Synopsis of Peer Review of TST Approach

A. Before EPA External Peer Review

The 1995 Pellston Conference on whole effluent toxicity (WET), attended by 50 WET experts from across the U.S., recognized that a bioequivalence testing approach to WET data analysis had many advantages that could address several limitations identified for the traditional hypothesis testing approach, and the point estimate approach used for WET analysis (Chapman et al. 1996). The consensus of these experts was that studies should be initiated to both explore the test of bioequivalence, and perform a critical assessment of the statistical properties of bioequivalence testing and the appropriate biological effect levels to be used. These experts also agreed that the test of bioequivalence should follow the procedures outlined in Erickson and McDonald (1995).

Erickson and MacDonald (1995) established the premise for the bioequivalence testing approach to be used for toxicity testing: the underlying hypothesis and t-test formula. However, they recognized that a clear regulatory management decision threshold regarding what is or is not acceptable toxicity (in terms of an effect in WET tests) was lacking. In other words, their paper did not establish an explicitly defined bioequivalence value (b). Without a regulatory management decision threshold for b, a bioequivalence approach could not be readily implemented in WET or receiving water monitoring and assessment programs.

Erickson W and McDonald L. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. *Environ. Toxicol. Chem.* 14:1247-1256.

B. EPA External Peer Review

External peer review was conducted using the guideline for EPA's peer review process (U.S. EPA 2006). EPA's Office of Wastewater Management (OWM) prepared the peer review charge questions (document #1). This review was conducted by the independent contractor, Avanti. Questions along with the draft TST report (document #2) were submitted to external peer reviewers by the contractor. The five reviewers were independently selected, based on their qualifications, by the contractor. Avanti summarized the comments and removed any information which would identify the reviewers and submitted the summary review document to OWM (document #3). Note that the initial TST approach set alpha at 0.05 for all toxicity test methods and varied the b value depending on the desired maximum false positive and false negative rates. The external peer reviewers recommended keeping b at a fixed risk management level based on ecological information and not dependent on test method performance or test design. Peer reviewers unanimously concurred that the bioequivalence approach used in the TST is a sound approach for the WET program. There was also consensus among peer reviewers that the analytical approach used to develop the TST, and the results of the TST analyses, are reasonable and defensible. Some peer reviewers commented that the dependence on empirical WET data used in the initial approach was somewhat limiting and that future analyses should also include simulations or other tools to obtain true population error rates when the TST is used.

As a result of the peer review, the final TST approach was refined. In particular, the document reflects established fixed b values at 0.75 for chronic WET test endpoints and 0.80 for acute WET test endpoints. These b values are regulatory management decisions for toxicity made by EPA. More extensive Monte Carlo simulation analyses where conducted to develop population false positive and false negative rates using regulatory management decision thresholds for unacceptable toxicity (25% effect for chronic endpoints and 20% effect for acute endpoints), and for acceptable toxicity (10% effect for all endpoints).

The Draft Technical Document was revised and updated to reflect the external peer review comments, editorial improvements, and internal Agency suggestions. OWM conducted various communication webinars with EPA Regions and States before the final release of the technical document (document #4) by EPA's OWM Director, Jim Hanlon, on June 2010 (document #5).

Document #1 (pages 4 – 8): EPA OWM charge questions.
Document #2 (pages 9 – 148): May 20, 2008 Draft TST Technical Document.
Document #3 (pages 149 – 204): Summary of Peer Review of EPA TST document submitted by Avanti to EPA OWM, dated October 23, 2008.
Document #4 (pages 205 – 320): U.S. EPA. 2010. National Pollutant Discharge Elimination System Test of Significant Toxicity Technical Document. EPA/833-R-10-004, U.S. Environmental Protection Agency, Office of Environmental Management, Washington, DC.
Document #5 (pages 321 – 322): Jim Hanlon, Director of OWM's transmittal of the final technical document, June 2010.

C. <u>After EPA External Peer Review: Journal Peer Review</u>

"Peer-reviewed journal articles (written by EPA or non-EPA authors) performed by a credible, referred scientific journal contributes to the scientific and technical credibility of the reviewed product. Generally, EPA considers peer review by such journals as adequate for reviewing the scientific credibility and validity of the findings (or data) in that article, and therefore a satisfactory form of peer review." (U.S. EPA 2006)

Debra Denton and Jerry Diamond submitted the 2010 final TST approach for consideration to the international peer reviewed journal, *Environmental Toxicology and Chemistry*. This article was reviewed by three independent, anonymous reviewers and accepted for publication with minor edits. This TST article was published in April 2011. Diamond et al. submitted a paper to the international peer reviewed journal, *Integrated Environmental Management and Assessment*, titled, "It is time for change in the analysis of whole effluent toxicity data." This article was reviewed by two independent reviewers and has been accepted for publication. (Note: Under the "Discussion" for change #2 in this article, the correct alternative hypothesis is: *Alternative Hypothesis:* $\mu_T > b \times \mu_{C.}$)

Denton D, Diamond J, Zheng L. 2011. Test of significant toxicity: a statistical application for assessing whether an effluent or site water is truly toxic. *Environ. Toxicol. Chem.* Vol. 30:1117-1126.

Diamond J, Denton D, Anderson B, Phillips B. in press. It is time for changes in the analysis of whole effluent toxicity data. *Integrated Environmental Management and Assessment*.

References:

Chapman, G., et al. 1996. Methods and Appropriate Endpoints. in: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins (editors). *Whole effluent toxicity testing: An evaluation of methods and prediction of receiving system impacts*, SETAC Press, Pensacola, Florida. pp. 51-82.

U.S. EPA. 2006. Peer Review Handbook: *Science Policy Council Peer Review Handbook, 3rd Edition (June 2006).* EPA/100-B-06-002. Office of Science Policy and Office of Research and Development, Washington, DC.

CHARGE TO EXTERNAL PEER REVIEWERS (8/5/08)

***Subject*: Evaluation of the Draft Test of Significant Toxicity Approach as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity**

The U.S. Environmental Protection Agency (EPA), Office of Wastewater Management/Water Permits Division/State and Regional Branch has requested a technical and scientific review of an alternative statistical approach to existing recommended WET data analyses in the EPA WET test methods and other EPA guidance documents. This draft alternative statistical approach has been described in detail in a draft document entitled," *Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity."* This draft document proposes and demonstrates a statistical approach which builds upon existing EPA approaches yet provides advancements and improvements in analyzing and interpreting results of whole effluent toxicity tests (WET) for the purposes of NPDES permitting regulation and compliance (see attachment). EPA is conducting this external peer review of the draft Test of Significant Toxicity (TST) document, in addition to a previous scientific and technical review by EPA's Office of Research and Development (ORD). Together these two reviews should support the Agency's overall objective to enhance the quality and credibility of the Agency's NPDES regulatory decisions by ensuring that the scientific and technical work products (tool box of statistical analyses) underlying these decisions receive appropriate levels of peer review by both EPA's scientific and technical experts and independent external experts in the field. This peer review charge was developed in accordance with EPA's 2006 peer review handbook, "Science Policy Council's Peer Review Handbook" (3rd Edition, USEPA, 2006) which was issued to ensure that the Agency uses credible and appropriate science including the evaluation of WET test method applications and implementation under EPA's NPDES permit program.

Peer Review Objective

The draft document, *"Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity"* provides technical information; a description of the proposed draft TST methodology; and a detailed discussion of the following technical points: how the draft TST statistical approach was developed for several representative EPA WET test methods; how results using the draft TST approach compare with those obtained using current recommended EPA analysis approaches; and how the draft TST approach performs in terms of common statistical metrics such as power, confidence, and sensitivity, as indicated by the rate at which the draft TST approach recognizes a truly toxic sample. In addition, this draft document is intended to demonstrate how the draft TST approach can be used to analyze both acute and chronic WET test data, as well as ambient (2 concentrations) and WET (multiple concentrations) test data.

The main function of this technical peer review is to assess whether the methodology described in draft TST document and the document's presentation of the proposed methodology provides usable and sound technical information and recommendations regarding this draft alternative statistical approach for WET analysis. As part of the main function in assessing the draft

document's technical approach EPA is particularly interested in a technical evaluation of the test method-specific bioequivalency "*b*" values proposed in the draft document. The draft TST document provides the underlying science for deriving the "*b*" values and their application to the NPDES program.

Background on WET:

EPA's WET testing policies and NPDES regulations are intended to support the goals of the Clean Water Act (CWA) or more specifically, to provide for the protection and propagation of fish, shellfish, and wildlife. The CWA (section 101(a)(3)) states that "it is the national policy that the discharge of toxic pollutants in toxic amounts be prohibited." EPA and States authorized to administer the NPDES permitting program have pursued this objective through the water quality standards (WQS) program and the NPDES permitting program. A major step forward for toxics control was the adoption of water quality-based permitting to integrate chemical and biological monitoring to protect receiving water quality. The integration of the effluent effects and receiving water exposure measurements resulted in the development of effluent hazard assessment approaches.

Acute and short-term chronic WET tests estimate the toxicity of wastewaters in order to protect States' aquatic life criteria. These tests measure the aggregate toxic effects of an effluent to standardized, aquatic plants, vertebrates or invertebrates. The standardized tests are used to monitor both effluents and receiving waters, and to measure compliance with WET limits in the NPDES permits. If a discharge causes, has a reasonable potential to cause, or contributes to an in stream excursion above a numeric water quality criterion for WET (i.e., demonstrates reasonable potential) then a WET limit is required in the NPDES permit (40 CFR 122.44(d)(1)(iv)). If a discharge demonstrates reasonable potential to exceed a narrative water quality criterion (e.g., a water quality criterion to prevent the discharge of toxic pollutants in toxic amounts) the permit must contain WET limits unless the permitting authority has identified the parameters causing toxicity and placed chemical specific limits in the permit to appropriately control those parameters (40 CFR 122.44(d)(1)(v)).

Currently, the Agency recommends either the hypothesis test (e.g., *No Observed Effect Concentration* or NOEC) or the point estimate (e.g., *Inhibition Concentration* such as the IC_{25}) approaches, as described in the *Technical Support Document for Water Quality-based Toxics Control* (USEPA, 1991). Additional method guidance is provided in (USEPA 1995, 2002a, 2000b, 2000c) to analyze WET data and to determine compliance with permit conditions or water quality standards. Both of the above approaches have strengths and limitations in terms of their ability to consistently identify truly toxic conditions or truly non-toxic conditions when they occur (Chapman et al., 1996). Chapman et al., (1996) recommended that studies be initiated to evaluate improvements in the statistical analysis of WET test data, such as exploring tests of bioequivalence using WET databases. The alternative approach discussed in this draft document addresses these limitations, advancing the Technical Support Document's (TSD) approach for WET analysis.

The results of a single test could be used to assess compliance with a permit limit for WET which are usually expressed in terms of acute or chronic toxic units or TUs. Commonly, numeric WET criteria or interpretations of narrative criteria are expressed as the equivalent of 0.3 toxicity unit ($0.3\ TU_a$) for

acute toxicity or one toxicity unit (1.0 TU_c) for chronic toxicity. To apply acute and chronic WET criteria in NPDES permits, a regulatory authority develops "wasteload allocations" (WLAs) which represent the pollutant discharge or toxicity of individual point sources that may be allowed while still attaining water quality standards for the receiving water. WLAs calculated for individual discharges use the applicable water quality criteria and account for dilution as allowed by the applicable water quality standards. Once WLAs are determined, the permitting authority calculates long-term averages (LTAs) of pollutant concentrations or effluent toxicity from the WLAs, and then calculates effluent limits from the most limiting LTA. Further explanation of how WET testing is used for regulatory purposes is found in EPA's WET test methods (40 CFR Part 136, 2002 edition as well as well as some earlier editions; USEPA 2000d) and in Chapter 3 of EPA's TSD.

Review Questions:

To enhance the technical and scientific quality as well as the credibility of the draft TST document, EPA asks that the following questions listed below be addressed. Note that the review questions are concerned with a technical evaluation of this draft alternative statistical approach for WET data analysis within the scope of the NPDES permit program and therefore is to be evaluated within the context of the NPDES permit program and its regulations. Please note that the draft TST approach does not address improvements to point estimate (e.g., IC_{25}) techniques used to analyze WET; rather the draft TST approach was developed to address improvements only to the current hypothesis testing approach (NOEC) used to examine WET. Also, the draft TST approach does not examine or suggest enhancements to the statistical flowcharts within EPA's WET test methods manuals. This report is concerned only with the calculation of the WET test result endpoint using the draft TST framework. Finally, EPA is very concerned about the presentation (or meeting *EPA's Plain English* requirements) of the draft TST document to its users, the permitting authorities in the NPDES States and EPA Regions who will be implementing the approach. In answering each of the review questions below, we ask that the questions be answered within this stated scope and context.

1) **Document's Merit:** Evaluate the conceptual soundness of the draft TST document's recommendations and the data analysis on which it is based. Is the draft TST approach an improvement over the current accepted hypothesis testing approach used in the NPDES WET program? If so, why, and if not, why not?

2) **Document's Responsiveness:** Assess whether the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis.

3) **Document's Data Analysis Basis:** Assess whether the data supporting the recommendations and conclusions on the draft TST document are technically correct and defensible. The draft TST document attempts to evaluate existing data comprehensively, but: (1) for the purposes of standardizing comparisons, relies on data developed after 1995; (2) to be comprehensive, evaluates data developed using EPA WET test methods conducted under the current 2002 edition, as well as some earlier editions; and (3) to ensure that conclusions are based on appropriate data, censors some data points. The Agency's reasoning behind each of these aspects of the

evaluation is explained in the draft document and related references (i.e., data test acceptance and quality assurance protocol).

4) **Document Conclusions**: Assess whether the draft TST approach as applied is technically defensible especially if challenged by either the NPDES regulated community, permitting authorities or expert consultants hired by permittees or other interested parties. Specifically, bioequivalency "*b*" values were derived for each test method using several risk management decision criteria which together, were intended to balance desired maximum alpha and beta errors at specific mean effect levels and within-test variability. Comment on the fact that this draft TST approach could be similarly used for additional WET test method(s) in the future. This draft TST approach builds upon EPA's earlier peer reviewed NPDES WET Variability document (USEPA 2000e) to derive and evaluate the "*b*" values. Evaluate the methodology used in the draft TST document to derive method-specific "*b*" values and apply the draft TST approach.

5) **Document Quality Overall:** Provide any recommendations for how this draft TST document should be presented to the public (or the users of this approach) particularly NPDES regulatory authorities such as NPDES States and EPA Regions (the document will be revised to accommodate readers with a more *Plain English* version). Suggest, if possible, how it's highly technical content should be translated into a version more readily understood by the NPDES regulatory public (again meeting EPA's *Plain English* requirements) and yet maintain its clarity given its potential scientific, regulatory, and technical applications. Also critique whether a regulatory authority and their permittees would clearly understand the draft TST document's recommendations and if not how specifically should it be revised to make it easier to implement under EPA's NPDES permit's program.

6) **Recommendations:** Provide any recommendations to improve the draft TST document's technical basis and approach for deriving the alternative WET statistical analysis method in the NPDES permitting program.

Schedule:

The external peer review should begin by or before August 11, 2008. EPA requests that peer review comments, suggestions and recommendations be submitted to the Avanti Corporation (External Peer Review contractor) no later than October 6, 2008.

Key References:

Chapman GA, Anderson BS, Bailer AJ, Baird RB, Berger R, Burton DT, Denton DL, Goodfellow WL, Heber MA, McDonald LL, Norberg-King TJ, Ruffier PJ. 1996. Methods and appropriate endpoints. In: Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts. Eds. Grothe DR, Dickson KL, Reed-Judkins DK. Special SETAC Publication.

USEPA. 1991. *Technical Support Document for Water Quality-based Toxics Control.* EPA/505-2-90-001. Office of Water, Washington, DC.

USEPA. 1995. *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to West Coast Marine and Estuarine Organisms*. EPA/600/R-95-136. Office of Research and Development, Cincinnati, OH.

USEPA. 2002a. *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*, 5th edition. EPA/821/R-02/012. Office of Science and Technology.

