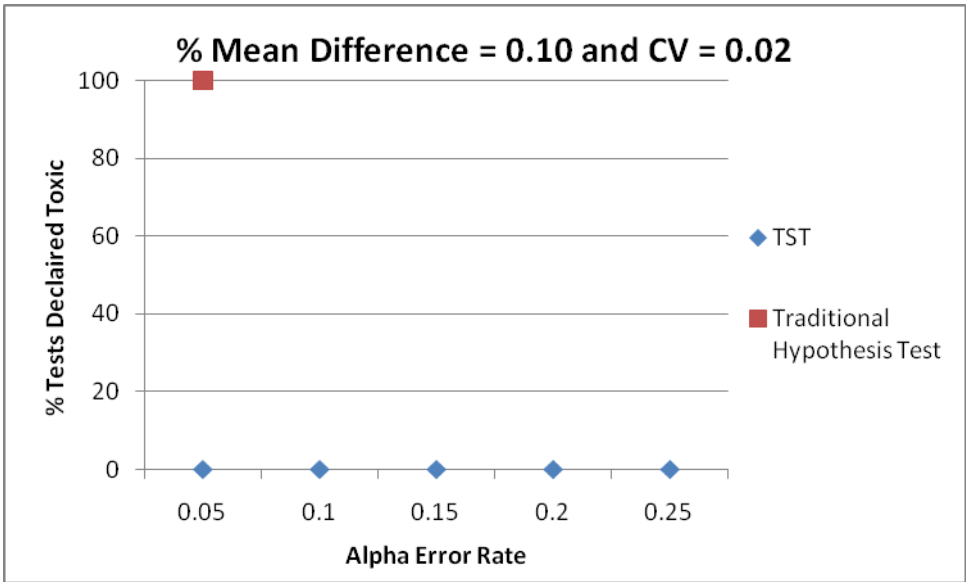

Figure 3-33. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

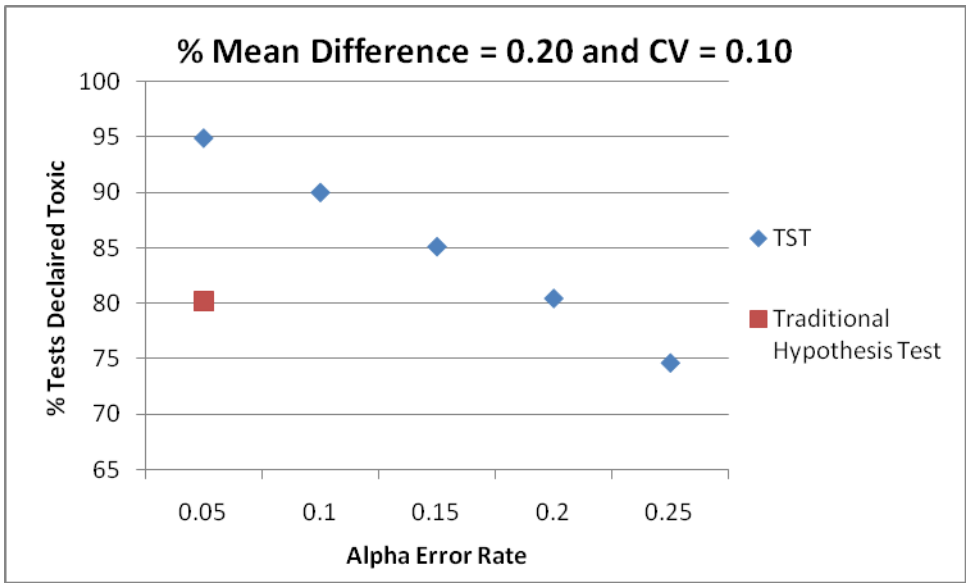

Figure 3-34. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

Effect of Increased Number of Within-Test Replicates

As with the fathead minnow acute method, increasing test replication from four (the minimum allowed in the EPA WET test methods for acute *Ceriodaphnia dubia* tests) to six replicates results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach and a *lower* rate of tests declared toxic using the TST approach at a 10 percent mean effect using *C. dubia* acute test data. For tests with a mean effect of 10 percent and a control CV of 0.06 (corresponding to between the 85th and 90th percentile), if replicates are added to the test design, the TST approach demonstrates an increased ability to declare such a test as non-toxic (Table 3-16). As the mean effect approaches 20 percent, the percentage of tests declared toxic is less affected by increased replication using TST because a 20 percent effect in the effluent is the RMD using TST. However, the percentage of tests declared toxic continues to increase with increased replication using the traditional hypothesis approach, even when there is a negligible effect (10 percent effect as defined by TST) of the effluent. Thus, increasing test replication increases TST's ability to confirm an acceptable effluent test with mean effect less than 20 percent.

Table 3-16. Percent of *Ceriodaphnia dubia* acute tests declared toxic using TST and a *b* value = 0.8 as a function of percent mean effect, number of replicates (4 or 6 replicates), and different alpha or Type I error levels

B value	CV	% effect	# reps	Alpha			
				0.05	0.1	0.15	0.2
0.8	0.06	0.10	4	23	12	7	5
0.8	0.06	0.20	4	95	90	85	80
0.8	0.06	0.10	6	8	4	3	2
0.8	0.06	0.20	6	95	90	85	80

4.0 SUMMARY OF RESULTS AND IMPLEMENTING TST

4.1 Summary of Test Method-Specific Alpha Values

On the basis of all the analyses conducted in this project, the test method-specific alpha levels summarized in Table 4-1 are used with the TST approach. The method-specific alpha values apply to all test endpoints for a given EPA WET test method (e.g., survival and reproduction for the *Ceriodaphnia* chronic WET test method). As noted in Section 2.3.1, alpha values were selected on the basis of simulation analyses using normally distributed data and equal variances in the control and the effluent. While additional analyses indicate that the alpha levels identified are robust to the type of heterogeneous variances and non-normal data observed in WET test data (see Appendix A), this issue is still acknowledged as a potential uncertainty.

The alpha values identified above provide as much protection under most circumstances as the current approved WET test analysis methods when the mean effect at the IWC exceeds the toxicity threshold of the TST approach.

At the chronic toxicity regulatory management threshold of 25 percent mean effect of the effluent and lower within-test control CVs (< 50th percentile), TST declares a greater percentage of tests non-toxic than the traditional hypothesis approach for some of the chronic WET test methods examined (e.g., fathead minnow chronic WET test) because of the higher alpha levels assigned to those test methods. At either higher within-test CVs or higher mean effect levels, results are more similar between the two approaches, as explained in Section 1.4 of this document. With more extreme within-test variability (≥ 80th percentile CV), results tend to be reversed with TST declaring a higher percentage of tests toxic at 25 percent mean effect of the effluent as compared to the traditional hypothesis approach; e.g., for the *Ceriodaphnia* reproduction endpoint, at the 80th percentile CV, TST declares ~20 percent of the tests non-toxic at a 25 percent mean effect, while the traditional approach declares 24 percent of the tests non-toxic. If test data are non-normal (a somewhat frequent condition for some WET endpoints such as acute and chronic survival, or when a high level of toxicity is observed in certain effluent concentrations within a test), additional research has indicated that use of Welch's t-test results in a lower rejection rate (i.e., is more conservative) using the TST approach, resulting in a higher percentage of tests declared toxic when the effluent effect $\geq b \times$ control mean (Appendix A). For the acute fathead minnow test method, at the acute toxicity regulatory management threshold of 20 percent mean effect of the effluent, both approaches had a similarly low percentage of tests declared non-toxic over all within-test CVs. Results of this comparison also demonstrate that for all WET test methods, the TST approach declares a lower percentage of tests as toxic at a 10 percent mean effect in the effluent, for most WET tests (i.e., within-test CV ≤ 75th percentile for a given WET test method). If within-test variability is lower (control data has greater precision), the result is further accentuated; i.e., an even greater percentage of tests are declared toxic at a 10 percent effect using the traditional hypothesis approach and an even lower percentage of tests declared toxic using TST.

Table 4-1. Summary of alpha (α) levels or false negative rates recommended for different EPA WET test methods using the TST.

EPA WET test method	*b* value	Probability of declaring a toxic effluent non-toxic: False negative (α) error[a]
Chronic Freshwater and East Coast Methods		
Ceriodaphnia dubia (water flea) survival and reproduction	0.75	0.20
Pimephales promelas (fathead minnow) survival and growth	0.75	0.25
Selenastrum capricornutum (green algae) growth	0.75	0.25
Americamysis bahia (mysid shrimp) survival and growth	0.75	0.15
Arbacia punctulata (Echinoderm) fertilization	0.75	0.05
Cyprinodon variegatus (Sheepshead minnow) and *Menidia beryllina* (inland silverside) survival and growth	0.75	0.25
Chronic West Coast Marine Methods		
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	0.75	0.05
Atherinops affinis (topsmelt) survival and growth	0.75	0.25
Haliotis rufescens (red abalone), *Crassostrea gigas (oyster), Dendraster excentricus, Strongylocentrotus purpuratus* (Echinoderm) and *Mytilus sp* (mussel) larval development methods	0.75	0.05
Macrocystis pyrifera (giant kelp) germination and germ-tube length	0.75	0.05
Acute Methods		
Pimephales promelas (fathead minnow), *Cyprinodon variegatus* (Sheepshead minnow), *Atherinops affinis* (topsmelt), *Menidia beryllina* (inland silverside) acute survival[b]	0.80	0.10
Ceriodaphnia dubia, *Daphnia magna, Daphnia pulex*, *Americamysis bahia* acute survival[b]	0.80	0.10

Notes:
a α levels shown are the probability of declaring an effluent toxic when the mean effluent effect = 25% for chronic tests or 20% for acute tests and the false positive rate (β) is $\leq$ 0.05 (5%) when mean effluent effect = 10%.
b. Based on a four replicate test design

4.2 Calculating Statistics for Valid WET Data Using the TST Approach

Appendix B includes a step-by-step guide for using the TST approach to analyze valid WET data. The appendix also includes a statistical flowchart. Note that the WET test method should follow the test condition requirements as specified in EPA's approved WET methods (USEPA 1995, 2002a, 2002b, 2002c).