USEPA. 2002b. *Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms*, 4th edition. EPA/821/R-02-13. Office of Science and Technology.

USEPA. 2002c. *Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms*, 3rd edition. EPA/821/R-02-14. Office of Science and Technology.

USEPA. 2000d. *Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing* (40 CFR Part 136). Office of Water, Office of Science and Technology. Washington, DC. EPA/821/B-00-004.

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

USEPA. 2006. Peer Review Handbook: *Science Policy Council Peer Review Handbook, 3rd Edition (June 2006).* EPA/100-B-06-002. Office of Science Policy and Office of Research and Development, Washington, DC.

Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity

Revised Report

Contract No. EP-C-05-046
Work Assignment 1-29

Prepared for

Laura Phillips
U.S. Environmental Protection Agency
Office of Wastewater Management
1201 Constitution Ave., NW
Mail Code 4203M
ICC Building – Room 7135
Washington, DC 20004

Prepared by

Tetra Tech, Inc.
400 Red Brook Blvd, Suite 200
Owings Mills, MD 21117

May 20, 2008

NOTICE AND DISCLAIMER

This document provides guidance to NPDES regulatory authorities and persons interested in analyzing whole effluent toxicity (WET) test data using hypothesis testing approach as part of the NPDES program under the Clean Water Act (CWA). This document describes what EPA believes is an improved alternative to current approaches for analyzing WET test data produced using EPA's 2002 WET test methods promulgated at 40 CFR Part 136 and is provided as additional national guidance to EPA Regions and States. The document does not, however, substitute for the CWA, an NPDES permit, or EPA or State regulations applicable to permits or whole effluent toxicity testing; nor is this document a permit or a regulation itself. The document does not and cannot impose any legally binding requirements on EPA, States, NPDES permittees, and/or laboratories conducting or using whole effluent toxicity testing for permittees (or for States in the evaluation of ambient water quality). EPA and State officials retain discretion to adopt approaches on a case-by-case basis that differ from this guidance based on site-specific circumstances to the extent that such approaches are consistent with EPA federal NPDES regulations and the CWA. This guidance may be revised 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.

Executive Summary

Point estimate (e.g., IC_{25}) and hypothesis test (e.g., NOEC) approaches, under the Environmental Protection Agency's (EPA) National Pollutant Discharge Elimination System (NPDES) program, offer different advantages in terms of analyzing and interpreting whole effluent toxicity (WET) test data. The Test of Significant Toxicity (TST; otherwise referred to in literature as the test of bioequivalence) is an attractive alternative because it potentially incorporates the advantages of hypothesis testing and the transparency of the risk management level (i.e., 25% or 20% effect) while providing a positive incentive to generate high quality WET data. EPA has evaluated the feasibility of using TST to analyze routine WET data and to determine if TST is as protective as current recommended EPA approaches. TST tests whether the organism response in a given effluent concentration is less than a pre-defined proportion (termed *b*) of organism response in the control. The choice of *b* is a risk management decision, and therefore , the particular *b* value used, and the bioequivalence approach in general, is influenced by the error rates one is willing to tolerate. Thus, the objective of this project was to define an appropriate *b* value for several common WET test methods, and their associated error rates, such that the bioequivalence approach provides a level of protection comparable to or better than the current hypothesis approach without being unduly burdensome. This objective was accomplished by selecting a *b* value for each WET test method such that: (1) TST declares samples as toxic when the mean effluent effect is ≥ a prescribed percent effect level at an alpha level of 0.05 (5% Type I error rate) given average within-test variability observed over many tests (termed *sensitivity* in this report), and at the same time, (2) declares samples as non-toxic when the mean effluent effect ≤ a prescribed mean percent effect at a beta level of 0.20 (i.e. 20% Type II error rate) given average within-test variability for the test method and endpoint (termed *specificity* in this report).

Data from seven commonly required EPA WET test methods were analyzed using TST and the current EPA hypothesis testing approach (see Table E-1). Over 2000 WET tests from four different State programs were examined in this study, and rigorous data quality criteria were used to ensure that results represent current laboratory practices.

Table E-2 summarizes the risk management criteria used to identify test method-specific *b* values. For the West Coast WET methods examined the risk management goals were as follows: (a) declare effluents having a 20% mean effect in a test as toxic with an error rate ≤ 5% when within-test variability is above average for the method (> 50^{th} percentile control coefficient of variation [CV]); (b) tests having ≥ 25% mean effect should be declared toxic 100% of the time; (c) declare effluents having ≤ 10% mean effect in a test as non-toxic with a β error rate ≤ 20% under average within-test variability (≤ 50^{th} percentile CV); (d) tests having 15-20% mean effect and > 50^{th} percentile CV should be declared toxic with increasing frequency (i.e., higher beta error) as within test CV increases. Both simulation analyses (Table E-3) and analysis of actual WET test data (Table E-4) indicated that a *b* value of 0.80 satisfied the risk management goals, including the desired sensitivity and specificity, for the different West Coast WET methods examined in this project.

For the East Coast WET methods examined, the risk management goals were as follows: (a) declare effluents having a 25% mean effect in a test as toxic with an α error rate = 5% when

within-test variability is above average (> 50th percentile control CV); (b) tests having ≥ 30% mean effect should be declared toxic 100% of the time; (c) declare effluents having ≤ 15% mean effect as non-toxic with a β error rate ≤ 20% under typical within-test variability (≤ 75th percentile CV); (d) tests having 15-20% mean effect and higher variability should be declared toxic with increasing frequency as within-test CV increases. Both simulation analyses (Table E-3) and analyses of actual WET test data (Table E-4) for these methods indicated that a *b* value between 0.65 and 0.70 satisfied these risk management goals for the freshwater chronic and East Coast mysid methods. In general these *b* values were similar to 1-90th percentile MSD values calculated in this project and similar to those previously reported by EPA for the East Coast methods (Denton and Norberg-King 1996; USEPA 2000a)

Both simulation and actual WET data analyses indicated that TST, using the *b* values derived in this project, was superior to the current NOEC approach. This was demonstrated by: (1) higher Type I error rates (lower specificity) using the NOEC approach (i.e., declaring a test toxic when in fact it is not) than TST when the mean percent effect was less than the risk management decision level and within-test variability was low (Figure E-1); and (2) higher Type II error rates (lower sensitivity) using the NOEC approach (i.e., declaring a test non-toxic when in fact it is toxic) when the mean percent effect exceeded the risk management decision level and test variability was high (Figure E-1). In fact, for some WET test endpoints (e.g., *Ceriodaphnia dubia* reproduction) the current NOEC approach had as much as a 44% Type II error rate in tests having as high as a 30% effect depending on within-test variability. TST, using the *b* values derived in this project, always declared a 30% mean effect as toxic, regardless of within-test variability. Data from the abalone WET test method is a West Coast example that demonstrates higher Type I error rates using the NOEC approach. At a mean percent effect level of 10%, the current NOEC approach declared 88% of the tests as toxic when there was average or lower than average within-test variability. Using the *b* value derived in this project, TST never declared a 10% mean effect as toxic unless the within-test variability exceeded the 85th percentile for this test method. Thus, TST provides greater protection than the current NOEC approach under typical test performance, when the risk management effect level is exceeded and TST reduces the incidence of non-problems (false positives) when the risk management effect level is not exceeded.

TST was also evaluated as an alternative for analyzing acute WET test data. Using *P. promelas* acute WET test data (Table E-1), a *b* value of 0.70 was derived in this project based on a toxicity risk management decision of 25% mean effect on survival and the same alpha and beta error rate goals used for other WET methods in this project. Similar to results for chronic endpoints, as within-test variability increases, or the mean percent effect exceeds the decision level (25%), the finding of toxicity approaches 100% using TST. Both simulation and actual WET analyses demonstrated that the current t-test approach has lower sensitivity at a 25-30% mean effect level and typical within-test variability, and lower specificity at a 15% mean effect level than TST (Tables E-3 and E-4).

TST also provided desired advantages for two-concentration test data, based on analyses of 665 ambient toxicity tests provided by California's Surface Water Ambient Monitoring Program (SWAMP). Appropriate *b* values calculated based on *Ceriodaphnia dubia* (freshwater invertebrate - water flea) and *Pimephales promelas* (freshwater vertebrate - fathead minnow)

chronic ambient toxicity test data were similar to those based on multi-concentration WET test data for these two species. TST again achieved high sensitivity and specificity, and low error rates than the current t-test approach (Figure E-2). These results demonstrate that TST is a useful approach for analyzing both multi-concentration and two-concentration test designs.

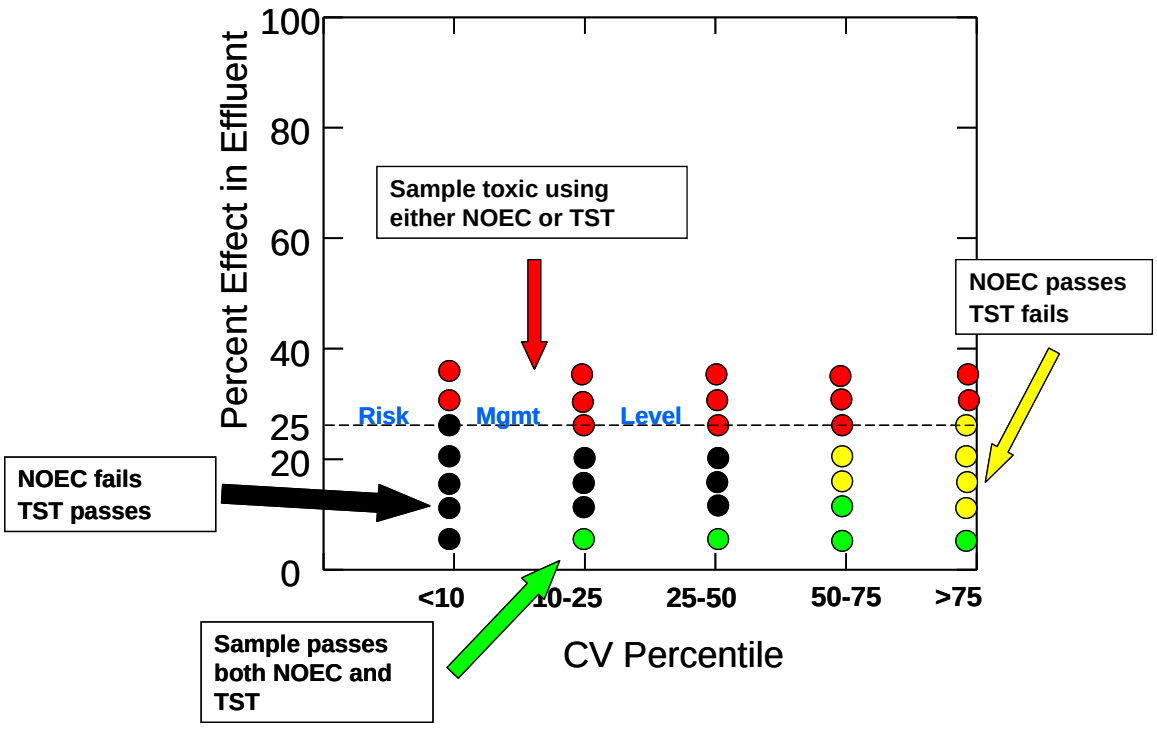


Figure E-1. Example using *Ceriodaphnia dubia* (freshwater water flea) EPA WET test results illustrating concordance observed between the current NOEC approach and the Test of Significant Toxicity (TST) as a function of coefficient of variation (CV) range observed within a test and percent effect observed in the effluent.

The alpha and beta error rates, and therefore, the *b* values derived in this project, are a function of the within-test variability commonly observed by many laboratories and many types of effluents. Given the important influence of lab performance on within-test variability and therefore, TST analyses, it is appropriate to periodically reevaluate the precision of WET test methods (especially the freshwater chronic and East Coast mysid methods) to determine whether the method specific *b* values should be modified to reflect improved method performance.

Results of this project suggest that TST could have several benefits to EPA's NPDES WET program including:

- Incentives for permittees to provide to permitting authorities high quality WET data upon which to base both WET reasonable potential decisions as well as compliance decisions with NPDES permit requirements (e.g., WET limits).

- TST is more likely to rank a sample as toxic if WET test data have greater within-test variability or inconsistencies
- Incorporation of error rates into the decision process, increasing confidence in test results.

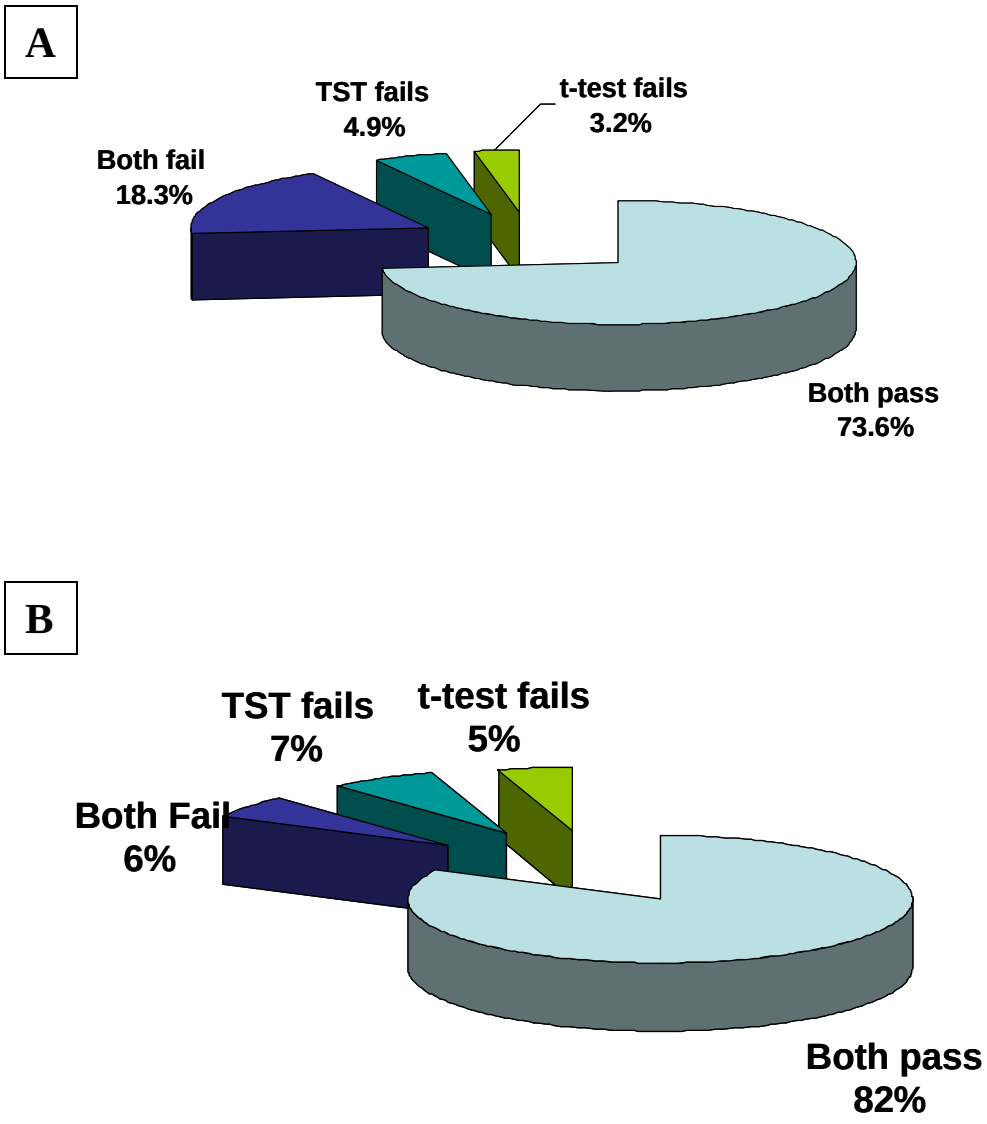


Figure E-2. Concordance between results of (A) *Ceriodaphnia dubia* (water flea) and (B) *Pimephales promelas (fathead minnow)* EPA chronic ambient toxicity tests using TST and a standard t-test analysis of the data. *b* value = 0.68 for *Ceriodaphnia* and 0.70 for *Pimephales* TST analyses.