The TST approach is used to statistically compare organism responses from two treatments of the WET test, the IWC and the control. Percent data (quantal data), such as percent survival or percent germination from a WET test, is first transformed as recommended in the EPA WET test manuals. Other types of WET data (e.g., growth or reproduction data) are not transformed (for

the rationale, see Appendix A). Data are then analyzed using Welch's t-test, a well-known modification of the traditional t-test (Zar 1996), which is appropriate for the TST approach (see Appendix A).

Appendix C lists the critical t values that apply to WET testing using the TST approach given the number of degrees of freedom and the α level that applies for a given WET test method from Table 4-1 of this document. If the calculated t value for the WET test is greater than the critical t value (given in Appendix C), the null hypothesis is rejected, i.e., the test result is a *pass* and **the effluent is declared *non-toxic***. If the calculated t value is less than the critical t value in Appendix C, the null hypothesis is not rejected, i.e., the test result is a *fail* and **the effluent *is* declared toxic**.

4.3 Benefits of Increased Replication Using TST

One of the intended benefits of the TST approach is that increasing the precision and power of the test increases the chances of rejecting the null hypothesis and declaring a truly acceptable sample as non-toxic. This increases the permittee's ability to demonstrate that a sample is acceptable. Results for the *Ceriodaphnia*, fathead minnow, and mysid chronic test methods presented in Section 3 indicate the benefits of increased replication within a test, especially when the mean effect of the sample is below about 20 percent in the case of chronic tests and about 15 percent for acute tests. As expected, increasing test replication (and thereby the power of the test) results in a *higher* rate of tests declared toxic using the traditional hypothesis testing approach but a *lower* rate of tests declared toxic using the TST approach.

Conducting tests with more replicates can help a permittee demonstrate that the effluent is acceptable if the mean effect at the IWC is truly less than the RMDs as defined by TST (25 percent effect for chronic and 20 percent for acute). Conversely, increasing replicates does not assist a permittee using the traditional hypothesis testing approach.

4.4 Applying TST to Ambient Toxicity Programs

In ambient and stormwater toxicity testing, a laboratory control and a single concentration (i.e., 100 percent ambient water or stormwater) are often tested. In those two-concentration WET tests, the objective is to determine if a sample or site water is toxic, as indicated by a significantly worse organism response compared to the control. In this WET testing design, the determination of pass or fail (i.e., toxic or non-toxic) is ascertained using a traditional t-test (USEPA 2002c). EPA WET test methods recommend that the statistical significance (i.e., pass/fail) of a two-sample test design for ambient and stormwater toxicity testing be determined by using only a modified t-test (if homogeneity of variance is not achieved) or a traditional t-test (if homogeneity of variance is achieved).

To demonstrate the value of the TST approach in ambient toxicity programs, ambient toxicity test data from California's SWAMP was used for 409 chronic tests for *Ceriodaphnia dubia* and 256 chronic tests for *Pimephales promelas* using EPA's 2002 WET test methods (USEPA 2002a). WET test data for each WET test method were subjected to the same statistical analyses as described in Section 2 of this document.

Chronic *Ceriodaphnia dubia* Ambient Toxicity Tests

Table 4-2 summarizes results from the 409 *Ceriodaphnia dubia* ambient toxicity tests analyzed and a $\alpha = 0.20$ for this test method. Although the majority of the tests examined resulted in the same decision using either the TST or the traditional t-test approach, approximately 6 percent of the tests (24 tests) would have been declared non-toxic using the traditional t-test approach with mean effect levels > 25 percent. In addition, 2 percent of the tests (7 tests) would have been declared toxic at mean effect levels < 15 percent and as low as 7 percent.

Table 4-2. Comparison of results of chronic *Ceriodaphnia* ambient toxicity tests using the TST approach and the traditional t-test analysis. $\alpha = 0.2$ and *b* value = 0.75 for the TST approach. $\alpha = 0.05$ for the traditional hypothesis testing approach

Both approaches declare toxic	Only TST declares toxic	Only traditional approach declares toxic	Both approaches declare non-toxic
19.8%	5.9%	1.7%	72.6%

Figure 4-1 shows ranges of CV values observed in *Ceriodaphnia dubia* ambient toxicity tests for those samples declared toxic using either the TST approach or the traditional t-test but not both approaches. As expected, within-test variability was relatively high (higher CVs) for those tests found non-toxic using a t-test but toxic using the TST approach. The results again demonstrate a limitation of the traditional hypothesis testing approach when control variability is relatively high. Under those conditions, the t-test did not have the power to detect toxicity when it was present. Figure 4-1 also demonstrates that the TST approach is superior to the traditional t-test when within-test variability is relatively low and the mean percent effect is well below the risk management level of 25 percent. Under such conditions, the traditional t-test declared some samples toxic using this WET test method, even when the mean effect was as little as 7 percent. The TST approach, however, declared all such samples non-toxic using the recommended $\alpha = 0.20$. Thus, the TST approach reduces the number of tests classified as toxic when effects are actually well below risk management levels of concern.

Similar to the *Ceriodaphnia* ambient test data, within-test variability was higher in those chronic fathead minnow ambient tests found non-toxic using a t-test but toxic using the TST approach (Figure 4-2). Similarly, those tests declared non-toxic by the TST approach but toxic using t-test had lower within-test variability and mean effect levels < 25 percent (Figure 4-2). Thus, as with the chronic *Ceriodaphnia* ambient tests, data from chronic fathead minnow ambient tests demonstrate that the TST approach provides better protection than the traditional t-test approach while also identifying those samples that are truly acceptable from a regulatory management perspective.

4.5 Implementing TST in WET Permitting under NPDES

The TST approach is an alternative statistical approach for analyzing and interpreting valid WET data; it is not an alternative approach to developing NPDES permit WET limitations. Using the TST approach does not result in any changes to EPA's WET test methods.

Chronic *Ceriodaphnia* ambient WET tests that are identified as non-toxic (pass) using the traditional hypothesis approach (t-test) generally have poor test sensitivity (high control CVs), masking effects, as compared to using the TST approach.

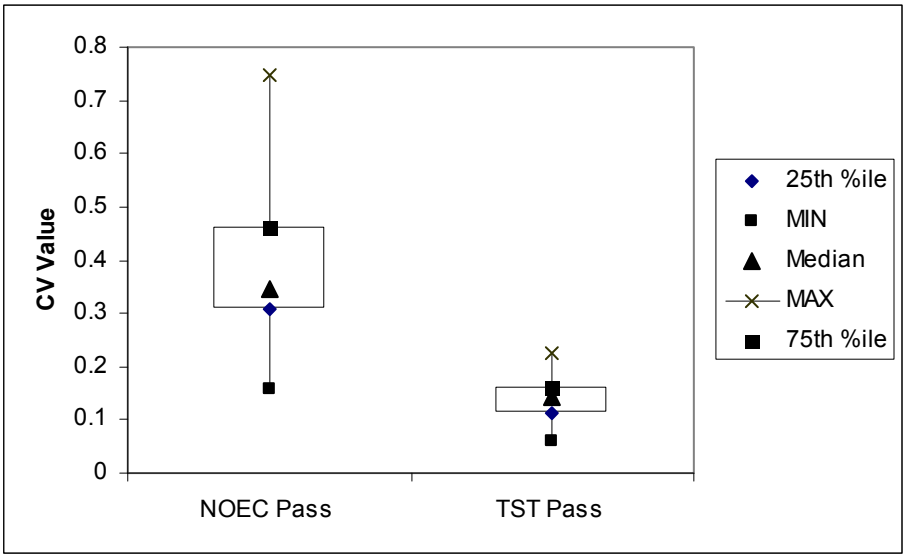

Figure 4-1. Range of CV values observed in chronic *C. dubia* ambient toxicity tests for samples that were found to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*). California's SWAMP WET test data.

Fish ambient WET tests that are identified as non-toxic using the traditional hypothesis approach (t-test) generally have poor test sensitivity (high control CVs), masking effects, as compared to using the TST approach.

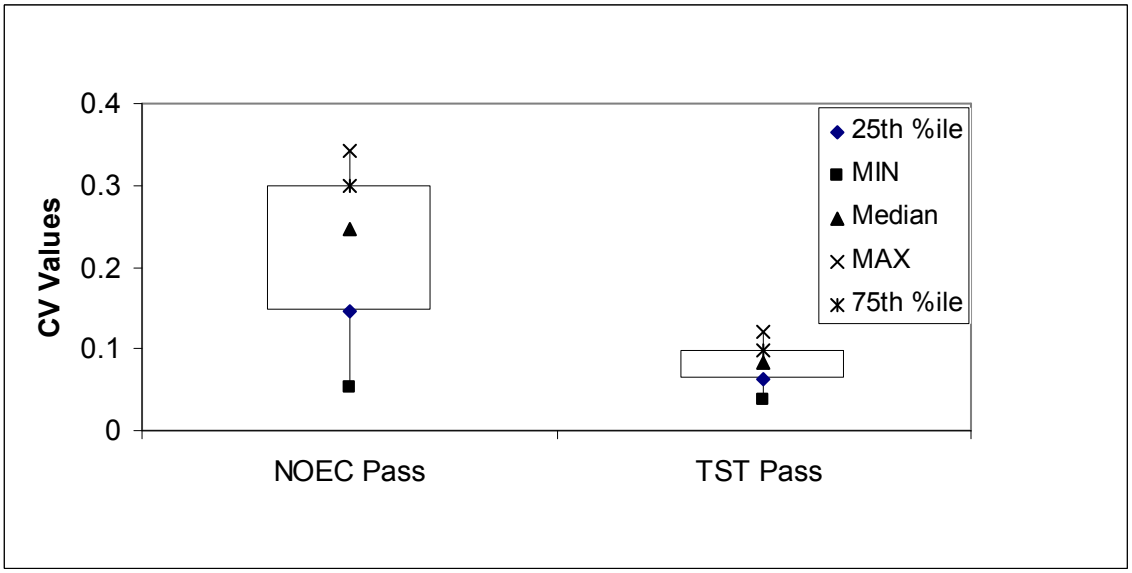

Figure 4-2. Range of CV values observed in chronic *P. promelas* ambient toxicity tests for samples that were declared to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*). California's SWAMP WET test data.

4.6 Reasonable Potential (RP) WET Analysis

NPDES permitting authorities conducting an RP analysis must follow 40 CFR 122.44(d)(1) to determine whether a discharge will, "cause, have the reasonable potential to cause, or contribute to" an excursion of a numeric criterion or a narrative WET criterion. Some states have state-specific WET RP approaches in their water quality control plan or other NPDES policy or guidance.

For RP calculations using the TST approach, EPA recommends that permitting authorities use all valid WET test data generated during the current permit term and any additional valid data that are submitted as part of the permit renewal application. The TST RP approach necessitates having at least a minimum of four valid WET tests to address effluent representativeness (see EPA's TSD, Chapter 3, pg. 57, under Step 2 in the section *Steps in Whole Effluent Characterization Process*). EPA also recommends that states request that their permittees provide the actual test endpoint responses for the control (i.e., control mean) and IWC concentration (i.e., IWC mean) for each WET test conducted to make it easier for permit writers to find the necessary WET test results when determining WET RP. WET test data are then analyzed according to the TST approach using the IWC and control test concentrations for all the valid WET test data available. If fewer than four valid WET test data points are available, permitting authorities should follow EPA's TSD RP approach because it addresses small WET data sets by incorporating an RP multiplying factor (see section 3.3.2 of the TSD, pg. 54) to account for effluent variability in small WET data sets. If sufficient, valid WET test data are available and the TST statistical approach indicates that the IWC is toxic in any WET test, RP has been demonstrated (40 CFR 122.44(d)(1)(i)). To address concerns regarding the "potential to cause or contribute to toxicity," an analysis of the mean effect at the IWC is also conducted to determine whether the effluent has RP, even if all test results are declared a *pass* using the TST approach (for more details, see EPA's *TST Implementation Document* EPA 833-R-10-003).

Note that using the TST approach might be to the permittee's advantage. If the permittee decides to incorporate additional test replicates for the control and the IWC when conducting the WET test, above the minimum required in the EPA WET test methods, the test power is increased. More test replicates increases test power, which means a lower probability of a false positive using the TST approach *if the effluent is truly non-toxic based on the RMDs in the TST approach*. Thus, using the TST approach, a permittee has a greater ability to *prove the negative* (i.e., its effluent does not have RP).

In those cases where the WET RP outcome is *yes*, a WET limit is expressed in the permit. In situations where the RP outcome is *no*, WET monitoring requirements should still be incorporated in the permit. A *fail* test result during monitoring could trigger additional steps if described in the permit. In either of those situations, if toxicity is demonstrated, states should specify an approach to address toxicity in the permit. This often includes initially accelerated toxicity tests (i.e., increased frequency of testing) and permit requirements to perform a toxicity reduction evaluation.

4.7 NPDES WET Permit Limits

Using the TST approach, WET NPDES permit limits would be expressed as *no significant toxicity of the effluent at the IWC using the TST analysis approach*. A test result of *Pass* is when

the calculated *t* value is greater than *the critical t value*. A test result of *Fail* is when the calculated *t* value is less than *the critical t value*.

Beyond assessing WET data for the NPDES Program, WET tests are used to assess toxicity of receiving water (watershed assessment for CWA section 303(d) determinations) and stormwater samples. Often as a first assessment of receiving or stormwater toxicity, researchers test a control and a single concentration (e.g., 100 percent receiving water or stormwater). In such cases, the TST approach can be used in the same way a t-test is used. Such analysis is used to determine whether organism response in a specified ambient concentration is significantly different than the control organism response.

5.0 CONCLUSIONS

Results of this project indicate that the TST is a viable additional option for analyzing valid acute and chronic WET test data. Given the RMDs and test-method specific alpha values specified in the TST approach, TST provides a transparent methodology for demonstrating whether an effluent truly is acceptable under the NPDES WET Program. The advantage of the TST approach is that it provides a structure in which it is easier to express, understand, and implement regulatory management goals. The alpha values identified in this project build on existing statistical information (such as data sources and analysis examining ability to detect toxic effects) on WET previously published by EPA, including *Understanding and Accounting for Method Variability in WET Applications Under the NPDES Program* (USEPA 2000).

More than 2,000 valid WET test results and thousands of simulations were conducted to develop the technical basis for the TST approach. This approach builds on the strengths of the traditional hypothesis testing approach, including using robust statistical analyses to determine whether an effluent is toxic (i.e., Welch's t-test), as well as published EPA documents regarding WET analysis and interpretation and the statistical literature. The TST approach yields a rigorous statistical interpretation of valid WET data by incorporating the transparent RMDs, established alpha and beta error rates, and thereby test power. Because this approach incorporates statistical test power, using TST will result in greater confidence in WET regulatory decisions. Additional benefits of using TST in WET analysis include the following:

- It provides a positive incentive for the permittee to generate high quality WET data to the permitting authority.
- It provides the ability to analyze a two-concentration test design (e.g., IWC versus control; stormwater and watershed assessments) using a streamlined statistical analysis flowchart. It is applicable to both NPDES WET permitting and section 303(d) watershed assessment programs.

In summary, the TST approach provides another option for permitting authorities and permittees to use in analyzing valid WET test data. The TST provides a positive incentive to generate high quality WET data to make informed decisions regarding NPDES WET RP and permit compliance determinations. By using TST, permitting authorities will be better able to identify toxic or non-toxic samples.

6.0 LITERATURE CITED

Anderson, S. and W. Hauck. 1983. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics—Theory and Methods* 12:2663–2692.

Aras, G. 2001. Superiority, non-inferiority, equivalence, and bioequivalence–revisited. *Drug Information Journal* 35:1157–1164.

Berger, R., and J. Hsu. 1996. Bioequivalence trials, intersection—union tests and equivalence confidence sets. *Statistical Science* 11:283–319.

Denton, D., and T. Norberg-King. 1996. Whole Effluent Toxicity Statistics: A regulatory perspective. In D. Groethe, K. Dickson, and D. Reed (eds). *Whole Effluent Toxicity Testing: An Evaluation of Methods and Predictability of Receiving System Responses*. SETAC Publications, Pensacola, FL, pp. 83–102.

Denton, D., J. Fox, and F. Fulk. 2003. Enhancing toxicity test performance by using a statistical criterion. *Environmental Toxicology and Chemistry* 22:2323–2328.

Erickson, W., and L. McDonald. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. *Environmental Toxicology and Chemistry* 14:1247–1256.

Erickson, W. 1992. Hypothesis testing under the assumption that a treatment does harm to the environment. Master Thesis, University of Wyoming, Laramie, WY.

Grothe, D., K. Dickson, and D. Reed. 1996. *Whole Effluent Toxicity Testing: An Evaluation of Methods and Predictability of Receiving System Responses*. SETAC Publications, Pensacola, FL.

Hatch, J. 1996. Using statistical equivalence testing in clinical biofeedback research. *Biofeedback and Self-Regulation* 21:105–119.

Shukla, R., Q. Wang, F. Fulk, C. Deng, and D. Denton. 2000. Bioequivalence approach for whole effluent toxicity testing. *Environmental Toxicology and Chemistry* 19:169–174.

Streiner, D. 2003. Unicorns *Do* Exist: A Tutorial on *Proving* the Null Hypothesis. *Canadian Journal of Psychiatry* 48(11):756–761.

Stunkard, C. 1990. *Tests of Proportional Means for Mesocosms Studies*. Technical Report. Department of Measurement, Statistics, and Evaluation. University Maryland, College Park, MD.