Table E-1. Summary of WET Test Data Analyzed

EPA WET Test Method	Number of Tests		Number of Laboratories	Number of Dischargers
	Effluent	Ref Tox		
Ceriodaphnia dubia (water flea) survival and reproduction[a]	554	238	44	68
Ceriodaphnia dubia ambient chronic tests (California SWAMP program)	409	--	--	--
Pimephales promelas (fathead minnow) acute Survival[b]	347	0	15	101
Pimephales promelas (fathead minnow) survival and growth[b]	275	197	28	50
Pimephales promelas ambient chronic tests (California SWAMP program)	256	--	--	--
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) germ-tube length and germination[d]	0	135	11	--
Haliotis rufescans (red abalone) larval development[c]	0	136	10	--

[a] FIV – Freshwater invertebrate
[b] FV – Freshwater vertebrate
[c] SIV – Saltwater invertebrate
[d] SA – Saltwater algae

Table E-2. Summary of risk management decision criteria used to identify an appropriate *b* value for each WET test method examined in this project.

East Coast WET Methods (*C. dubia*, *P. promelas*, and *A. bahia* chronic test methods)		
Criterion	**Non-toxic (Bioequivalent) Effect Level**	**Within-test Variability Level (percentile of control CV)**
β error $\leq$ 0.20	15%	$\leq$ 75th
α error $\leq$ 0.05	25%	> 50th
α error = 0.00	30%	All levels
West Coast WET Methods (Sea Urchin fertilization, *H. rufens*, *M. pyrifera* chronic test methods)		
β error $\leq$ 0.20	10%	$\leq$ 50th
α error $\leq$ 0.05	20%	> 50th
α error = 0.00	25%	All levels
***P. promelas* Acute WET Method**		
β error $\leq$ 0.20	15%	$\leq$ 90th
α error $\leq$ 0.05	30%	> 90th
α error = 0.00	40%	All levels

Table E-3. Summary of hypothesis rejection rates using Test of Significant Toxicity and the current NOEC approaches for seven EPA WET test methods and Monte Carlo simulation analyses.

EPA WET Test Method	Number of Tests	Current NOEC Approach		Draft TST	
Chronic Freshwater and East Coast Mysid Chronic Methods		**Fraction deemed toxic at a 15% mean effect level[1]**	**Fraction deemed non-toxic at a 25% mean effect level[2]**	**Fraction deemed toxic at a 15% mean effect level[1]**	**Fraction deemed non-toxic at a 25% mean effect level[2]**
Ceriodaphnia dubia (water flea) 7-d survival and reproduction	792	0.89	0.20	0.05	0.00
Pimephales promelas (fathead minnow) 7-d survival and growth	472	0.78	0.40	0.14	0.00
Americamysis bahia (mysid shrimp) 7-d survival and growth	210	0.87	0.39	0.04	0.00
West Coast Marine Methods		**Fraction deemed toxic at a 10% mean effect level[3]**	**Fraction deemed non-toxic at a 20% mean effect level[4]**	**Fraction deemed toxic at a 10% mean effect level[3]**	**Fraction deemed non-toxic at a 20% mean effect level[4]**
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	177	0.85	0.76	0.19	0.00
Haliotis rufescans (red abalone) larval development	136	0.90	0.31	0.16	0.00
Macrocystis pyrifera (giant kelp) germination	135	1.00	0.00	0.00	0.00
Macrocystis pyrifera (giant kelp) germ-tube length	135	0.92	0.07	0.04	0.00
Pimephales promelas (fathead minnow) acute		**Fraction deemed toxic at a 15% mean effect level[5]**	**Fraction deemed non-toxic at a 30% mean effect level[6]**	**Fraction deemed toxic at a 15% mean effect level[5]**	**Fraction deemed non-toxic at a 30% mean effect level[6]**
Pimephales promelas (fathead minnow) acute survival	347	0.58	0.02	0.07	0.21

[1] For tests having $\leq 75^{th}$ percentile within-test CV
[2] For tests having $> 50^{th}$ percentile within-test CV
[3] For tests having $\leq 50^{th}$ percentile within-test CV
[4] For tests having $> 50^{th}$ percentile within-test CV
[5] For tests having $\leq 90^{th}$ percentile within-test CV
[6] For tests having $\geq 90^{th}$ percentile within-test CV

Table E-4. Summary of test performance characteristics using Test of Significant Toxicity analyses for seven EPA WET test methods examined. *b* values and performance characteristics using actual WET data are based on multi-concentration tests, consistent with current EPA WET test method protocols.

EPA WET Test Method	**Number of Tests**	**Recommended *b* value**		
Chronic Freshwater and East Coast Mysid Chronic Methods			**Relative Specificity [1] %**	**Relative Sensitivity[2] %**
Ceriodaphnia dubia (water flea) 7-d survival and reproduction	792	0.68	92	99
Pimephales promelas (fathead minnow) 7-d survival and growth	472	0.70	92	100
Americamysis bahia (mysid shrimp) 7-d survival and growth	210	0.70	96	99
West Coast Marine Methods			**Relative Specificity [3] %**	**Relative Sensitivity[4] %**
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	177	0.80	82	100
Haliotis rufescans (red abalone) larval development	136	0.80	83	100
Macrocystis pyrifera (giant kelp) germination	135	0.80	80	100
germ-tube length	347	0.80	80	100
Pimephales promelas (fathead minnow) acute			**Relative Specificity [1] %**	**Relative Sensitivity[2] %**
Pimephales promelas (fathead minnow) acute survival	347	0.70	90	100

1 Percentage of actual WET tests in which TST declared the sample as non-toxic (i.e., bioequivalent) when ≤ 25% difference in mean response was observed between the control and the effluent.
2 Percentage of actual WET tests in which TST declared the sample as toxic (i.e., not bioequivalent) when > 25% difference in mean response was observed between the control and the effluent.
3 Percentage of actual WET tests in which TST declared the sample as non-toxic (i.e., bioequivalent) when ≤ 20% difference in mean response was observed between the control and the effluent.
4 Percentage of actual WET tests in which TST declared the sample as toxic (i.e., not bioequivalent) when > 25% difference in mean response was observed between the control and the effluent.

List of Acronyms

ANOVA	Analysis of Variance
CETIS®	Comprehensive Environmental Toxicity Information System
CV	Coefficient of Variation
EPA	Environmental Protection Agency
FDA	Food & Drug Administration
IC_{25}	25% Inhibition Concentration
IWC	Instream Waste Concentration
LC_{50}	50% Lethal Concentration
LOEC	Lowest Observed Effect Concentration
LSEC	Lowest Significant Effect Concentration
MSD	Minimum Significant Difference
NOEC	No Observed Effect Concentration
NPDES	National Pollutant Discharge Elimination System
NSEC	No Significant Effect Concentration
PMSD	Percent Minimum Significant Difference
QAPP	Quality Assurance Project Plan
QA/QC	Quality Assurance/Quality Control
SWAMP	Surface Water Ambient Monitoring Program (California)
TAC	Test Acceptability Criteria
TMDL	Total Maximum Daily Load
TSD	Technical Support Document
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 death) on a group of test organisms during a short-term exposure (e.g., 24, 48, or 96 hours). Acute toxicity is measured 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.

ANOVA is analysis of variance.

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. Chronic toxicity is defined in terms of Toxicity Units (TU) as TUc = 100/NOEC or TUc = 100/ECp or ICp.

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. It is also called the relative standard deviation (RSD). The CV can be used as a measure of precision within (within-laboratory) and between (between-laboratory) laboratories, or among replicates for each treatment concentration.

Confidence Interval is the numerical interval constructed around a point estimate of a population parameter.

Effect Concentration (EC) is a point estimate of the toxicant concentration that would cause an observable adverse effect (e.g., death, immobilization, or serious incapacitation) in a given percent of the test organisms, calculated from a continuous model (e.g., Probit Model). EC_{25} is a point estimate of the toxicant concentration that would cause an observable adverse effect in 25 percent of the test organisms.

Hypothesis Testing is a statistical technique (e.g., Dunnett's test) for determining whether a tested concentration is statistically different from the control. Endpoints determined from hypothesis testing are NOEC and LOEC. The two hypotheses commonly tested in WET are: **Null hypothesis (Ho)**: The effluent is not toxic. **Alternative hypothesis (Ha)**: 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). IC_{25} is a point estimate of the toxicant concentration that would cause a 25 percent reduction in a non-lethal biological measurement.

Instream 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).

LC_{50} (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 adverse effects on the test organisms (i.e., where the values for the observed endpoints are statistically different from the control).

Lowest Significant Effect Concentration (LSEC) is the lowest tested concentration of an effluent or toxicant that causes significant adverse effect on the test organisms using the Test of Significant Toxicity (i.e., the lowest concentration of toxicant at which the values for the observed responses are not bioequivalent to the controls). ***** (New term and acronym for purposes of this project)**

Minimum Significant Difference (MSD) is the magnitude of difference from control where the 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 controls).

No Significant Effect Concentration (NSEC) is the highest tested concentration of an effluent or toxicant that causes no significant adverse effect on the test organisms using the Test of Significant Toxicity (i.e., the highest concentration of toxicant at which the values for the observed responses are bioequivalent to the controls). ***** (New term and acronym for purposes of this project)**

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 CWA.

Power is the probability of correctly detecting an actual toxic effect (i.e., declaring an effluent toxic when, in fact, it is toxic).

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 based on a number of factors.

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.

Significant Difference is defined as 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).

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 100 percent effluent). 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) is the error of rejecting the null hypothesis that should have been accepted.

Type II Error (beta) is the error of accepting the null hypothesis 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.

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

Table of Contents

Table of Contents (Continued)

Appendices

1.0 Introduction

1.1 Summary of Current EPA-Recommended WET Analysis Approaches

Some freshwater and marine chronic WET testing analyses examine both lethal and sub-lethal responses of test organisms along an effluent dilution series (USEPA 1995; 2002a; 2002b). The principal response endpoints used in acute WET testing are the lethal concentration to 50% of the test organisms (LC_{50}) and significant difference from control (e.g., t-test). The response endpoints commonly used in chronic testing are the no observed effect concentration (NOEC), the lowest observed effect concentration (LOEC), and the 25% inhibition concentration (IC_{25}). The NOEC endpoint is a hypothesis test approach that identifies the maximum effluent concentration at which the effect is not significantly different from the control (using an appropriate statistical test). The t-test or NOEC approach answers the question: Does the effluent at a critical concentration show a statistically significant decrease in organism response as compared to the control? The effluent is deemed not toxic for WET if the NOEC concentration is greater than the permitted instream waste concentration (IWC). The LC_{50} or IC_{25}, by contrast, is a point-estimation approach. It identifies the concentration at which the effect response is either 50 or 25% below the control value, respectively, and interpolates the effluent concentration for this response. The LC_{50} or IC_{25} approach answers the question: At what effluent concentration is a 50 or 25% effect observed, respectively, and is the critical effluent concentration less than this value? For chronic WET tests, an effluent is deemed not toxic if the permitted IWC is less than the IC_{25} concentration. For acute WET testing, the critical effluent concentration is usually 100% effluent. For either hypothesis or point estimate statistical approaches, the control performance (and control response), as well as the effluent response, has an influence on the endpoint values.

1.2 Advantages and Disadvantages of Recommended Statistical Approaches

Many researchers have reported several advantages and disadvantages of the hypothesis and point estimate approaches currently recommended. The Pellston Workshop on WET (Grothe et al. 1996) discussed these and they are summarized in Table 1-1. While several enhancements of

WET statistical procedures have been incorporated by EPA (e.g., USEPA 1995; 2002a; 2002b), an alternative approach that brings together the strengths of each approach is desirable.

Table 1-1. Summary of advantages and disadvantages of current recommended statistical approaches for analyzing WET data (adapted from Pellston Workshop on WET, Grothe et al. 1996).

Hypothesis Testing (e.g., NOEC, t-test)		**Point Estimate (e.g., IC_{25}, LC_{50})**	
Advantages	**Disadvantages**	**Advantages**	**Disadvantages**
(1) Well-suited for comparing critical effluent concentration to control	(1) Results dependent on effluent concentrations tested	(1) Provides and uses data from all treatments	(1) Computationally intensive
(2) Computationally easy, with common statistical tools	(2) Does not explicitly evaluate statistical power	(2) Endpoint does not need to be one of the effluent concentrations tested	(2) Endpoint can be concentration-dependent
(3) False positive rate controlled	(3) No incentive for permittee to increase test precision	(3) Precision estimates provided	(3) Endpoint is model-fit dependent
	(4) Confounded by hormesis and other non-linear concentration response patterns	(4) Confidence limits provided	(4) Confidence intervals may be affected by choice of EC_p value desired
	(5) Often violate equal variance and normality requiring use of less powerful non-parametric tests	(5) Wide choice of models can be used to derive endpoint	(5) Level of effect (ECp) must be specified
	(6) No direct relationship between statistical significance and biological significance of results	(6) Can be applied to all types of data	(6) False negative rate not explicitly addressed
		(7) Specifies a maximum allowable effect level that is biologically relevant	(7) Linear interpolation method subject to inaccuracies depending on concentration-response and data "smoothing" manipulations.

1.3 Use of Percent Minimum Significant Difference

One enhancement that has been implemented in chronic WET testing to help address the limitations noted in Table 1-1 is the use of a percent minimum significant difference (PMSD) value as part of the test review process and a minimum PMSD value as a "threshold" in the interpretation of effluent toxicity in a given test (Denton and Norberg-King 1996; USEPA 2000a; 2002a;2002b).

To minimize the degree of within-test variability, USEPA (1995) included a PMSD criterion that must be achieved in the seven West Coast WET test methods. The PMSD is a measure of the within-test variability and represents the difference from the control response that can be detected statistically. The PMSD is expressed as a percentage of the control response (PMSD = minimum significant difference (MSD)/control mean × 100). PMSD values from multiple tests for a test method are used to determine the level of sensitivity that can be achieved by most of the tests and most labs. This approach helps to standardize the NOEC endpoint and provides an incentive for testing laboratories to control test variability. The incorporation of a maximum PMSD into traditional hypothesis testing (in which the null hypothesis is that there is no effect of the effluent) as part of the test acceptability criteria for several West Coast WET test methods (USEPA 1995) is recognized as a step forward to address this issue. However, it increases the complexity of statistical analyses and still lacks inclusion of known test power in the analysis.

In addition to a maximum allowed within-test variability (maximum PMSD), a minimum test sensitivity level can also be established for tests demonstrating high precision (small PMSD). A minimum PMSD can help address issues that may arise when WET test controls are precise; in these cases, an effluent may have a statistically significant effect that may not be biologically significant.

1.4 Test of Significant Toxicity (TST)

The endpoints recommended in EPA's Technical Support Document for Water Quality-Based Toxics Control (TSD; USEPA 1991) and current EPA recommended procedures for analyzing

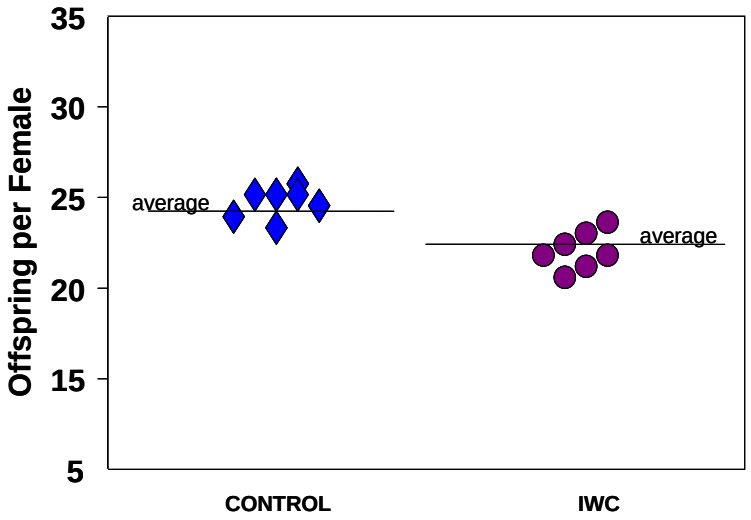
chronic WET data were not required to document statistical power of the test. Indeed, most statistical approaches used in environmental science are designed to minimize false positives (Type I or α error); i.e., concluding that the test treatment (e.g., effluent) is worse than the control when in fact it is not. False negatives (Type II or β error), concluding a sample is not toxic when it truly is, are not typically documented or known. Using current recommended procedures, if the replicate-to-replicate variability is high in the control, statistical power of the test is generally low and a truly toxic effluent may be incorrectly classified as not toxic (e.g., false negative; see Figure 1-1). Conversely, if the control replicate variability is very low (i.e., test is very precise), power may be high and an effluent may be considered toxic when in fact it is not (e.g., false positive; Figure 1-1; also see Table 1-2A for a summary of these relationships based on the traditional hypothesis testing approach). Unfortunately, this contributes to situations in which the discharger has no incentive to increase the precision of a test, and indeed may be penalized for achieving a high level of precision (see Figure 1-1). This project advances the TSD by making additional recommendations and demonstrating an alternative approach to WET test analysis that meets acceptable α and β error rates.

The Test of Significant Toxicity (TST) appears to be a viable alternative to the current EPA recommended hypothesis statistical method for analyzing WET test data. This statistical approach has been referred to as bioequivalence by other researchers (Erickson and McDonald 1995; Shukla et al. 2000) and is similar in principle to the test of non-inferiority, which has been used extensively in the medical field. The Food and Drug Administration (FDA) uses the test of non-inferiority to determine whether a new drug or therapy is at least as good as the existing treatment (Rogers et al. 1993; Anderson and Hauck 1983; Hatch 1996; Aras 2001; Streiner 2003). Using an approach similar to TST, the drug manufacturer has incentives to generate higher-quality (i.e., less variable) data upon which to base a decision.

In the context of the WET regulatory program, the bioequivalence test is structured to assess whether the effluent concentration of concern (e.g., instream waste concentration or IWC) and the control differ by a biologically significant amount. The null hypothesis in TST is that the effluent is significantly more toxic (i.e., results in lower organism response) compared to the control (Table 1-2B). Appropriate statistical tests are taken to reject the null hypothesis and

A **Very Small Intra-Test Variability**

NOEC approach: IWC different from control, *effluent is toxic**, Insignificant Difference Biologically***
TST approach: IWC equivalent to control, *effluent is non toxic*



B **High Intra-Test Variability**

NOEC approach: IWC same as control, *effluent is non toxic* ***, False Negative***
TST approach: IWC not bioequivalent to control, ***effluent is toxic***

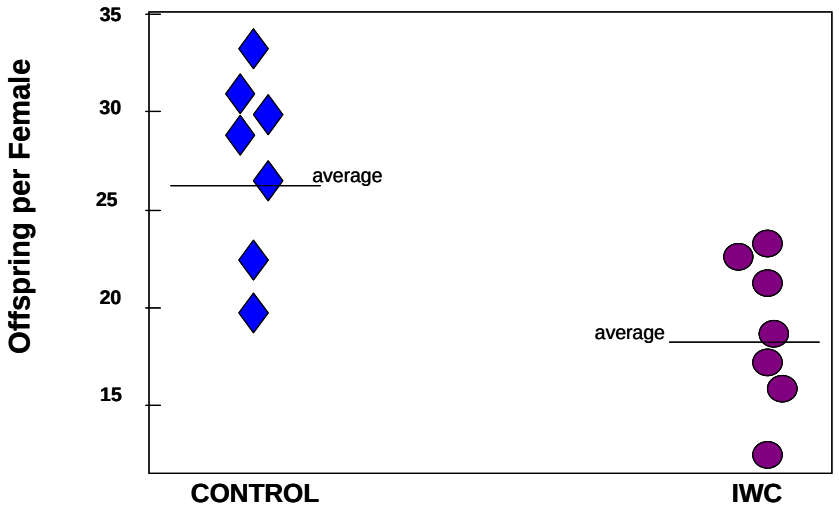


Figure 1-1. Two examples of WET test data demonstrating how the Test of Significant Toxicity (TST) can advance EPA's TSD (USEPA 1991) approaches for analyzing WET (USEPA 1995; 2002).

thereby accept the alternative hypothesis that the effluent is non-toxic. This approach shifts the responsibility on the discharger to demonstrate that the effluent is not toxic (i.e., effluent is bioequivalent to control). The "re-stated null hypothesis" used in the TST bioequivalence approach stands in contrast to the standard statistical test approach in which the null hypothesis is that the two concentrations (IWC compared to control) are not statistically significant. Use of the bioequivalence approach allows regulators to minimize the occurrence of false negatives (i.e., declaring an effluent safe when it is actually toxic), improving sensitivity of the test, and to minimize the occurrence of false positives (i.e., declaring an effluent toxic when it is actually not toxic) improving specificity of the test. Fairweather (1991) stated: the commitment of time, energy and people to a false positive will only continue until the mistake is discovered. In contrast, the cost of a false negative might have both short and long-term costs (e.g., ensuing environmental degradation). The TST bioequivalence approach also has the added advantage of providing dischargers with a clear incentive to improve the precision of test results (e.g., decrease within-test variability and/or increasing replication) to reach a definitive conclusion as to whether significant toxicity is observed in a test or not.

Table 1-2A. Relationships between false positive and false negative rates and resulting decisions based on the traditional hypothesis testing approach. From Denton and Norberg-King (1996).

Decision	**True condition**	
	Treatment = Control	**Treatment > Control**
Treatment = Control	Correct decision $(1 - \alpha)$	False negative Type II error (β)
Treatment > Control	False positive Type I error (α)	Correct decision $(1 - \beta)$ (power)

Entries correspond to the probability decision given in parentheses. Alpha, α, represents the probability of a Type I statistical error (i.e., false positive) and beta, β, is the probability of making a Type II statistical error (i.e., false negative).

Table 1-2B. Relationships between false positive and false negative rates and resulting decisions based on the Test of Significant Toxicity approach.

	True Condition	
Decision	Effluent ≤ b*Control (not bioequivalent)	Effluent > b*Control (bioequivalent)
Effluent is not bioequivalent (= toxic)	Correct decision ($1-\alpha$)	Type II error (β)[1]
Effluent is bioequivalent (= non-toxic)	Type I error (α)[2]	Correct decision ($1-\beta$) (power)

[1]Type II is the error of accepting the null hypothesis that should have been rejected.
[2]Type I is the error of rejecting the null hypothesis that should have been accepted.

Current EPA WET analysis approaches address specificity but not sensitivity because the latter depends on statistical power, which is not explicitly controlled in current WET analysis methods. However, this is not the case using TST because statistical power is incorporated into the TST approach by re-stating the hypothesis as shown in Figure 1-2 and by incorporating the bioequivalent parameter *b* that ensures maximum desired α and β rates. Thus, TST is more protective than current EPA methods for those tests where greater sensitivity is needed (i.e., highly variable within-test data), and it is less sensitive to minor differences between control and effluent that are statistically but probably not biologically significant.

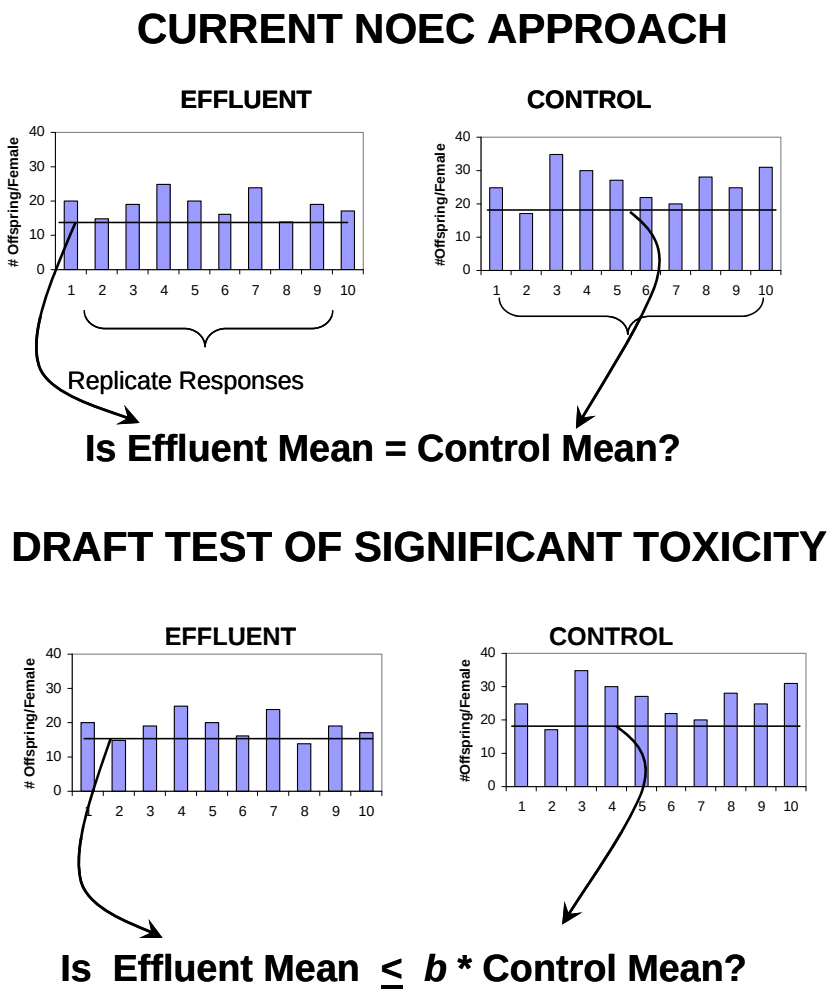


Figure 1-2. Pictorial depiction comparing the current recommended hypothesis testing approach and the Test of Significant Toxicity (TST) approach for evaluating WET data.

1.5 Project Objectives

The primary purpose of this project was to evaluate whether TST is a useful alternative data analysis approach for WET data, in addition to the approaches currently recommended in EPA's TSD and test method manuals. The principal objectives addressed were:

1. Identify an appropriate bioequivalent factor *b* for several common WET test methods based on desired α and β error rates at pre-specified risk management effect levels and normal test performance.
2. Determine the degree of protectiveness of TST compared to the current hypothesis testing approach. In this report, "as protective as" means equal ability to declare a sample toxic when toxicity is present and not declare a sample toxic if less than the desired regulatory risk management level.

In this project, emphasis was placed on comparing results of TST to current hypothesis testing approaches and not to linear interpolation (i.e., IC_{25}). TST, like any hypothesis test approach, is not a point estimate technique and therefore can not be directly compared to the IC_{25} approach. A distinction to be made regarding the utility of either a point estimate vs hypothesis testing approaches is that hypothesis testing can utilize either multi-concentration or two-concentration (IWC vs control) testing design. A two-concentration design can be advantageous for situations in which the critical effluent concentration is at or near 100% effluent, and conducting stormwater and watershed (i.e., ambient) toxicity testing. To address the above objectives, EPA first analyzed a limited set of *Ceriodaphnia dubia* chronic WET data and then conducted a more comprehensive analysis on a larger number of WET tests and several commonly required chronic freshwater and marine EPA WET test methods, as well as one commonly used acute freshwater vertebrate WET test method with both multi-concentration and two-concentration designs. The focus of these analyses was on chronic WET test methods and sublethal endpoints because many different types of alternative analysis procedures have been proposed for these tests. This document provides a summary of the results of this comprehensive analysis.

2.0 Data Characteristics

2.1 Test Methods and Endpoints Evaluated

Table 2-1 summarizes the seven EPA WET test methods evaluated under this project. Preference was given to WET data generated using the EPA 1995 WET test methods for the EPA West Coast marine species and for all other species the 2002 EPA WET test methods was the basis for data selected (USEPA 2002a, 2002b). However, a few of the effluent and reference toxicant tests were not based on the 2002 methods. Examination of the inter-lab reference toxicant data for *C. dubia* by year indicated significantly more precise data from 1996 on as compared to pre-1995 (Figure 2-1). This result is not unexpected because the EPA freshwater WET test methods were substantially refined as of 1995 and laboratories had more experience with the chronic test methods by this time. Therefore, only post-1995 data were used in this project for all EPA WET test methods. These post-1995 data are likely to be consistent with data generated using the 2002 methods because the test methods examined in this project were not substantially revised between 1995 and 2002.

All seven WET test methods listed in Table 2-1 are commonly used by regulatory authorities in making regulatory decisions such as determining WET reasonable potential or to determine compliance with acute and chronic WET limits or monitoring triggers. These seven 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 analytical laboratories. Therefore, results obtained in this project, using these seven EPA test methods, should be applicable to other EPA WET methods not examined. For example, results of analyses for the freshwater fathead minnow larval survival and growth test, one of the EPA test methods examined in this project, can be extrapolated to other EPA fish survival and growth tests (e.g., *Menidia sp,*, *Atherinops affinis*, *Cyprinus variegatus*) because all of these EPA test methods use a similar test design (e.g., number of replicates, number of organisms tested) and measure the same endpoints. Denton and Norberg-King (1996) reported similar statistical properties of test endpoints for various EPA WET tests using the fish test design. Similarly, the freshwater fish acute WET test analyzed in this project can be extrapolated to other acute test methods because they use a similar test design and measure

mortality. The use of both EPA saltwater and freshwater WET tests ensured that there was adequate representation of different types of discharge situations and laboratories.

For each set of test data received, additional metadata information were required including:

- Discharger 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

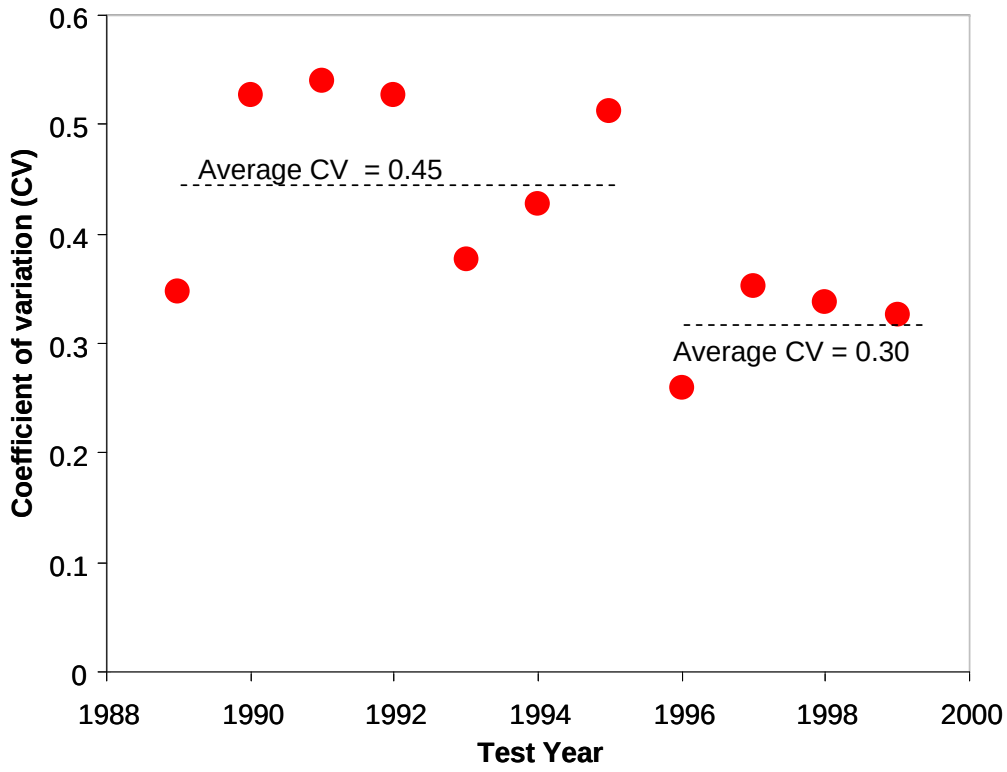


Figure 2-1. Summary of the test variability (expressed as the 90th percentile coefficient of variation or CV) observed between 1989 and 2000 for the *Ceriodaphnia dubia* (freshwater water flea) chronic EPA WET test. This figure illustrates and supports the basis for utilizing test data post 1995, as test precision improved from an average CV = 0.45 to an average CV = 0.30.