USEPA (U.S. Environmental Protection Agency). 1988. *Methods for Evaluating the Attainment of Cleanup Standards*. Volume 1: Soils and solid media. U.S. Environmental Protection Agency, Statistical Policy Branch (PM-223), Office of Policy, Planning and Evaluation, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 1989. *Guidance Document for Conducting Terrestrial Field Studies*. U.S. Environmental Protection Agency, Ecological Effects Branch, Hazard Evaluation Division, Office of Pesticides Programs, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 1991. *Technical Support Document for Water Quality-based Toxics Control*. EPA/505/2-90-001. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (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*. Eds: G. Chapman, D. Denton, and J. Lazorchak. EPA/600/R-95-136. U.S. Environmental Protection Agency, National Exposure Research Laboratory, Cincinnati, OH, and Office of Research and Development, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2000. *Understanding and Accounting for Method Variability in Whole Effluent Toxicity Applications Under the NPDES Program*. EPA/833-R-00-003. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2002a. *Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms*. EPA/821/R-02-013. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2002b. *Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms*. 3rd ed. EPA/821/R-02-14. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2002c. *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*. 5th ed. EPA/821/R-02-012. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

Zar, J.H. 1996. *Biostatistical Analysis*. 3rd ed. Prentice Hall Publishers, Princeton, NJ.

APPENDIX A

RATIONALE FOR USING WELCH'S T-TEST IN TST ANALYSIS OF WET DATA FOR TWO-SAMPLE COMPARISONS

APPENDIX A

RATIONALE FOR USING WELCH'S T-TEST IN TST ANALYSIS OF WET DATA FOR TWO-SAMPLE COMPARISONS

This appendix demonstrates that the Welch modification of the t-test is suitable for WET test data and applicable to the TST approach. It also provides the evaluation and justification for certain WET test data that do not strictly adhere to the assumptions of the Welch t-test.

The Welch t-test accounts for different variances in two groups and assumes data are normally distributed (Welch 1938, 1947; Moser et al. 1989; Coombs et al. 1996; Zar 1996). For non-normal data that have skewed, long-tailed distributions, the Welch's t-test is known to have poor coverage (Zimmerman 2006). (By poor coverage, EPA means that the realized error rate, alpha, under the null hypothesis, is greater than the intended, nominal value of alpha). It is demonstrated below that WET data to which the TST will be applied typically have moderately unequal variances in the control and the IWC. That fact motivates use of the Welch t-test rather than the t-test (which assumes equal variances). It is also demonstrated that WET test data are typically non-normal but in a way that does not substantially compromise coverage of the Welch test—the data are leptokurtic and typically held within some range by the test design of the EPA WET test methods. Such data are known to have little effect on coverage for the Welch t-test (Zimmerman 2006; Zar 1996).

So as not to rely on previous literature alone, simulations were conducted to demonstrate that the Welch t-test applied to the TST is suitable for WET test data. Simulated data were generated, having variances and non-normal distributions similar to WET test data for control and IWC groups. It is demonstrated that (a) moderately unequal variances (similar to WET data) have little effect on coverage of the t-test or Welch t-test (for normally-distributed data), and (b) for non-normally distributed data (similar in distribution to WET data) representing control and IWC groups, the TST using the Welch t-test has close to nominal coverage, on the basis of simulations with up to a nine-fold difference in variance between IWC and control (a relatively high difference in variances on the basis of observed WET test data).

Therefore, published studies provide ample evidence, the analysis of WET data and simulations described here, that the Welch t-test can be applied with confidence using the TST approach.

Characterization of WET Data

Because various WET test methods have a different experimental design, and thus could represent different distribution functions, a range of WET test methods (six) was examined to determine the frequency and magnitude of unequal variances between control and IWC as well as the frequency and type of non-normality in these methods. In addition, standard data transformations were used for tests when data were non-normal to see whether transformed data would meet assumptions of normality.

Unequal Variances

Standard F-tests ($p = 0.01$) were conducted for each valid WET test (IWC and control) to determine whether variances were unequal. Some WET test methods and endpoints demonstrated a higher frequency of unequal variances than other test methods (Table A-1).

Table A-1. Number (and percent) of tests with non-normal distribution and unequal variances for different types of WET tests, as well as the effect of data transformation on distribution, including skew and kurtosis

Test name	Number of tests	Data transformation	# (%) of non-normal tests ($p \leq 0.01$)	# (%) tests failing f-test for unequal variances ($p \leq 0.01$)	Range of skewness statistic for non-normal tests	# (%) tests failing D'Agostino test for skewness ($p \leq 0.01$)	Range of kurtosis statistic for non-normal tests	# (%) tests failing Anscombe test for kurtosis ($p \leq 0.01$)
C. dubia reproduction	1,382	Raw	285 (20.6)	390 (28.2)	-1.529 – -0.26	33 (2.4)	3.821 – 6.571	159 (11.5)
		Sqrt trans	418 (30.2)	545 (39.4)	-1.790 – -0.385	89 (6.4)	4.013 – 7.45	268 (19.4)
		Log +1	525 (37.9)	630 (45.6)	-2.058 – -0.564	143 (10.3)	4.06 – 8.43	343 (24.9)
Fish growth	108	Raw	2 (1.9)	18 (16.7)	-1.253 – 1.250	0 (0)	3.261 – 4.213	0 (0)
Mysid growth	907	Raw	10 (1.1)	37 (4.0)	-0.423 – 1.443	1 (0.1)	2.52 – 4.912	7 (0.77)
Kelp growth	100	Raw	9 (9.0)	22 (22)	-1.478 – 1.548	0 (0)	4.025 – 5.456	6 (6)
		Log+1	8 (8.0)	30 (30)	-1.571 – 1.234	0 (0)	4.25 – 6.080	8 (8)
		sqrt	9 (9.0)	29 (29)	-1.625 – 1.381	0 (0)	4.238 – 6.068	8 (8)
Kelp germination	100	Raw	3 (3.0)	15 (15)	-0.9 – 1.281	0 (0)	3.465 – 4.697	3 (3)
		arcsin(sqrt)	1 (1.0)	9 (9)	-0.872 – 1.04	0 (0)	3.465 – 4.698	0 (0)
Fish survival	108	percent	44 (40.7)	61 (56.5)	-1.633 – 0.654	0 (0)	2 – 4.67	3 (2.8)
		arcsin(sqrt)	42 (38.9)	61 (56.5)	-1.633 – 0	0 (0)	2 – 4.67	3 (2.8)

For example, over half of the *P. promelas* (fish) acute survival tests had unequal variances. That result is expected because control acute survival typically has little or no variance (i.e., all control replicates display 100 percent survival). *Ceriodaphnia* reproduction had the next highest frequency of tests with unequal variances (28.2 percent). The giant kelp growth or germination, and *P. promelas* (fish) chronic growth WET endpoints each had a lower frequency of tests with unequal variances (15–22 percent) while the mysid growth endpoint had the lowest frequency of unequal variances of the six test endpoints evaluated (4 percent). Using the *Ceriodaphnia* test method as an example of a WET method having a higher frequency of heterogeneous variances, the variance ratio between IWC and control was generally < 9:1 (95^{th} percentile ratio) with a median variance ratio of 2.5. Examination of data using other growth/reproduction methods indicates that most tests have a variance ratio < 10:1 (95^{th} percentile) and median variance ratio < 3.0. Percent data (germination) are subject to higher variance ratios (20~30:1); however, the fish acute test method has a variance ratio generally < 6.2:1 (95^{th} percentile).

Non-Normality

Shapiro's normality test was used to evaluate if WET test data were normally distributed. A measure of skewness was then used and Pearson's measure of kurtosis (R moments package) to examine if skewness or kurtosis or both are the major sources of non-normality. The critical values of those moments for a normal distribution are shown in Table A-2. A skewness measure significantly less than 0 indicates that the sample comes from a population that is skewed to the left, and a skewness measure significantly larger than 0 indicates that the distribution is skewed to the right. A kurtosis measure significantly larger than the median value (50^{th} percentile) for a given test design in Table A-2 indicates an underlying leptokurtic distribution. EPA also used the D'Agostino test of skewness (D'Agostino 1970) and Anscombe–Glynn test of kurtosis (Anscombe and Glynn 1983) for hypothesis testing.

Table A-2. Distribution of critical skewness and kurtosis ranges for different sample size (N) based on 1,000,000 simulation runs. N = 20 corresponds to *C. dubia* reproduction test (10 replicates in IWC and control); N = 16 corresponds to the Mysid chronic test (8 replicates per treatment); N = 10 corresponds to the two giant kelp chronic test endpoints (5 replicates per treatment); N = 8 corresponds to fathead minnow acute and chronic tests (four replicates per treatment)

N	Statistic	Percentiles						
		1%	5%	10%	50%	90%	95%	99%
20	Skewness	-1.152	-0.771	-0.587	0	0.588	0.772	1.155
	Kurtosis	1.645	1.831	1.951	2.551	3.667	4.151	5.361
16	Skewness	-1.244	-0.834	-0.635	0	0.635	0.833	1.247
	Kurtosis	1.562	1.746	1.866	2.477	3.629	4.126	5.351
10	Skewness	-1.407	-0.956	-0.729	0	0.726	0.953	1.404
	Kurtosis	1.387	1.563	1.679	2.289	3.463	3.940	4.972
8	Skewness	-1.453	-0.998	-0.766	0	0.766	0.997	1.450
	Kurtosis	1.318	1.470	1.583	2.173	3.319	3.731	4.567

The number of tests failing the hypothesis tests at 1 percent probability is reported in Table A-1. About 21 percent of the *Ceriodaphnia* reproduction tests (285 out of 1,382 cases) failed Shapiro's normality test (Table A-1). Both square root transformation and logarithm

transformation did not correct the non-normal distribution problem and instead increased the total number of tests failing the normality test (Table A-1). The D'Agostino test of skewness indicated that 33 tests (< 3 percent) were highly skewed. A test of kurtosis found 11 percent of tests (160) had significantly leptokurtic distribution (Table A-1). Apparently, most of the *Ceriodaphnia* test data failed the normality test because of kurtosis (leptokurtic distribution) and that occasional asymmetric distribution was mostly from outliers (Figure A-1). In general, most WET test growth data (i.e., *Pimephales promelas* growth, mysid growth, or kelp growth) were normally distributed. Both fish and mysid growth data exhibited non-normal distribution in only a very few cases (< 2 percent) and those were generally related to leptokurtic distributions that were short-tailed. Almost half of the acute fish survival tests had non-normally distributed data. Zero variance in many tests for either the control (34 cases) or IWC (26 cases) were the main cause of failing the normality test. Non-normality in acute fish survival data was because of leptokurtic data distribution (Table A-1).

The above analyses indicate that WET data in general do not have the distribution characteristics indicative of when Welch's t-test would be inappropriate (long-tail, highly skewed distribution).

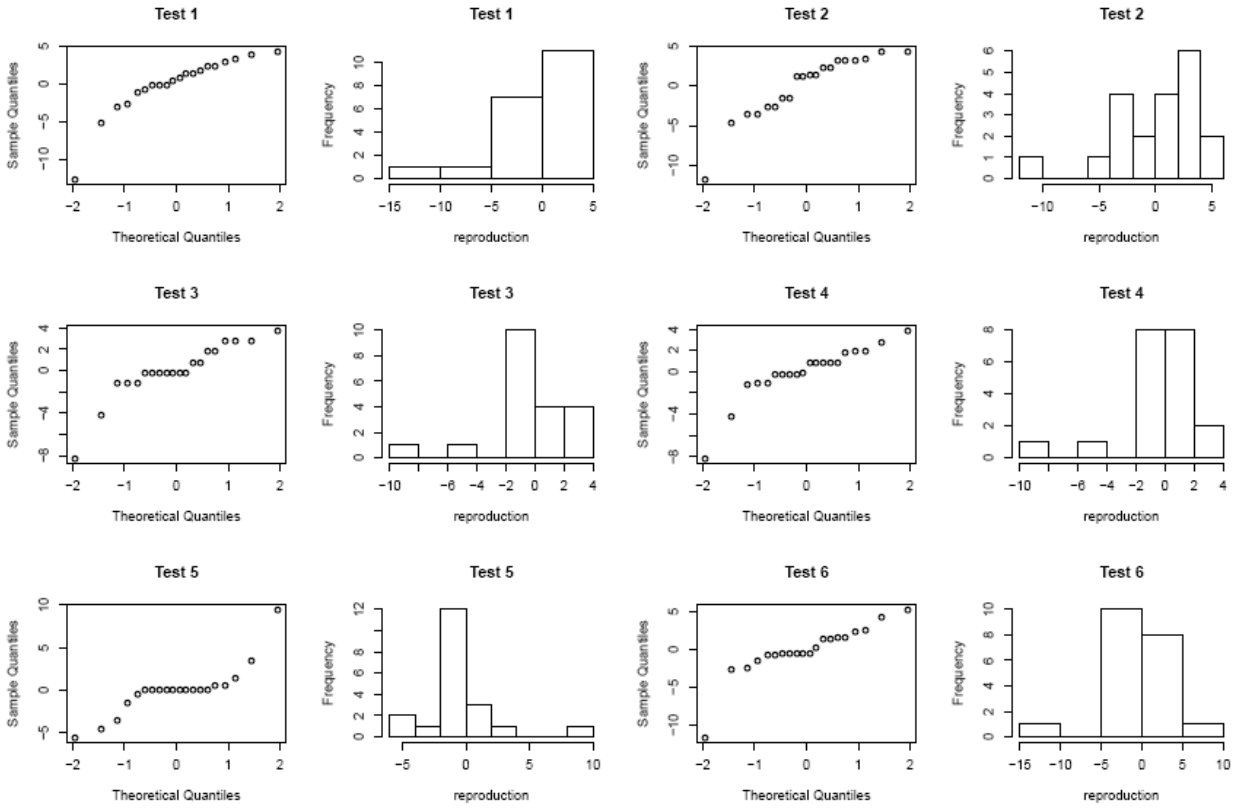

Figure A-1. Probability plots and histograms of examples of *Ceriodaphinia* chronic reproduction test data showing non-normal distribution and especially leptokurtic distribution.

Simulations

Unequal Variances

Various simulations were conducted using the chronic *Ceriodaphnia* test method as an example, to examine alpha error rate using either the traditional hypothesis t-test or Welch's t-test with data having different relationships between control and effluent variance. From analyses of more than 2,000 WET tests presented in Table A-1, a variance ratio (IWC/control) of 9:1 (95^{th} percentile of variance ratio) is a reasonable upper limit. Therefore, simulation scenarios examined included (1) equal variances and no mean difference between control and effluent; (2) IWC with 9 times the control variance and no mean difference; (3) equal variance and a 25 percent mean effect of the IWC; and (4) IWC with 9 times the control variance and a 25 percent mean effect. Equal sample size (N = 10 using *Ceriodaphnia* chronic test method as the example) was assumed for both control and treatment group which is most often the case in WET analyses. Results are shown in Table A-3.

Table A-3. Results of Monte Carlo simulations evaluating alpha error rate using either the traditional t-test or Welch's t-test with data having different relationships between control and effluent variances. S_c^2 = control variance, S_t^2 = IWC variance, μ_c = control mean, and μ_t = IWC mean. Results are based on 1,000,000 simulation runs per scenario.

		$\mu_c = \mu_t$		$\mu_t = 0.75\ \mu_c$	
	Alpha	**T-test**	**Welch t-test**	**T-test**	**Welch t-test**
$S_c^2 = S_t^2$	0.010	0.0098	0.0093	0.0099	0.0095
	0.050	0.0498	0.0490	0.0497	0.0491
	0.100	0.0996	0.0988	0.1000	0.0992
	0.150	0.1493	0.1486	0.1501	0.1506
	0.200	0.1996	0.1991	0.2000	0.1997
	0.250	0.2498	0.2493	0.2502	0.2498
$S_c^2 = S_t^2/9$	0.010	0.0132	0.0105	0.0204	0.0103
	0.050	0.0550	0.0503	0.0725	0.0503
	0.100	0.1050	0.1001	0.1269	0.1002
	0.150	0.1543	0.1501	0.1774	0.1499
	0.200	0.2037	0.2003	0.2260	0.1999
	0.250	0.2526	0.2499	0.2732	0.2499

When there are equal variances and the true difference is equal to 0, the observed error rates from both the traditional t-test and Welch's t-test are very close to the expected error rates. When control and treatment groups have unequal variance, (effluent variance = 9 times the control variance), the traditional t-test has a slightly higher Type I error rate, but Welch's t-test has a Type I error rate similar to the expected value. When the true response at the IWC is 0.75 × control mean, and both populations have equal variances, alpha error rates are very similar to expected using both the traditional t-test and Welch's t-test. When the true response at the IWC is 0.75 × control mean and population variances are not equal (i.e., effluent variance is 9 times

the control variance), the error rates are about 2–3 percent higher than expected using the traditional t-test but are similar to expected alphas using Welch's t-test.

While the specific results pertain to the *Ceriodaphnia* reproduction endpoint, the general conclusions of this analysis would apply to all WET methods and endpoints. Such results confirm that Welch's t-test has better coverage than the traditional t-test using the TST approach when variances are unequal.

Non-Normality

The objective of the simulations was to confirm that the alpha error rate is relatively stable against deviations from non-normal distribution when variances are unequal as well for both the traditional hypothesis test and Welch's t-test.

EPA examined the distribution of control and effluent reproduction data from 281 *C. dubia* multiple concentration tests (Figure A-2). While most tests indicate that control reproduction follows a normal distribution (mean = 24.5, standard deviation = 5.56), effluent data tend to deviate from a normal distribution: effluents with low toxicity have less skewed data, while effluents with data that have high toxicity are more likely to deviate from normal distribution. To address this observation, two populations were simulated on the basis of the shape of the frequency distribution in the highest effluent concentration in each *C. dubia* test (Figure A-3). The first simulated effluent population had a mean = 25 (equal to the population mean for the control group) and a standard deviation = 7.7, while the second one had a population mean of $b \times 25$ (where $b = 0.75$ for chronic test methods), resulting in an effluent mean of 18.75. The variance of those two effluent populations was the same. Random samples taken from these two populations were used to compare with the control population data (mean = 25, standard deviation = 5.56).