Table 2-1. Summary of test condition requirements and test acceptance criteria for each EPA WET method evaluated in Test of Significant Toxicity 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)
1000.0	Fathead Minnow *(Pimephales promelas)*	survival	Acute	10	2	5	48-96h	≥ 90% survival in controls
	Fathead Minnow *(Pimephales promelas)*	survival and growth	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 *(Mysidopsis bahia)*	survival, 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 eggs	4	4	40 min (20 min plus 20 min)	≥ 70% egg fertilization in controls; and appropriate sperm counts; %MSD <25%;
1018.0	Giant kelp germination *(Macrocystis pyrifera*)	germination and germ-tube length	Chronic	100 spores for germination and 10 spores for length	5	4	48 hrs	≥ 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 rufescans*)	larval development	Chronic	100 larvae	5	4	48 hrs	≥ 80% normal larval development in controls must have a statistica; significant effect at 56 μg/L zinc and a MSD < 20%

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. These data were instrumental in establishing *b* thresholds for TST analyses. 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 2000a). 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, 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' (DNR) ambient monitoring program). Laboratory and test method metadata required were the same as those noted above for effluent toxicity data.

2.2 Data Sources

EPA received WET data from multiple sources including: Washington State Department of Ecology, EPA Headquarters' Office of Science & 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 data desired were identified and documented in the Data Management Plan (Appendix A) and the QAPP for this project. Nearly 2,500 WET tests of interest were incorporated, representing many dischargers and laboratories (Table 2-2). Data were received in a variety of formats and compiled by test type in the database program CETIS® (Tidepool Software, v. 1.0). The CETIS program is designed to analyze, store, and manage WET data.

Table 2-2. Summary of WET test data analyzed.

EPA WET Test Method	Number of Tests		Number of Laboratories	Number of Dischargers
	Effluent	Ref Tox		
Ceriodaphnia dubia (water flea) survival and reproduction[a]	554	238	44	68
Ceriodaphnia dubia ambient chronic tests (California SWAMP program)	409	--	--	--
Pimephales promelas (fathead minnow) acute survival[b]	347	0	15	101
Pimephales promelas (fathead minnow) survival and growth[b]	275	197	28	50
Pimephales promelas ambient chronic tests (California SWAMP program)	256	--	--	--
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) germ-tube length and germination[d]	0	135	11	--
Haliotis rufescans (red abalone) larval development[c]	0	136	10	--

[a] FIV – Freshwater invertebrate
[b] FV – Freshwater vertebrate
[c] SIV – Saltwater invertebrate
[d] SA – Saltwater algae

2.3 Representativeness of WET Data

The usefulness of the results obtained in this project depended on having representative WET 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 facility types, including both industrial and municipal dischargers
- Represent many facilities for a given EPA WET test method (i.e., no one facility dominates the data for a given test method)
- Cover a range of target (design) effluent dilutions upon which WET reasonable potential and compliance are based, ranging from 10% to 100% 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% to 100% effluent)

Attempts were made to ensure that no one laboratory or permittee had > 10% of the test data for a given test type. For any EPA WET test method, only the 20 most recently conducted tests were used if more than 20 tests were available for a given laboratory or facility. The summary information presented in Table 2-2 demonstrates that WET data were received from numerous laboratories and facilities for all EPA WET test methods analyzed under this project.

2.4 Data Processing

Processing of raw WET 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 WET test was assigned a unique code and each laboratory was also uniquely coded. A tracking system was used to help evaluate whether WET data were needed for certain types of EPA WET test methods and to help increase representativeness of laboratories or types of facilities for a given method.

WET data received in either ToxCalc® or CETIS® were imported directly into the CETIS® database dedicated to this project. WET data received in Excel or other spreadsheet format were also directly imported into CETIS®. In cases where the source organization had not yet entered their WET data electronically, they were supplied with a template so their data could be readily transferred to CETIS® to minimize transcription errors. Data in CETIS® were checked on 10% of the tests received from each source to document proper data transfer.

WET 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 next section and summarized in Table 2-1. 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.

2.5 TST Data Analyses

2.5.1 General Approach

For evaluation of WET compliance, a one-sided test of bioequivalence is used, in which the null hypothesis is H_O: $\mu_T \leq b\ \mu_C$, where μ_T is the treatment mean (i.e., the response of the effluent concentration of concern; e.g., IWC), μ_C is the control mean (i.e., the response of the control), and *b* is the ratio that measures whether or not a response (e.g., decreased growth, decreased reproduction) is biologically significant and thus defines bioequivalence. To statistically test the null hypothesis, a modified *t*-test is performed (e.g., see Erickson and MacDonald 1995) in which the value "*b*" is incorporated to account for the fact that one is not testing whether two means are equal but rather whether the means are within a specified level of change of each other.

There are three decisions to be made in setting up the bioequivalence *t* test:

1. The probability of Type I error (α), which, in the bioequivalence formulation, is the probability of declaring the test water "bioequivalent" to the control when in fact there is a significant effect (see Table 2-3);
2. The probability of Type II error (β), which is the probability of declaring that a significant effect exists when in fact it does not (see Table 2-3) ; and
3. The level of change (*b*) that is used in the t-test to define whether an effect is not bioequivalent.

All three decisions are in essence risk-management policy decisions. To meet Clean Water Act mandates to protect beneficial uses, Type I error should be controlled to a small value, typically specified as α = 5%. Both regulatory agencies and dischargers are interested in minimizing Type II errors (incorrectly declaring that an effect exists) because these errors result in unnecessary expenditure of resources on non-problems. Therefore, in the draft TST approach, a Type II (or β) error rate $\leq$ 20% was also included as a risk management goal, a commonly accepted value for power in statistical analyses (e.g., Fairweather 1991; Denton and Norberg-

King 1996; USEPA 2000a). The level of change (*b*) that defines bioequivalence using TST is a 25% effect for the chronic freshwater and East Coast mysid methods and 20% for the chronic West Coast marine methods. The 25% effect threshold is a risk management goal which is consistent with EPA's Technical support Document (USEPA 1991) and the promulgated WET methods (USEPA 2002a, USEPA 2002b) in the form of the IC_{25} (effluent concentration exhibiting a 25% effect as compared to the control). The slightly lower risk management criteria proposed for many of the West Coast methods (e.g., Echinoderm, giant kelp and red abalone) is based on an analysis of the approximate value of "p" for the West Coast methods and was found be below 25% (Denton et al 1994; Denton and Norberg-King 1996). These methods employ an experimental design comprised of four to five replicates and counting a hundred cells or organisms per replicate. Therefore, these tests have greater ability to detect a lower toxicity threshold (i.e., 20 % effect).

Table 2-3. Definition of α and β errors under Test of Significant Toxicity.

Given the null hypothesis is: H_O: $\mu_T \leq b\ \mu_C$,

where μ_T is the treatment mean (i.e., the response of the effluent concentration of concern), μ_C is the control mean, alpha and beta and Type I and II error rates are defined as:

	True Condition	
Decision	Effluent ≤ b*Control (not bioequivalent)	Effluent > b*Control (bioequivalent)
Effluent is not bioequivalent (= toxic)	Correct decision (1-α)	Type II error (β)[1]
Effluent is bioequivalent (= non-toxic)	Type I error (α)[2]	Correct decision (1-β) (power)

[1]Type II is the error of accepting the null hypothesis that should have been rejected.
[2]Type I is the error of rejecting the null hypothesis that should have been accepted.

Two general types of analyses were conducted to determine an appropriate *b* value for each test method. The first approach used Monte Carlo to simulate WET data, and examined many different test scenarios (mean percent effect of the effluent and within-test variability) to determine Type I and Type II error rates for TST using different *b* values. Error rates were also computed for the same simulated tests using the current NOEC approach. This analysis

established *b* values for each WET test method that achieved desired error rates given different test scenarios and provided a comparison with the current NOEC approach.

The second analysis analyzed Type I and Type II error rates for TST using actual WET data obtained in this project. This analysis was used to help ground truth the Monte Carlo results and to further support the selection of a *b* value for each WET test method and endpoint. The following provides further details on how each approach was conducted.

2.5.2 Simulation Analyses

For each WET test method and endpoint, Monte Carlo analyses were conducted using all combinations of several different *b* values, coefficient of variation (CV) ranges for the control and effluent treatments, and several different effect levels to characterize α and β errors under these different scenarios. This analysis helps to identify an appropriate *b* value that meets risk management goals. Values of *b* examined in simulation were based on the distribution of minimum significant difference (MSD) values observed using actual test data as well as those reported previously in USEPA (2000a). For each test method and endpoint (e.g., growth, survival) the simulated control mean was a randomly chosen value between the 10^{th} and 90^{th} percentile of observed control mean responses from actual WET tests for a given method. For chronic test endpoints, the simulated effluent mean was set between 65% and 95% of the respective control mean at 5% intervals (5%, 10%, 15%, 20%, 25%, 30%, and 35% effect). Effect levels higher than 35% were not examined for chronic endpoints in this analysis because previous work has shown that such large effects are typically declared as toxic 100% of the time using either the current NOEC approach or TST. For the *P. promelas* acute WET test, effect levels ranging between 15 and 40% were examined.

Due to the wide range and skewed distribution of effluent CVs observed in actual WET test data for a given chronic method, within-test variability in simulations was based first on control variability. Control CV was categorized based on the range of observed CV values for a given method and endpoint (0-10^{th}, 10-25^{th}, 25-50^{th}, 50-75^{th}, 75-85^{th}, and 85-95^{th} percentiles of observed control CV). Effluent CV for a given simulated test was then randomly chosen based

on the ratios of control to effluent CV observed in actual WET tests for a given method and endpoint. The range of ratios used corresponded to between the 25th and 75th percentiles of the ratio distribution observed. For most of the test methods and endpoints examined, the effluent CV value was between 0.5 and 2 times the control CV value. Thus, by simulating control CV, effluent variability was also simulated for each WET test method and endpoint.

For each simulated test, a mean percent effect (e.g., 25% effect) was applied, as well as one of the CV categories, and one of the several *b* values. Each combination of percent effect, CV, and *b* value was run 1000 times for each method and endpoint to determine Type I (α) and Type II (β) error rates for specified percent effect levels and within-test variability, and to determine the *b* value that yielded error rates consistent with the desired risk management goals.

Freshwater Chronic and East Coast Mysid WET Method Risk Management Criteria

Table 2-4 summarizes the risk management decisions used to identify an appropriate *b* value for each WET method examined in this study. For the freshwater chronic and East Coast mysid WET methods examined, the risk management toxicity threshold is 25% effect, consistent with the current IC_{25} approach which also uses a 25% effect threshold for determining compliance with Clean Water Act goals. Therefore, one criterion for selecting an appropriate *b* value for these methods was that TST, using a given *b* value, had an α error rate ≤ 0.05 (5%) when the mean effect level = 25% under above average within-test variability (> 50th percentile CV range). This criteria ensures that effluents having effects at or near the threshold level are not declared as toxic when test variability is relatively low or average and therefore the decision more certain. A second criterion for selecting an appropriate test-specific *b* value for these methods was that TST, using the same *b* value identified above, also has a β error rate ≤ 0.20 (20%) at an effect level considered not toxic (i.e., bioequivalent to the control). For the freshwater chronic and East Coast mysid methods, this effect level was 15% given typical normal within-test variability (≤ 75th percentile CV). This second criterion ensures that truly non-toxic effluents are declared as such most of the time unless data quality is relatively poor (i.e., relatively high within-test variability). The third decision criterion used to select the method-specific *b* value for freshwater chronic and East Coast mysid methods is that the *b* value

meeting the previous two criteria must also result in TST declaring all (100%) tests toxic at a mean percent effect level of 30%, regardless of within-test variability. This criterion ensures that a truly toxic effluent is always declared as such by TST given the *b* value identified.

Using the three risk management criteria above, it was understood that tests having a mean percent effect between 15 and 25% will tend to be declared toxic at a β error rate > 0.2, except when within-test variability is very low (e.g., CV within the 0-25th percentile for a given WET test method). This result is in keeping with the overall risk management goal of having TST be as or more protective than current WET analysis approaches when tests have high within-test variability, and therefore, uncertain results.

Table 2-4. Summary of risk management decision criteria used to identify an appropriate *b* value for each WET test method examined in this project.

Chronic Freshwater and East Coast mysid WET Methods (*C. dubia*, *P. promelas*, and *A. bahia* chronic test methods)		
Criterion	**Non-toxic (Bioequivalent) Effect Level**	**Within-test Variability Level (percentile of control CV)**
β error ≤ 0.20	15%	≤ 75th
α error ≤ 0.05	25%	> 50th
α error = 0.00	30%	All levels
West Coast Marine WET Methods (Sea Urchin fertilization, *H. rufens*, *M. pyrifera* chronic test methods)		
β error ≤ 0.20	10%	≤ 50th
α error ≤ 0.05	20%	> 50th
α error = 0.00	25%	All levels
***P. promelas* Acute WET Method**		
β error ≤ 0.20	15%	≤ 90th
α error ≤ 0.05	30%	≥ 90th
α error = 0.00	40%	All levels

West Coast Marine WET Method Risk Management Criteria

For the West Coast saltwater chronic WET methods examined, risk management levels regarding non-toxic and toxic effects are slightly lower than for the freshwater chronic and East Coast mysid WET methods as discussed above. The non-toxic level, the threshold level, and toxic level for these methods are 10, 20, and 25% mean effect, respectively (Table 2-4). Desired error rates for α and β were the same as those used for the freshwater chronic and East

Coast mysid methods (0.05 and 0.20, respectively). The error rate for α was evaluated at above-average within-test variability (≥ 50th percentile CV) while β error was evaluated at average within-test variability or lower (≤ 50th percentile CV). Similar to the methodology summarized above, the approach for West Coast marine WET methods resulted in a higher β error rate (i.e., TST would declare more effluents toxic) for mean effect levels between 10 and 20% when within-test variability is relatively high.

Student t-test was used to compare mean endpoints between control and effluent treatments. Exhibit A demonstrates the t-test using both the current hypothesis approach and the TST approach, using *C. dubia* reproduction data from two different examples. These examples also demonstrate the advantages of TST over the current hypothesis testing approach. In cases of unequal variance, Welch's adjusted t-test was used. The F-statistic was used to test for unequal variances and to determine whether a normal t-test or Welch's t-test was used for a given simulated test. For proportional data (e.g., survival, development, and germination data), arcsine square root transformation was used (as specified in the 2002 EPA WET methods) to address non-normally distributed data. CETIS® was used to generate current recommended NOEC values for each WET test as well to compare error rates with TST results.

For the *Pimephales promelas* acute toxicity test method, only two replicates per test (minimum required number of replicates in the test method) were typically available using actual WET data. In order to perform the statistical comparisons required for multiple concentrations, a minimum of four replicates is required. Therefore, mean and standard deviation estimates for this method were based on values from WET tests using 2 replicates but data were simulated using 4 replicates. For the *P. promelas* acute WET test, the range of CV values is much more constrained than for other WET methods examined in this project because: (1) the test acceptability criterion for this and other EPA acute WET test methods is ≥ 90% survival in the control, which limits the variability allowed in terms of control survival; and (2) acute WET tests use no more than four replicates, each having typically five organisms (USEPA 2002c). The combination of these two factors results in a relatively narrow range of mortality that can occur in each control replicate and therefore, a narrow range for the control CV overall in these methods. Control CVs computed from *P. promelas* acute WET tests were all < 20% and nearly

90% of the tests had a CV = 0%. Decision criteria used to determine the *b* value for the *P. promelas* acute WET test method included: (a) β error ≤ 20% when mean effect = 15% at a CV ≤ 90th percentile; and (b) α error ≤ 5% when mean effect = 30% at a CV ≥ 90th percentile (Table 2-4). At a mean effect level ≥ 40%, the α error = 0.00%, regardless of within-test variability. The decision to identify mean effect levels of 30-40% as toxic most or all of the time is more conservative than the current LC_{50} approach, in which at least 50% mortality may result in a decision of no acute toxicity.

2.5.3 Analyses of Actual WET Data

In addition to simulation analysis, actual WET data were analyzed to determine whether a given test was declared toxic or not based on exceeding a specified mean percent effect level, as a function of several *b* values. For each test method and endpoint, mean percent effect ranges of 10-15%, 15-20%, 20-25%, and 25-30% were used as thresholds for determining whether a given sample was declared toxic. Values for *b* bracketed 1-90th percentile MSD for a given test method and endpoint and also included several *b* values used in simulation analyses. For each method, the fraction of WET tests declared as toxic using TST was calculated as a function of those tests having either > or ≤ the respective risk management toxicity threshold (i.e., 25% for freshwater chronic and East Coast mysid chronic methods and *P. promelas* acute methods, and 20% for West Coast marine methods) as well as at lower effect levels (see Figure 2-2 for an example). Those tests declared as toxic by TST, but having a mean percent effect ≤ the toxicity threshold, were used to calculate the β or Type II error rate (referred to as specificity). Those tests declared as non-toxic but exceeding the toxicity threshold were used to calculate the α or Type I error rate (referred to as sensitivity).

It is understood that the above analysis was dependent on the test data available and was therefore not as robust statistically as the simulation analyses. However, for several test methods, the number of tests was fairly large (> 200 tests) and therefore, this analysis provides some ground truthing of the simulation results.

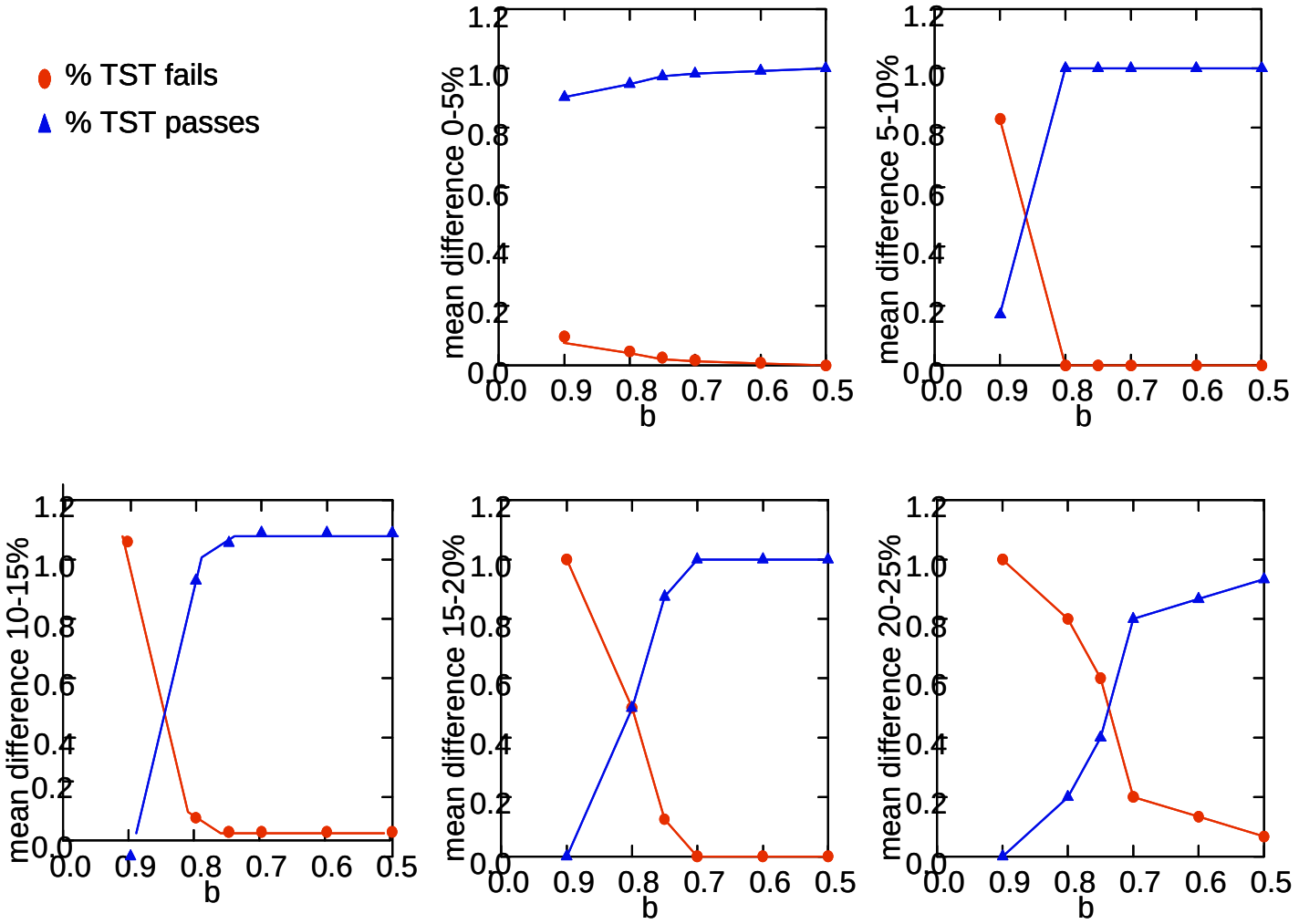
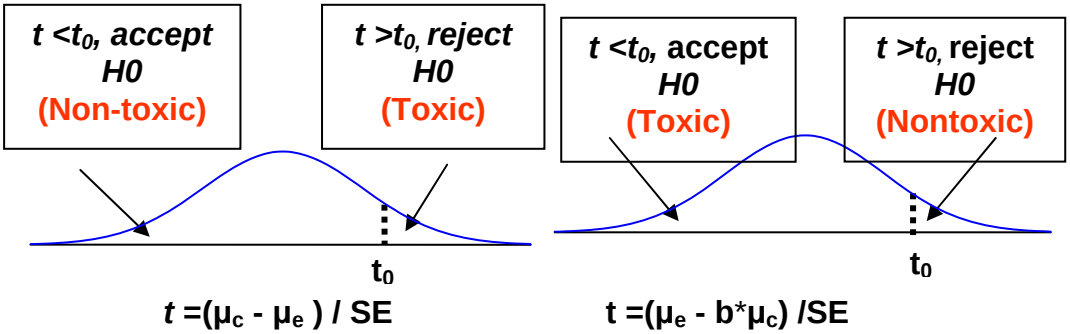


Figure 2-2. Example of detailed TST analysis from the red abalone larval development WET test showing the fraction of tests passed or failed (i.e., either bioequivalent, or not bioequivalent, respectively, to the control) as a function of *b* value.

Calculation of TST vs Current Hypothesis Testing Approach

Both the standard NOEC approach and Test of Significant Toxicity (TST) approach involve t statistics. Their calculations are similar. Once a t-value is calculated, whether a test "is declared toxic" or "declared non toxic" depends on whether one uses the standard hypothesis testing approach or the TST approach.



Standard hypothesis testing NOEC (Left); TST approach (Right)

The following shows calculations using both the standard hypothesis testing and the TST approaches for two different examples based on *Ceriodaphnia* chronic test reproduction data: one where there is relatively high within-test variability and one where there is relatively low within-test variability. These examples are used to illustrate the calculations used for each approach, and to show how the outcome compares using these two approaches, given two extremes in test variability.

EXAMPLE 1: TEST WITH HIGH WITHIN-TEST VARIABILITY.

Treatment	N	Effluent Concentration	Mean	St. Dev	Coefficient Variation
Control	**10**	**0%**	**33.6**	**10.06**	**30%**
IWC	**10**	**13.5%**	**27.0**	**10.83**	**39%**

Step 1 for both approaches: Compute Variance:
$Sp^2 = (S_c^2 + S_e^2)/2 = [(SDEV_c)^2 + (SDEV_e)^2]/2 = [(10.06)^2 + (10.83)^2]/2 = 109.2$
Where: Sp is the total variance; S_c is control variance; S_e is variance in effluent treatment; $SDEV_c$ is the standard deviation of the control data; $SDEV_e$ is the standard deviation of the effluent data

Standard NOEC Approach (*Null hypothesis: Control mean = Effluent mean*)

Step 2: Compute Standard error of the mean
Stand. Error = square root [Sp^2 * ($1/N_c$+$1/N_e$)]
= square root [109.2* (0.1 + 0.1] = 4.673
Where: N_c is the number of control replicates; N_e is the number of treatment (effluent) replicates

Step 3: Compute t-value
t= (μ_c - μ_e) / Stand. Error = (33.6-27.0)/4.67= **1.41**
where: $\boldsymbol{\mu_e}$ is the mean reproduction at IWC, $\boldsymbol{\mu_c}$ is mean control reproduction.

The table (critical) value t (α<0.05) = 1.73

Step 4: Compare computed t value with table t-value
1.41 < 1.73. Therefore, ***accept null hypothesis***: effluent is equal to control
Effluent is declared not toxic

TST Approach
Null hypothesis: Effluent mean ≤ b* Control mean

Step 1: Same as for NOEC approach above.
Step 2: Compute Standard error of the mean
Stand. Error = square root [Sp^2 * (1/ N_c + b^2/ N_e)]
= square root [109.2* (0.1 + 0.68^2/10)] = 3.99
where *b* is defined based on α and β error rates established as risk management criteria. For this test method, *b* = 0.68, which is also similar to 1- the 90^{th} percentile MSD for this test method.

Step 3: Compute t-value
t = (μ_e – *b*μ_c) / Stand. Error = (27.0 – 0.68* 33.6)/ 3.99 = **1.04**
The table (critical) value t (α < 0.05) = 1.73

Step 4: Compare computed t value with table t-value
1.04 < 1.73. Therefore, ***accept null hypothesis***: difference between control and effluent > 25%
Effluent is declared toxic.

EXAMPLE 2: TEST WITH LOW WITHIN-TEST VARIABILITY

Treatment	N	Effluent Concentration	Mean	St. Dev	Coefficient Variation
Control	**10**	**0%**	**26.5**	**2.12**	**8%**
IWC	**10**	**13.5%**	**23.6**	**2.99**	**13%**

Step 1: Compute Variance:
$Sp^2 = (Sc^2 + Se^2)/2 = [(SDEVc)^2 + (SDEVe)^2]/2 = [(2.12)^2 + (2.99)^2]/2 = 6.72$

Standard NOEC Approach (*Null hypothesis: Control mean = Effluent mean*)

Step 2: Compute Standard error of the mean
Stand. Error = square root $[Sp^2 * (1/N_c + 1/N_e)]$
= square root $[6.72 * (0.1 + 0.1] = 1.16$

Step 3: Compute t-value
$t = (\mu_c - \mu_e) /$ Stand. Error = (26.5-23.6)/1.16= **2.50**
where: $\boldsymbol{\mu_e}$ is the mean reproduction at IWC, $\boldsymbol{\mu_c}$ is mean control reproduction.
The table (critical) value t, $\alpha<0.05$ = **1.73**

Step 4: Compare computed t value with table t-value
2.50 > 1.73; Therefore, ***reject null hypothesis***: effluent is NOT equal to control;
Effluent is declared toxic

TST Approach
Null hypothesis: Effluent mean ≤ b* Control mean

Step 1: Same as for standard NOEC approach above
Step 2: Compute Standard error of the mean
Stand. Error = square root $[Sp^2 * (1/N_c + b^2/N_e)]$
= square root $[6.72 * (0.1 + 0.68^2/10)] = 0.991$

where *b* is defined based on α and β error rates established as risk management criteria. For this test method, *b* = 0.68, which is also similar to 1- the 90th percentile MSD for this test method.

Step 3: Compute t-value
$t = (\mu_e - b^*\mu_c) /$ Stand. Error = (23.6 – 0.68* 26.5)/ 0.991= **5.63**

Step 4: Compare computed t value with table t-value
5.63 >1.73. Therefore ***reject null hypothesis***: difference between control and effluent < 25%
Effluent is declared not toxic.

3.0 Quality Assurance

Prior to conducting any statistical analyses of WET data in this project, all data were screened to confirm that data quality requirements were met as summarized in Table 3. As described below, several quality requirements had to be met for each set of WET test data in order to be included in the analyses. These were:

- Test meets the specific EPA toxicity test method's test acceptability criteria (TAC)
- Required minimum number of test concentrations were used in the test
- WET data were reported for all endpoints relevant to a given EPA WET test method(s)
- Required minimum number of replicates were used as prescribed by the EPA WET test methods

For Methods 1000.0 and 1002.0, all tests with less than six concentrations (including the control) were removed from analysis because these methods require a minimum of 6 concentrations. For other EPA WET test methods, all tests with less than five concentrations (including the control) were removed from analysis because those tests require at least five concentrations.

All test estimate values using EPA recommended approaches were produced using CETIS®. For each EPA WET test method, the results for various endpoints were calculated according to EPA flowcharts in the 2002 WET test method manuals. Test results were reviewed to ensure that (1) data were properly reported; and (2) proper estimates were generated according to the statistical method employed.

WET test estimate values produced using CETIS® or Excel, along with related WET test information, were subjected to a variety of quality assurance procedures as per the EPA quality assurance/quality control requirements (e.g., QAPP) to ensure that the WET data were properly imported and exported into the databases, and that the number of WET tests and laboratories agreed between initial and final WET data sets. Furthermore, all statistical analyses were passed through several stages of quality assurance to ensure that formulae were calculated correctly and that derived endpoints, as well as statistical properties measured for a WET test, were accurate given the statistical approach used. Finally, several checks were used to ensure that compilation

(i.e., meta-analysis) of test statistics (e.g., power, confidence, *b*) for a given statistical approach were complete and accurate.

4.0 Results

4.1 *Ceriodaphnia dubia* (water flea) Chronic Reproduction

For *Ceriodaphnia dubia* (freshwater invertebrate – water flea) reproduction, 75% of the tests (594 tests) had a MSD ≤ 0.25, and the 90th percentile MSD was 0.32 (Table 4-1). These MSDs are slightly lower (i.e., indicative of less within-test variability) than those reported previously by EPA using test data prior to 1995 (e.g., 90th percentile MSD = 0.37; USEPA 2000a). For tests meeting the 75th percentile MSD of 0.25, 65% of the tests exhibited a power ≤ 0.8 (Appendix B) to detect a 25% effect; i.e., only 35% of the tests achieved a power ≥ 0.8 to detect a 25% effect. Using pre-1995 WET data, Denton and Norberg-King (1996) showed that at a power = 80% and MSD = 0.25, most *Ceriodaphnia dubia* reference toxicant tests could not statistically distinguish as much as a 10 offspring difference (30-50% effect) between the control and a test concentration. Oris and Bailer (1993) reported a similar result. Results of this project indicate that precision of this test method has improved since the late 1990s, as exemplified by lower MSDs or control CVs in this project (Figure 2-1).

Monte Carlo simulation analyses indicated that a *b* between 0.68 (1-90th percentile MSD) and 0.70 satisfied all three risk management decision criteria (Table 4-2). Using a *b* value = 0.7, greater protection is afforded for mean effect levels at 20-25% however, the fraction of tests deemed toxic at average CV levels for this endpoint exceeded 0.20 at a 20% mean effect level and was 0.62 at a mean effect level of 25% when within-test variability was relatively low (CV < 25th percentile). Using *b* = 0.68, TST maintains a lower β error rate under these test conditions and meets the β error risk management decision level established for truly non-toxic conditions.

Table 4-1. Summary of statistical variability parameters for EPA WET methods examined in this study. Coefficient of Variation values were calculated based on the mean control response (e.g., offspring per female) and its standard deviation. Minimum significant difference values were calculated based on the no observed effect concentrations for each test at various percentiles (50, 75, and 90th percentile).