Simulation results (Table A-4) indicate that when the two populations had the same mean but had a different distribution shape as compared to a normal distribution (control population), the alpha error rate using the traditional t-test was about 1 percent higher than expected. Welch's modified t-test slightly corrected the error rate (Table A-4). When the true population mean difference between control and effluent is 25 percent of the control mean and when the effluent population is not normally distributed, the alpha error rate is almost identical to the expected value using traditional t-test (Table A-4). Welch's t-test resulted in a decrease in the nominal alpha error rate by 2–3 percent using the TST approach. That is, when data are extremely non-normal (for WET test data) and variances are heterogeneous between control and effluent, Welch's t-test is less likely to reject the null hypothesis and slightly more likely to declare a sample toxic than expected (i.e., the analysis will be more conservative). As data approach a normal distribution, α error rates using Welch's t-test will be closer to nominal values.

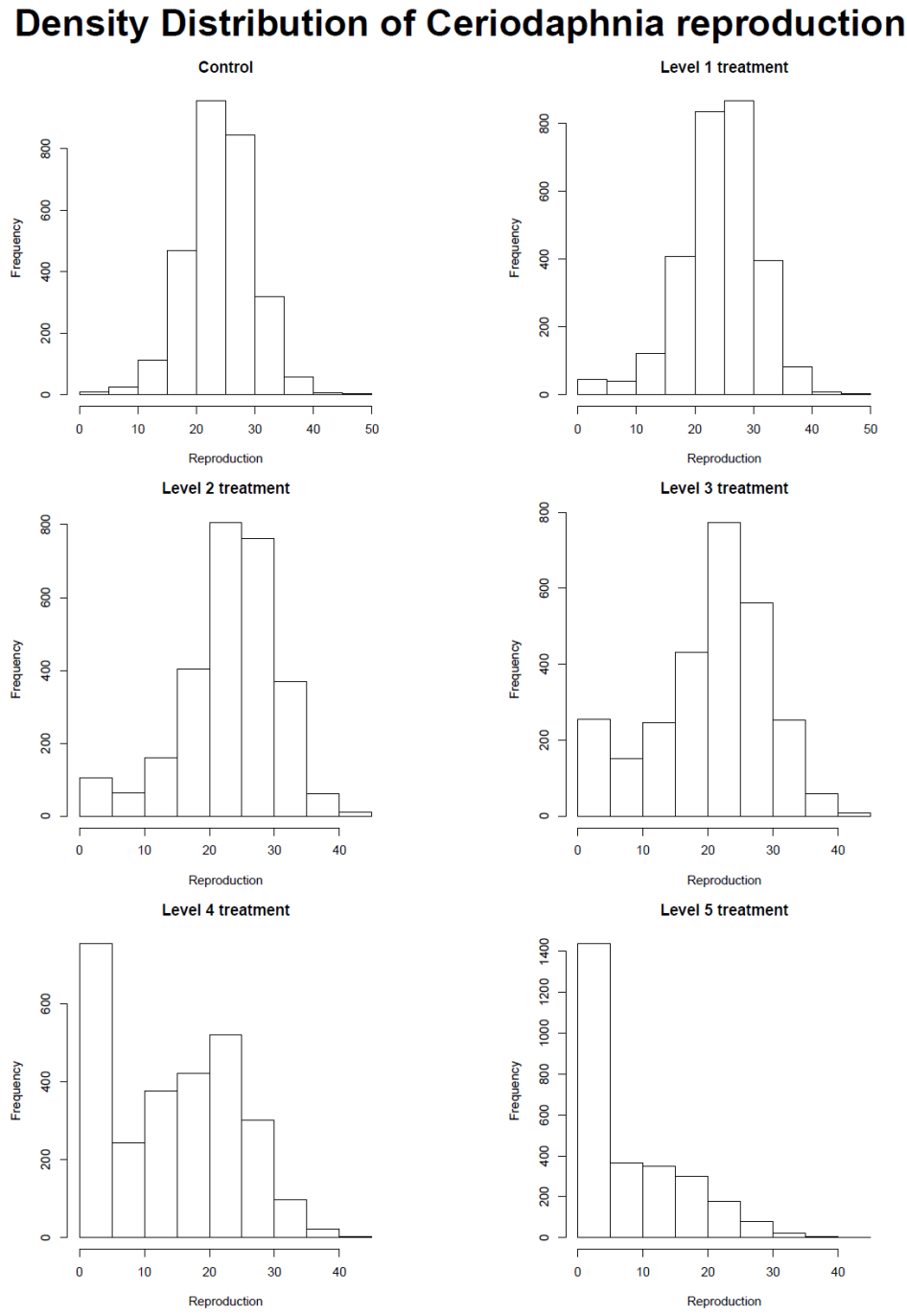

Figure A-2. Histogram of observed *Ceriodaphnia* reproduction at different level of effluent concentrations based on 281 multiple concentration tests.

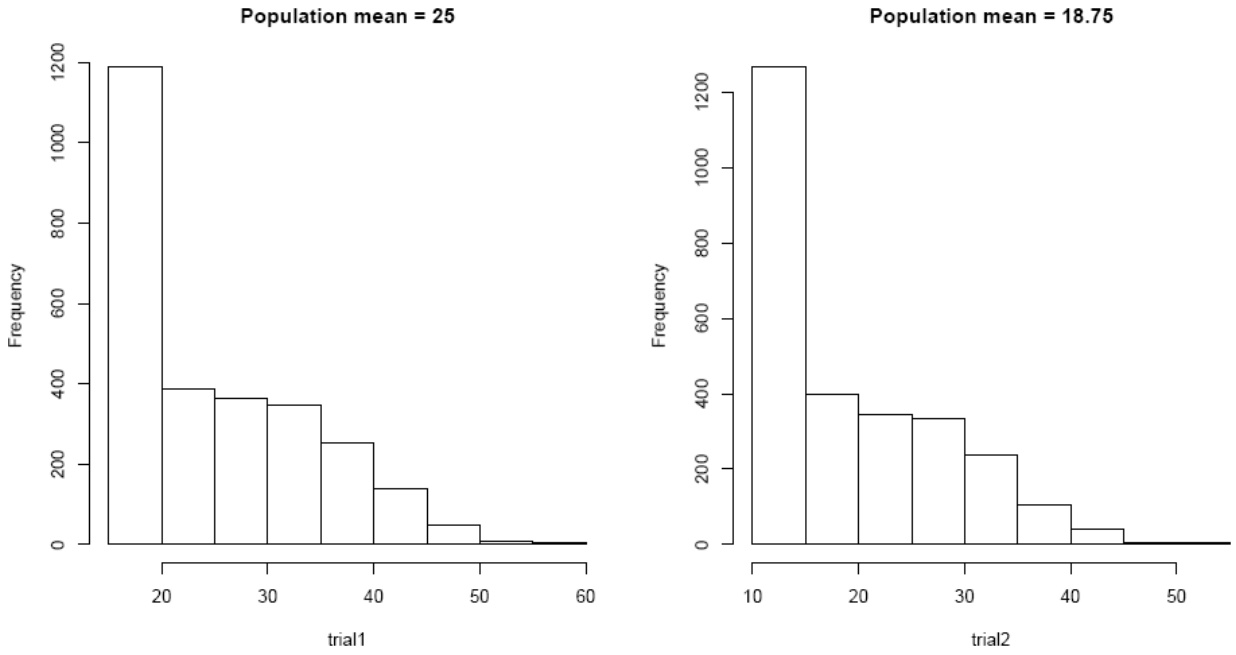

Figure A-3. Simulated frequency distributions of Ceriodaphnia reproduction data with two populations having non-normal data and different means. Both populations have a standard deviation of 7.7.

Table A-4. Results of Monte Carlo simulation analyses (100,000 simulations per scenario) indicating alpha error rates based on comparisons between two non-normally distributed populations and a normal distribution (control population, mean = 25, standard deviation = 5.65). The population means are 25 and 18.75, respectively, and the standard deviation is 7.7 in both populations.

Alpha	Welch's (μ = 25)	Traditional t (μ = 25)	TST t-test (μ = 18.75, b = 0.75)	TST Welch's (μ = 18.75, b = 0.75)
0.05	0.053	0.059	0.043	0.031
0.10	0.104	0.108	0.090	0.074
0.15	0.151	0.155	0.140	0.122
0.20	0.199	0.203	0.191	0.173

Although the simulated population does not necessarily represent the true population of effluent groups, EPA's examination of sample distribution indicates that effluent populations with low toxicity are less likely to deviate from normal distribution. The simulation also indicates that the alpha error rate using Welch's t-test under severely non-normal distributions and heterogeneous variances is less than the expected/critical values. That is, Welch's t-test is more conservative when toxicity is high (a desirable attribute for WET analysis) than when effluent toxicity is low. When effluent toxicity is low, results of analyses using *Ceriodaphnia* reproduction WET test data indicate that the effluent data are less likely to be non-normally distributed, and the observed alpha error rate approaches the expected error rate. On the basis of the foregoing results, the type of non-normal distribution observed in WET tests should not affect the overall performance of simulation analyses used to derive test method alpha values for the TST approach.

Rationale/Conclusions

When population variances are not equal or test samples are non-normally distributed (or both), concerns could be raised in using the two concentration t-test or the bioequivalence t-test (Erickson and McDonald 1995) because statistical assumptions might not be met. EPA WET test methods specify that if the data fail Shapiro-Wilks's normality test or Bartlett's homoscedasticity test (or both), a non-parametric test such as Wilcoxon Rank sum test should be used in such situations. Extension of such nonparametric tests to TST is, however, complicated because the null hypothesis for those tests is that results from control and effluent are from same population. This is stated as the null hypothesis of no difference among treatments. Because an effect size $1-(b \times \mu_0)$ is specified in the TST approach that is related to the control population mean, a non-parametric equivalent to a t-test approach using a bioequivalence formulation (such as with the TST approach), has been difficult to demonstrate (Zimmerman and Zumbo 1993; Manly 2004).

Data compiled from more than 2,000 valid WET tests in this project confirmed that the type of distributions exhibited by most test data do not seriously compromise the use of a t-test. The data can be dealt with appropriately using Welch's t-tests for unequal variances, as shown in simulation analyses. Use of Welch's t-test for TST analysis is supported on the basis of analysis of actual WET test data, which indicate that the majority of WET test data are normally distributed or have a leptokurtic distribution with short tails such that the use of Welch's t-test produces Type I error rates very close to expected error rates. Statistical literature indicates that actual power of the t-test (and by extension Welch's t-test) is *greater* when populations are leptokurtic, especially for small sample sizes (Zar 1996).

WET test data are biologically expected to have short-tailed distributions supporting the use of Welch's t-test because of the test method's required test acceptability criteria and test termination times, which constrain the range of endpoint responses encountered. For example, a chronic *Ceriodaphnia dubia* test must have 80 percent or greater survival and an average of 15 or more young per surviving female in the control for the test to meet the required test acceptability criteria (i.e., a valid test). Additionally, test termination is prescribed in the method as the time at which at least 60 percent or more of the surviving control females generate at least three broods, which can be 6–8 days (maximum is 8 days), also a test requirement. That results in a lower distribution bound (e.g., reproduction responses in controls start at 15). In addition, the upper part of the distribution cannot go to infinity, even if populations were to survive and reproduce beyond the prescribed test requirements because of biological constraints. Similar test method and biological constraints apply to all other WET test endpoints (e.g., growth, survival).

Furthermore, Welch's t-test is robust to non-normal distributions when the underlying distribution is symmetric and skewness is low, especially with sample sizes > 10 (Tiku 1971; Lee and D'Agostino 1976; Tiku and Akkaya 2004). For the West Coast WET methods examined and the *Ceriodaphnia* and Mysid chronic WET method evaluated, those conditions are met. Therefore, at least for those WET methods and others with similarly large sample sizes, Welch's t-test should not result in a substantial underestimation of the Type I error rate.

In addition, the Type I error rate using TST for several WET methods is set ≥ 0.05. The higher α levels include WET test methods that have smaller sample sizes such as the fathead minnow acute test. For those methods, the slight overestimation of the nominal Type I error rate that can occur using Welch's t-test when WET test data are not normally distributed is insignificant given

the higher nominal α levels established. For the West Coast WET test methods that have α levels set at 0.05, effect size examined in those test methods is large and, in many cases, data are normally distributed even without data transformation (e.g., giant kelp germination and tube-length endpoints, Table A-1).

The observed sample distribution from 281 *C. dubia* multiple concentration tests indicates that test populations at low effluent concentrations are less likely to deviate from normal distribution. A similar trend is expected for other WET endpoints such as growth. The simulation based on the distribution shape of the high effluent concentration population also indicates that the alpha error rate using Welch's t-test is less than expected. That is, Welch's t-test is more conservative when toxicity is high. Therefore, the type of non-normal distribution observed in WET tests should not negatively affect the outcome of TST analyses.

Analyses used to develop the TST analysis approach indicate that data transformation (log or square root) does not help the non-normality issue for WET test data (Table A-1). That is usually because of the leptokurtic distribution observed rather than because of skewness of data (Table A-2). Therefore, data transformation before TST analysis is not recommended except for percent data, which should be arcsine square root transformed before TST analysis (consistent with current EPA analysis recommendations). This precaution is suggested because percent data (especially acute percent survival) is most prone to non-normality.

In conclusion, given the leptokurtic and short-tailed distribution of most WET test data, as well as the other factors noted above, Welch's t-test is appropriate to use for one-tailed, two-sample comparisons using TST. Furthermore, because Welch's t-test performs as effectively as the t-test in terms of Type I error when data are normally distributed and variances are equal (Moser et al. 1989; Coombs et al. 1996), Welch's t-test should be used for all WET test data analysis using TST. Furthermore, many researchers have shown that the combination of using a preliminary variance test (e.g., F-test) plus a t-test does not control Type I error rates as well as simply always performing an unequal variance t-test such as Welch's t-test (Gans 1992; Moser and Stevens 1992). That is one reason why it is generally unwise to decide whether to perform one statistical test on the basis of the outcome of another (Smith 1936; Markowski and Markowski 1990; Zimmerman 2004).

Literature Cited

Anscombe, F. and W. Glynn. 1983. Distributions of the kurtosis statistic b_2 for normal statistics. *Biometrika* 70:227–234.

Coombs, W., J. Algina, and D. Oltman. 1996. Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not necessarily equal. *Review Educational Research* 66:137–79.

D'Agostino, R. 1970. Transformation to normality of the null distribution of g_1. *Biometrika* 58:341–348.

Erickson, W., and L. McDonald. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. *Environmental Toxicology and Chemistry* 14:1247–1256.

Gans, D. 1992. Preliminary tests on variances. *American Statistics* 45:258.

Lee, A., and R. D'Agostino. 1976. Levels of significance of some two-sample tests when observations are from compound normal distributions. *Communications In Statistics* A5(4):325–342.

Manly, B. 2004. One-sided tests of bioequivalence with non-normal distributions and unequal variances. *Journal of Agricultural, Biological, and Environmental Statistics* 9(3):270–283.

Markowski, C., and E. Markowski. 1990. Conditions for the effectiveness of a preliminary test of variance. *American Statistics* 44:322–6.

Moser, B., and G. Stevens. 1992. Homogeneity of variance in the two-sample means test. *American Statistics* 46:19–21.

Moser, B. G. Stevens, and C. Watts. 1989. The two-sample t-test versus Satterwaite's approximate F test. *Communications in Statistics—Theory and Methods* 18:3963–75.

Smith, H. 1936. The problem of comparing the results of two experiments with unequal errors. *Journal of Scientific & Industrial Research* 9:211–922.

Tiku, M. 1971. Student's *t* distribution under nonnormal situations. *Australian Journal of Statistics* 13:142–148.

Tiku, M., and A. Akkaya. 2004. *Robust Estimating and Hypothesis Testing*. New Age International Limited, Publishers New Delhi, India.

Welch, B. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29:350–362.

Welch, B. 1947. The generalisation of students problem when several different population variances are involved. *Biometrika* 34:23–35.

Zar, J. 1996. *Biostatistical Analysis*. 3rd ed. Prentice Hall International, Princeton, NJ.

Zimmerman, D. 2004. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology* 57:173–181.

Zimmerman, D. 2006. Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology* 3(4):351–374.

Zimmerman, D., and B. Zumbo. 1993. Rank transformations and the power of the Student t-test and Welch's t-test for non-normal populations. *Canadian Journal of Experimental Psychology* 47:523–539.

APPENDIX B

STEP-BY-STEP PROCEDURES FOR ANALYZING VALID WET DATA USING THE TST APPROACH

APPENDIX B

STEP-BY-STEP PROCEDURES FOR ANALYZING VALID WET DATA USING THE TST APPROACH

The following is a step-by-step guide for using the TST approach to analyze valid WET data for the NPDES WET Program. This guide is applicable for a two-concentration data analysis of an IWC or a receiving water concentration compared to a control concentration. For further information regarding conducting WET tests and proper quality assurance/quality control needed, see the EPA WET method manuals. As you proceed through this guide, refer to the flowchart shown in Figure B-1 of this appendix.

Step 1: Conduct WET test following procedures in the appropriate EPA WET test method manual. This includes following all test requirements specified in the method (USEPA 1995 for chronic West Coast marine methods, USEPA 2002a for chronic freshwater WET methods, USEPA 2002b for chronic East Coast marine WET methods, and USEPA 2002c for acute freshwater and marine methods).

Step 2: For each test endpoint specified in the WET test method manual (e.g., survival and reproduction for the *Ceriodaphnia* chronic WET test method), follow Steps 3–7 below. Note that the guide refers to an effluent concentration tested, which is assumed to be the IWC as specified in the permit or a receiving water concentration for ambient testing. For example, if no mixing zone is allocated, the IWC is 100 percent effluent.

Note: If there is no variance (i.e., zero variance) in the endpoint in both concentrations being compared (i.e., all replicates in each concentration have the same exact response), then skip the remaining steps in the flowchart and do the following. Compute the percent difference between the control and the other concentration (e.g., IWC) and compare the percent difference against the RMD values of 25% for chronic and 20% for acute endpoints. Percent mean effect is calculated as:

$$\%\ \text{Effect at IWC} = \frac{\text{Mean Control Response} - \text{Mean Response at IWC}}{\text{Mean Control Response}} \times 100$$

If the percent mean response is $\geq$ the RMD, the sample is declared toxic and the test is "*Fail*". If the percent mean response is < the RMD, the sample is declared non-toxic and the test is "*Pass*".