	Coefficient of Variation (%)			**Minimum Significant Difference**		
Test Method	**50**	**75**	**90**	**50**	**75**	**90**
Ceriodaphnia dubia (water flea) 7-d survival and reproduction	15	22	35	0.18	0.25	0.32
Pimephales promelas (fathead minnow) 7-d survival and growth	8	12	16	0.16	0.20	0.25
Americamysis bahia (mysid shrimp) 7-d survival and growth	14	15	18	0.12	0.16	0.20
Dendraster excentricus and *Strongylocentrotus purpuratus* fertilization	2	3	4	0.09	0.10	0.16
Haliotis rufescans (red abalone) larval development	4	11	20	0.09	0.13	0.19
Macrocystis pyrifera (giant kelp) germination germ-tube length	 4 4	 5 5	 7 6	 0.09 0.10	 0.16 0.11	 0.24 0.18
Pimephales promelas (fathead minnow) acute survival	Not calculable*		Not calculable*	0.08	0.15	0.21

* CV for control of many tests was 0.0.

Using either a b = 0.68 or 0.70 TST yields greater protection (higher fraction of tests deemed toxic; lower α error) than the current NOEC approach for truly toxic effluents (25 and 30% mean effect) and/or highly variable test results (Table 4-2). TST also yields a lower β error rate (smaller fraction of tests deemed toxic) than the current NOEC approach for mean effect levels ≤ 20% when test variability is low and results are therefore, more certain (Table 4-2). Thus, when within-test variability is abnormally high, TST, using b = 0.68 or 0.70 will tend to declare a sample as toxic. When within-test variability is relatively low, TST at b = 0.68, will tend to find the sample non-toxic, unless there is a 25% effect or more.

Table 4-2. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the *C. dubia* chronic reproduction WET endpoint in comparison with risk management decision criteria as a function of different values of *b* using TST analysis or using the NOEC approach and 15, 20, 25, and 30% effect levels.

Effect Level (%)	Risk Management		*b* Value									
			0.63		0.68		0.70		0.75		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
15	-	≤ 0.2[1]	-	0.00	-	0.001	-	0.003	-	0.35	-	0.99
20	-	-	-	0.00[5]	-	0.02	-	0.21	-	1.00	-	1.00
25	≤ 0.05[2]	0	0.99	0.9[6]	0.00	0.19	0.00	0.62	0.00	1.00	0.78	1.00
30	0.0[3]	-	0.99[4]	-	0.00	-	0.00	-	0.00	-	0.22*	-

1. at within-test CV ≤ 75th percentile
2. at within-test CV > 50th percentile
3. all tests deemed toxic regardless of within-test CV
4. α error rate observed for 0-10th percentile CV
5. at within-test CV between 25-50th percentile
6. at within-test CV ≤ 25th percentile

* CV ≥ 85th percentile.

Analyses of actual WET data support a *b* value of 0.68 for *C. dubia* reproduction (Figure 4-1). At a *b* = 0.68, α and β error rates = 4% and 9%, respectively using a toxic threshold of 25% effect. Higher *b* values had greater sensitivity but lower specificity (higher β error rates). A lower *b* value resulted in much lower sensitivity (higher α error rates). Analyses of lower or higher percent effect thresholds yielded results similar to those obtained from simulation analyses (Appendix B).

4.2 *Pimephales promelas* (fathead minnow) Chronic Growth

Pimephales promelas (freshwater vertebrate - fathead minnow) chronic growth data indicated an MSD somewhat lower (i.e., more precise) than that observed for the freshwater water flea (*C. dubia*); 75% of the tests achieved an MSD ≤ 0.2 (Table 4-1). The 90th percentile MSD was 0.25 (Table 4-1) similar to the MSD previously reported by EPA (USEPA 2000a). Slightly better precision was observed in effluent WET tests as compared to reference toxicant tests (Appendix C). Lower MSD values in this WET test method as compared to the *C. dubia* chronic WET method was anticipated because each replicate in a fish chronic test represents the weight of several fish, which tends to smooth out variability in individual fish weight. However, power of

the chronic fish test to detect a 25% decrease in growth from control was either similar to or less than that observed in the *C. dubia* WET tests (Appendix C). Most (75%) of the fish chronic WET tests had ≤ 70% power to detect a 25% effect when MSD = 0.20. This is because fish tests use fewer replicates per effluent concentration than the *C. dubia* WET test (four versus 10 replicates, respectively).

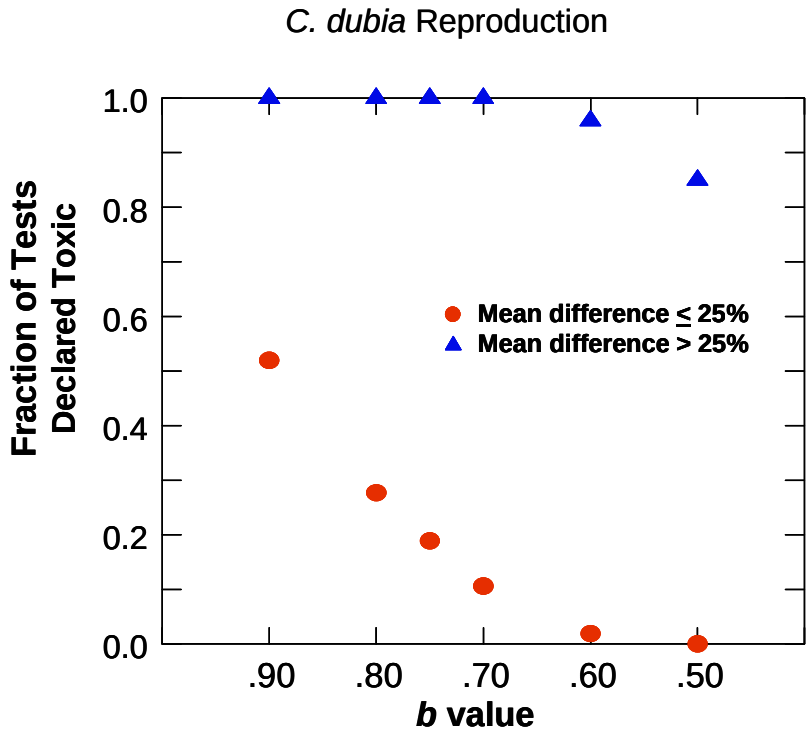


Figure 4-1. Percentage of *Ceriodaphnia dubia (freshwater water flea)* EPA chronic WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether the actual mean difference between control and effluent was less than or greater than 25%. For example, by selecting *b* = 0.68, approximately 99% of the tests with greater than a 25% effect were declared as toxic, whereas for those tests with ≤ a 25% effect, approximately 8% of those tests were deemed as toxic.

Monte Carlo simulation analyses indicated that of the *b* values examined, only *b* = 0.70 met the α and β error risk management decision criteria (Table 4-3; Table C-1, Appendix C). Similar to the *C. dubia* chronic test method, at *b* = 0.75, greater sensitivity was achieved (lower α error) at lower-average mean effect levels, but less specificity than that observed using a *b* = 0.70 (all tests having a 20% mean effect and low CV were deemed toxic; Table 4-3). At *b* values lower than 0.70 sensitivity was poorer, with many tests declared non-toxic at a 30% mean effect level, when within-test variability was low (Table 4-3).

Examining actual WET data yielded results similar to the simulation. Both specificity and sensitivity were greatest (β and α respectively) near $b = 0.7$ (Figure 4-2; Appendix C). Using a $b = 0.7$, nearly all tests having > 25% mean effect were declared as toxic (high sensitivity) and less than 5% of the tests exhibiting ≤ 25% mean effect were declared as non-toxic (high specificity).

Table 4-3. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the *P. promelas* chronic growth WET endpoint with risk management decision criteria as a function of different *b* values for TST analysis or the NOEC approach and 15, 20, 25, and 30% effect levels. See footnotes for Table 4-2 for CV percentiles relevant to each decision levels.

Effect Level (%)	Risk Management		*b* Value									
			0.60		0.65		0.70		0.75		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
15	-	≤ 0.2	-	0.00	-	0.00	-	0.00	-	0.28	-	0.46
20	-	-	-	0.00	-	0.02	-	0.51	-	1.00	-	1.00
25	≤ 0.05	-	0.35	0.00	0.02	0.00	0.00	0.624	0.00	1.00	0.65	1.00
30	0.00	-	0.99	1.00	0.71	-	0.00	-	0.00	-	0.69*	-

* α error rate at CV > 85th percentile

If it was desired to achieve a β error rate ≤ 0.20 for a 20 or 25% mean effect level and average within-test variability, along with the other decision criteria already identified, a *b* value between 0.65 and 0.70 is more appropriate.

Using $b = 0.70$, TST yields greater protection (lower α error) than the current NOEC approach for the *P. promelas* growth endpoint and mean effect ≥ 25% and/or tests having relatively high variability (Table 4-3). TST also yields a lower β error rate using $b = 0.70$ than the current NOEC approach for mean effect levels ≤ 20% when test variability is low-average.

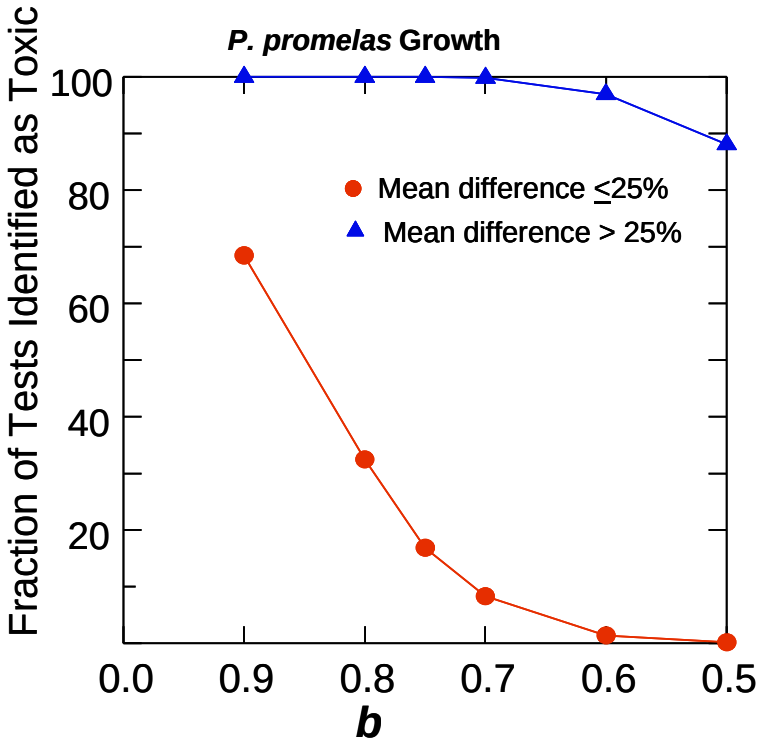


Figure 4-2. Percentage of *Pimephales promelas* (fathead minnow) EPA chronic WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether the actual mean difference between control and effluent was less than or greater than 25%. For example, by selecting *b* = 0.7, approximately 100% of the tests with greater than a 25% effect were declared as toxic, whereas for those tests with < a 25% effect, less than 10% of those tests were deemed as toxic.

4.3 *Americamysis bahia* (mysid shrimp) Chronic Growth

Most (75%) chronic *Americamysis bahia* (saltwater invertebrate - mysid shrimp) tests yielded MSD values < 0.3 (Table 4-1). Similar to results from both *Ceriodaphnia dubia* reproduction and *Pimephales promelas* growth analyses, slightly better precision (lower MSD values) was observed in effluent WET tests than in reference toxicant tests (Appendix D). Like the *C. dubia* reproduction and the *P. promelas* growth endpoints, the mysid growth endpoint is not generally precise enough to meet risk management criteria for α and β error rates using TST at *b* = 0.75. A slightly lower *b* value is needed to simultaneously meet all risk management criteria.

Similar to the *P. promelas* chronic endpoints evaluated above, Monte Carlo simulation analyses indicated that of the *b* values examined only *b* = 0.70 resulted in TST meeting all risk management decision criteria (Table 4-4). Lower *b* values had unacceptably low sensitivity at mean effect levels ≥ 25% (α error rates ≥ 0.39). A higher *b* value of 0.75 had an unacceptably high β error at mean effect levels ≤ 20% (Table 4-4).

Table 4-4. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the *A. bahia* chronic growth WET endpoint with risk management decision criteria as a function of different *b* values using TST or the NOEC approach and 15, 20, 25, and 30% effect levels. See footnotes for Table 4-2 for CV percentiles relevant to each decision levels.

Effect Level (%)	Risk Management		*b* Value									
			0.60		0.65		0.70		0.75		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
15	-	≤ 0.20	-	0.00	-	0.00	-	0.00	-	0.24	-	1.00
20	-	-	-	0.00	-	0.00	-	0.41	-	1.00	-	1.00
25	≤ 0.05	-	0.39	0.00	0.04	0.00	0.00	0.95	0.00	1.00	0.27	1.00
30	0.00	-	1.00	-	0.16	-	0.00	-	0.00	-	0.44*	-

* α error rate at CV > 85th percentile

If, in addition to the risk management criteria identified it was also desirable to reduce β error ≤ 0.20 for a mean effect level of 20% when within-test variability is low (CV < 25th percentile for this WET endpoint), a *b* value slightly lower than 0.70 might be more appropriate.

Using *b* = 0.70, TST yielded greater protection (lower α error) than the current NOEC approach for the *A. bahia* growth endpoint and effluent effects ≥ 25% (Table 4-4). TST, using *b* = 0.70 also yielded a lower β error rate than the current NOEC approach for this WET endpoint and mean effect levels ≤ 20%, especially, when with-test variability is low-average (Table 4-4).

Using actual WET data and a 25% mean effect as the toxicity threshold (risk management goal) for mysid growth, a *b* value near 0.70 appeared to be appropriate because sensitivity and specificity were both simultaneously highest (i.e., lowest and β error, respectively) (Figure 4-3; Appendix D). These results are consistent with the simulation results described above.

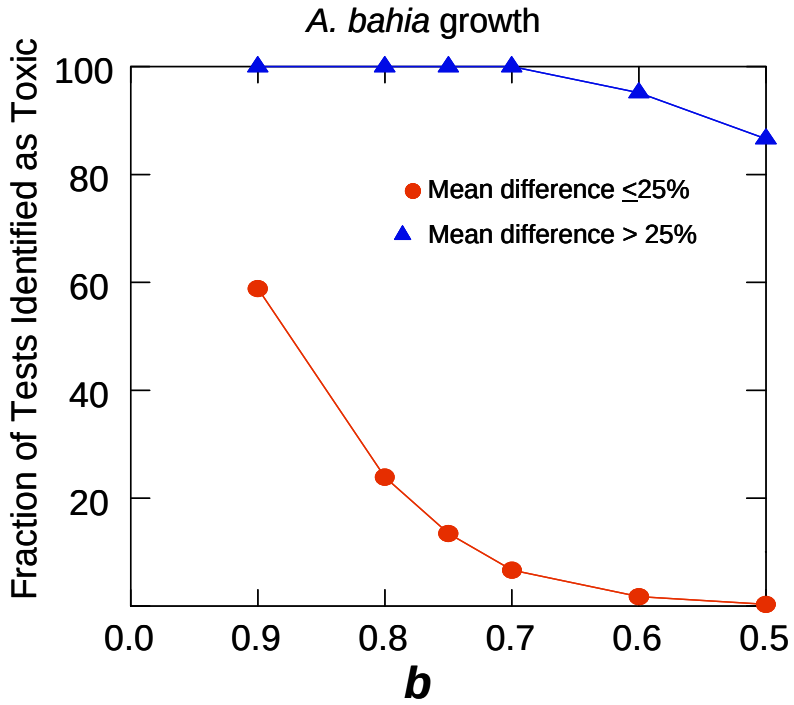


Figure 4-3. Percentage of *Americamysis bahia* (mysid shrimp) EPA chronic WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether the actual mean difference between control and effluent was less than or greater than 25%. For example, by selecting a $b = 0.7$, approximately 99% of the tests with greater than a 25% effect were declared as toxic, whereas for those tests with ≤ 25% effect, less than 5% of those tests were deemed to be toxic.