Step 3: For data consisting of proportions from a binomial (response/no response; live/dead) response variable, the variance within the *i*th treatment is proportional to $P_i (1 - P_i)$, where P_i is the expected proportion for the treatment. That clearly violates the homogeneity of variance assumption required by parametric procedures such as the TST procedure because the existence of a treatment effect implies different values of P_i for different treatments, *i*. Also, when the observed proportions are based on small samples, or when P_i is close to zero or one, the normality assumption might be invalid. The arcsine square root (arcsine $\sqrt{P}$) transformation is used for such data to stabilize the variance and satisfy the normality requirement. The square root of percent data (e.g., percent survival, percent fertilization), expressed as a decimal fraction (where 1.00 = 100 percent) for each treatment, is first calculated. The square root value is then

arcsine transformed before analysis in Step 4. Note: Excel and most statistical software packages can calculate arcsine values.

Step 4: Conduct Welch's t-test (Zar 1996) using Equation 1:

Equation 1

$$t = \frac{\overline{Y_t} - b \times \overline{Y_c}}{\sqrt{\frac{S_t^{\,2}}{n_t} + \frac{b^2 S_c^{\,2}}{n_c}}}$$

where

$\overline{Y_c}$ = Mean for the control

$\overline{Y_t}$ = Mean for the IWC

$s_c^{\,2}$ = Estimate of the variance for the control

$s_t^{\,2}$ = Estimate of the variance for the IWC

n_c = Number of replicates for the control

n_t = Number of replicates for the IWC

b = 0.75 for chronic tests; 0.80 for acute tests

Note on the use of Welch's t-test: Welch's t-test is appropriate to use when there are an unequal number of replicates between control and the IWC. When sample sizes of the control and treatment are the same (i.e., $n_t = n_{c)}$, Welch's t-test is equivalent to the usual Student's t-test (Zar 1996).

Step 5: Adjust the degrees of freedom (df) using Equation 2:

Equation 2

$$\upsilon = \frac{(\frac{S_t^{\,2}}{n_t} + \frac{b^2 S_c^{\,2}}{n_c})^2}{\frac{(\frac{S_t^{\,2}}{n_t})^2}{n_t - 1} + \frac{(\frac{b^2 S_c^{\,2}}{n_c})^2}{n_c - 1}}$$

Using Welch's t-test, df is the value obtained for *v* in Equation 2 above. Because *v is* most likely a non-integer, round *v* to the next smallest integer, and that number is the df.

Step 6: Using the calculated t value from Step 4, compare that *t* value with the critical *t* value table in Appendix C using the test method-specific alpha values shown in Table 4-1. To obtain the correct *t* value, look across the table for the alpha value that corresponds to the WET test method (for the appropriate alpha value, see Table 4-1 of this document) and then look down the table for the appropriate df.

Step 7: If the calculated *t* value is less than the critical *t* value, the IWC is declared toxic, and the test result is *Fail*. If the calculated *t* value is greater than the critical *t* value, the IWC is not declared toxic and the test result is *Pass*.

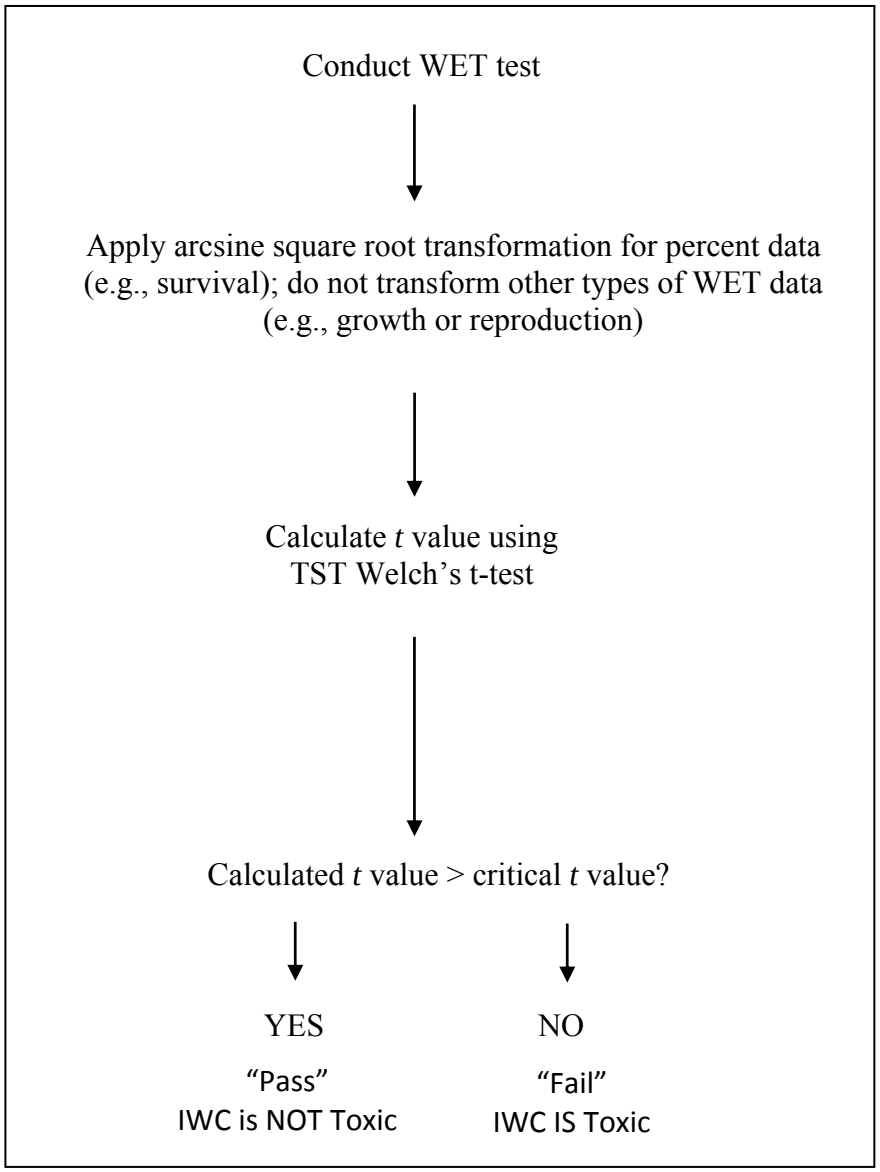

Figure B-1. Statistical flowchart for analyzing valid WET data using the TST approach for control and the IWC, receiving water, or stormwater.

Literature Cited

USEPA (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*. Eds. G. Chapman, D. Denton, and J. Lazorchak. EPA/600/R-95-136. U.S. Environmental Protection Agency, National Exposure Research Laboratory, Cincinnati, OH, Office of Research and Development, Washington, D.C.

USEPA (U.S. Environmental Protection Agency). 2002a. *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms*. 4th ed. EPA/821/R-02-013. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2002b. *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms*. 3rd ed. EPA/821/R-02-14. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

USEPA (U.S. Environmental Protection Agency). 2002c. *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*. 5th ed. EPA/821/R-02-012. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

Zar, J. 1996. *Biostatistical Analysis*. 3rd ed. Prentice Hall Publishers, Princeton, NJ.

APPENDIX C

CRITICAL *t* VALUES FOR THE TEST OF SIGNIFICANT TOXICITY APPROACH

Table C-1. Critical values of the *t* distribution. One tail probability is assumed.

	Alpha				
Degrees of freedom	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**
1	1	1.3764	1.9626	3.0777	6.3138
2	0.8165	1.0607	1.3862	1.8856	2.92
3	0.7649	0.9785	1.2498	1.6377	2.3534
4	0.7407	0.941	1.1896	1.5332	2.1318
5	0.7267	0.9195	1.1558	1.4759	2.015
6	0.7176	0.9057	1.1342	1.4398	1.9432
7	0.7111	0.896	1.1192	1.4149	1.8946
8	0.7064	0.8889	1.1081	1.3968	1.8595
9	0.7027	0.8834	1.0997	1.383	1.8331
10	0.6998	0.8791	1.0931	1.3722	1.8125
11	0.6974	0.8755	1.0877	1.3634	1.7959
12	0.6955	0.8726	1.0832	1.3562	1.7823
13	0.6938	0.8702	1.0795	1.3502	1.7709
14	0.6924	0.8681	1.0763	1.345	1.7613
15	0.6912	0.8662	1.0735	1.3406	1.7531
16	0.6901	0.8647	1.0711	1.3368	1.7459
17	0.6892	0.8633	1.069	1.3334	1.7396
18	0.6884	0.862	1.0672	1.3304	1.7341
19	0.6876	0.861	1.0655	1.3277	1.7291
20	0.687	0.86	1.064	1.3253	1.7247
21	0.6864	0.8591	1.0627	1.3232	1.7207
22	0.6858	0.8583	1.0614	1.3212	1.7171
23	0.6853	0.8575	1.0603	1.3195	1.7139
24	0.6849	0.8569	1.0593	1.3178	1.7109
25	0.6844	0.8562	1.0584	1.3163	1.7081
26	0.684	0.8557	1.0575	1.315	1.7056
27	0.6837	0.8551	1.0567	1.3137	1.7033
28	0.6834	0.8546	1.056	1.3125	1.7011
29	0.683	0.8542	1.0553	1.3114	1.6991
30	0.6828	0.8538	1.0547	1.3104	1.6973
inf	0.6745	0.8416	1.0364	1.2816	1.6449