4.4 *Dendraster excentricus* and *Strongylocentrotus purpuratus* (Echinoderm) Fertilization Test

In contrast to results presented for the previous two freshwater chronic and one East Coast mysid WET test endpoints, the West Coast marine WET method for *Dendraster excentricus* and *Strongylocentrotus purpuratus* (saltwater echinoderm - sand dollar and sea urchin, respectively) fertilization endpoint exhibited substantially higher test precision (Table 4-1). Seventy-five percent of the tests had MSD values < 0.10 and 90% of the tests had MSD values < 0.16. Unlike the three WET tests discussed above, the echinoderm reference toxicant tests exhibited slightly better precision (lower MSDs) than effluent tests (Appendix E). Power analysis showed that 75% of the tests could detect a 20% effect with a power of 0.8 and 80% of the tests could detect

a 25% effect with a power of 0.8 (Appendix E). This demonstrates that the echinoderm test has relatively high precision and high power to detect toxicity when present.

Simulation analyses indicated that of the *b* values examined, a *b* = 0.8 resulted in TST meeting all risk management decision criteria for West Coast WET endpoints (Table 4-5). Using *b* = 0.80 for this test method, TST was more protective (lower α error rate) than the current NOEC approach for mean effect levels ≥ 20% when within-test variability was relatively high. Also, TST had greater specificity (lower β error rate) at mean effect levels ≤ 15% particularly when within-test precision was high and therefore, certainty in the decision of non-toxicity greater.

Table 4-5. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the Sea urchin fertilization chronic growth WET endpoint with risk management decision criteria as a function of different *b* values using TST or the NOEC approach and 10, 15, 20, and 25% effect levels.

Effect Level (%)	Risk Management		*b* Value									
			0.70		0.75		0.80		0.85		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
10	-	≤ 0.20[1]	-	0.01	-	0.14	-	0.27	-	0.59	-	0.79
15	-	-	-	0.00[5]	-	0.03	-	0.22	-	0.89	-	1.00
20	≤ 0.05[2]	-	0.13	0.02[6]	0.04	0.27	0.00	0.76	0.00	1.00	0.28	0.95
25	0.00[3]	-	0.04[4]	-	0.44	-	0.00	-	0.00	-	0.86*	-

1. at CV ≤ 50th percentile
2. at CV ≥ 50th percentile
3. all tests declared as toxic regardless of CV
4. at CV = 0-10th percentile
5. at CV between ≤ 25th percentile
6. at CV < 25th percentile

* α error rate at CV > 85th percentile

Analyses based on actual WET data for this test method indicate that a *b* value near 0.8 results in simultaneously lowest α and β error rates using TST and a mean effect threshold of 20% (Figure 4-4). These results are consistent with the simulation results.

4.5 *Haliotis rufescans* (red abalone) Larval Development

The *Haliotis rufescans* (saltwater invertebrate - red abalone) West Coast WET test method demonstrated high precision and statistical power to detect relatively small effect levels, similar

to the Sea Urchin fertilization test method. Only reference toxicant test data were available for this WET test method. The 90th percentile MSD for larval development was 0.2 (Table 4-1; Appendix F). More than 80% of the tests could detect a 25% effect with a power of at least 0.8, indicating that most tests could detect a toxic effluent most of the time (Appendix F).

Simulation analyses indicated that a b = 0.8 resulted in TST meeting all risk management decision criteria (Table 4-6; Appendix F). In addition, using b = 0.8 for this test method resulted in greater protection than the current NOEC approach (Table 4-6). Lower α and β error rates were observed using TST at critical risk management effect levels (15-25% effect, Table 4-6).

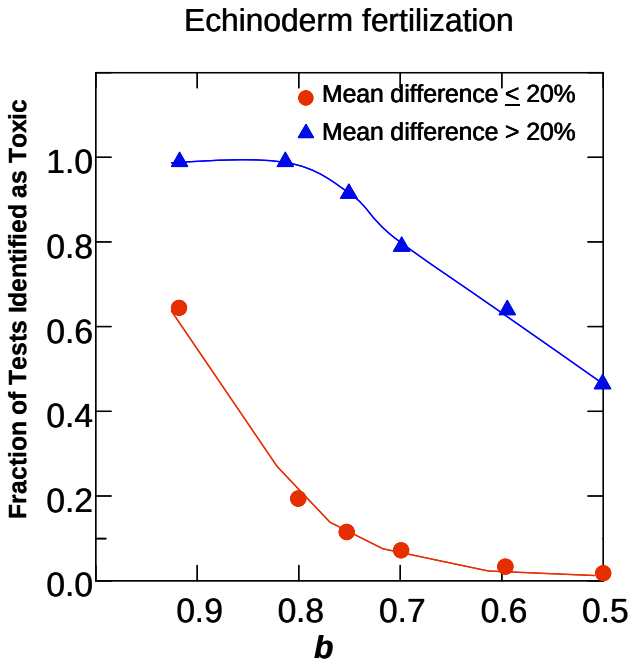


Figure 4-4. Percentage of *Dendraster excentricus* and *Strongylocentrotus purpuratus* (Echinoderm) EPA chronic WET tests in which the effluent is declared to be toxic using TST and different b values in relation to whether the actual mean difference between control and effluent was less than or greater than 20%. For example, by selecting a b = 0.8, approximately 100% of the tests with greater than a 20% effect were declared as toxic, whereas for those tests with ≤ 20% effect, approximately 20% of those tests were deemed to be toxic.

Table 4-6. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the Red abalone larval development chronic growth WET endpoint with risk management decision criteria as a function of different *b* values using TST or the NOEC approach and 10, 15, 20, and 25% effect levels. See Table 4-5 for explanation of footnotes.

Effect Level (%)	Risk Management		*b* Value									
			0.70		0.75		0.80		0.85		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
10	-	≤ 0.20[1]	-	0.00	-	0.02	-	0.18	-	0.67	-	0.88
15	-	-	-	0.00	-	0.06	-	0.33	-	0.97	-	1.00
20	≤ 0.05[2]	-	0.36	0.03	0.11	0.38	0.00	0.91	0.00	1.00	0.24	1.00
25	0.00[3]	-	0.90	-	0.35	-	0.00	-	0.00	-	0.35*	-

* α error rate at CV > 85th percentile

Analyses of reference toxicant test data for this method indicated similar results as the simulation analyses (Figure 4-5). At *b* near 0.8, sensitivity and specificity appeared to be simultaneously greatest compared to other *b* values.

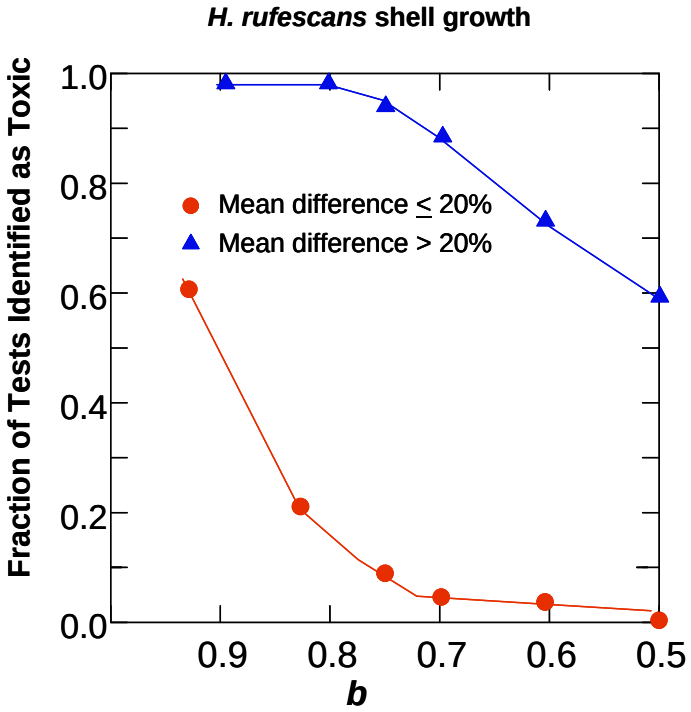


Figure 4-5. Percentage of *Haliotis rufescans* (Red abalone) EPA chronic WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether the actual mean difference between control and effluent was less than or greater than 20%. For example, by selecting a *b* = 0.8, approximately 98% of the tests with greater than a 20% effect were declared as toxic, whereas for those tests with ≤ 20% effect, approximately 20% of those tests were deemed to be toxic.

4.6 *Macrocystis pyrifera* (giant kelp) Germ-tube Length and Germination

There were very few effluent tests available for this West Coast marine test method. Therefore, analyses focused on reference toxicant test data only. Similar to results observed with the two other West Coast methods examined, the *Macrocystis pyrifera* (saltwater plant – giant kelp) WET test method has relatively high precision and high statistical power to detect small differences in response between the effluent and a control. Two different responses are measured in this test: germ-tube length and germination; therefore, these responses were analyzed separately. The 90th percentile MSD was 0.13 and 0.22 for germ-tube length and germination response, respectively (Table 4-1). Almost 90% of the tests could detect a 75% effect with a power of at least 0.8, indicating that most tests could detect a toxic effluent most of the time (Appendix G).

Simulation analyses indicated that a $b = 0.8$ yielded results consistent with risk management criteria using TST for both endpoints, similar to findings for the two other West Coast marine WET methods (Table 4-7; Appendix G). In addition, using $b = 0.8$ for this test method resulted in TST having greater protection than the current NOEC approach (Table 4-7).

Table 4-7. Fraction of tests deemed toxic or non-toxic by Monte Carlo simulation analyses for the Kelp germ-tube length and germination WET endpoints with risk management decision criteria as a function of different *b* values using TST a current NOEC approach and 10, 15, 20, and 25% effect levels. See Table 4-5 for explanation of footnotes.

Effect Level (%)	Risk Management		*b* Value									
			0.70		0.75		0.80		0.85		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
10	-	≤ 0.20[1]	-	0.00	-	0.00	-	0.08	-	0.69	-	0.83
15	-	-	-	0.00	-	0.05	-	0.68	-	1.00	-	1.00
20	≤ 0.05[2]	-	0.74	0.04	0.00	0.66	0.00	1.00	0.00	1.00	0.11	1.00
25	0.00[3]	-	1.00	-	0.00	-	0.00	-	0.00	-	0.00*	-

* α error rate at CV > 85th percentile

Analyses of reference toxicant test data for this method indicated similar results (Figure 4-6). At a *b* near 0.8, sensitivity and specificity appeared to be simultaneously greatest compared to at other *b* values.

4.7 *Pimephales promelas* (fathead minnow) Acute Survival

The *Pimephales promelas* (freshwater vertebrate - fathead minnow) acute test results were based on 347 effluent tests, all of which used EPA's 2002 WET test methods (USEPA 2002c). This WET test exhibited relatively lower MSD values than the fathead minnow chronic test (Table 4-1), which may be because acute tests are based on survival only, an easily measured endpoint, and test acceptability criteria limit the range of control mortality allowed by the method (≤ 10% mortality). The 75th and 90th percentile MSD values for this test method were 0.15 and 0.21, respectively (Appendix H). Approximately 90% of the tests could achieve a power of 0.9 to distinguish a 30% effect.

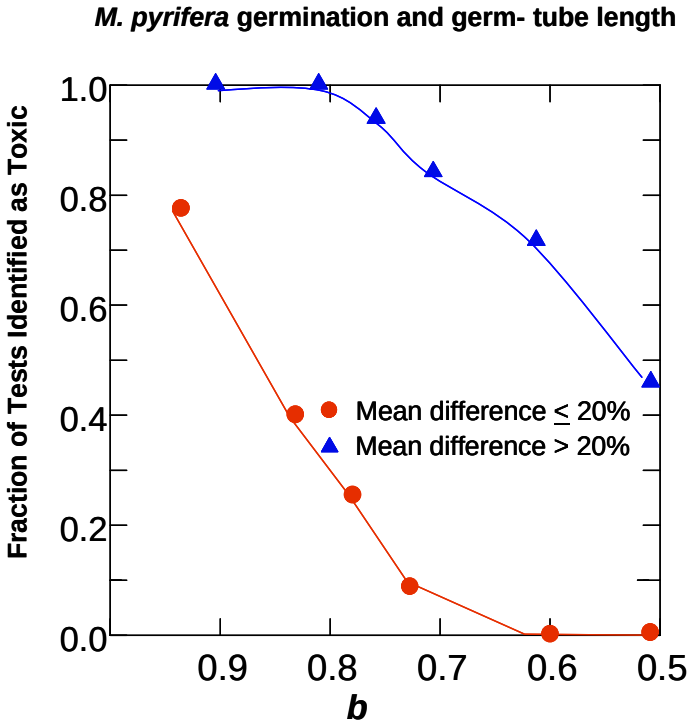


Figure 4-6. Percentage of *Macrocystis pyrifera* (Giant kelp) EPA chronic WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether the actual mean difference between control and effluent was less than or greater than 20%. For example, by selecting a *b* = 0.8, approximately 100% of the tests with greater than a 20% effect were declared as toxic, whereas for those tests with ≤ 20% effect, 25% of those tests were declared as toxic.

Simulation analyses indicated that a *b* = 0.65-0.70 met risk management decision criteria for this acute WET test method, identifying > 30% mean effect as toxic 100% of the time, regardless of within-test variability and a 25% mean effect as toxic most of the time ($\alpha < 0.01$) under average

or higher within-test variability (Table 4-8). In addition, a *b* = 0.65-0.70 met the third decision criterion of identifying a 15% mean effect as non-toxic most of the time ($\beta \leq 0.20$) under normal within-test variability (Table 4-8). These results are similar to those observed based on actual WET data. At a *b* between 0.60 and 0.70, > 25% mean effect on survival was declared as toxic nearly 100% of the time (α error = 0.03; Figure 4-7). In this same *b* value range, TST declared < 10% of the tests toxic when the mean effect was ≤ 25% (Figure 4-7).

Table 4-8. Fraction of tests deemed toxic or non-toxic obtained using simulation analysis of *P. promelas*, acute, survival data in relation to a range of *b* values in TST analysis and using the current NOEC approach for several mean percent effect levels.

Effect Level (%)	Risk Management		*b* Value									
			0.60		0.65		0.70		0.75		NOEC	
	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)	Non-Toxic (α)	Toxic (β)
15	-	≤ 0.20[1]	-	0.00	-	0.00	-	0.07	-	0.46	-	0.58
25	-	-	0.36[2]	0.00[1]	0.07	0.22	0.00	0.79	0.00	1.00	0.50	0.98
30	≤ 0.05[2]	-	0.36	-	0.00	-	0.00	-	0.00	-	0.27	-
40	0.00[3]	-	0.00	-	0.00	-	0.00	-	0.00	-	0.02[2]	-

1 CV: ≤ 90th percentile
2 CV ≥ 90th percentile
3 all tests regardless of within-test CV

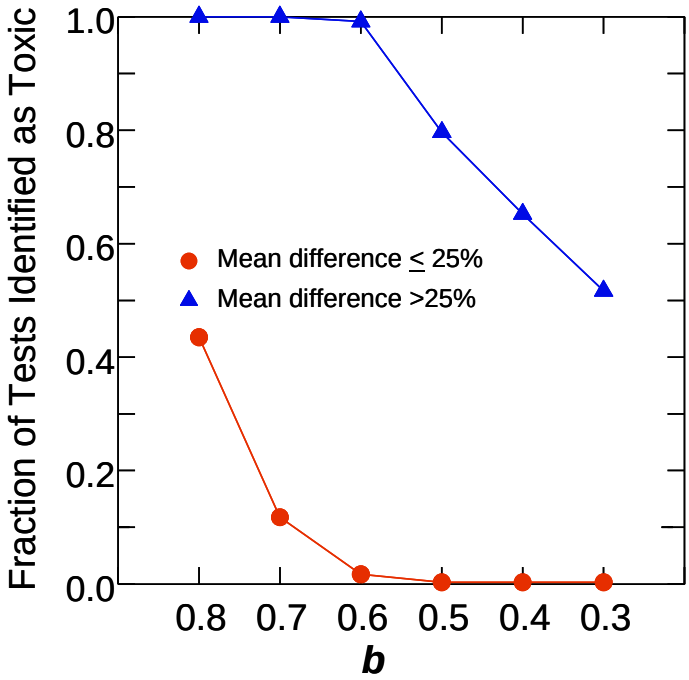


Figure 4-7. Percentage of *Pimephales promelas* (fathead minnow) EPA acute WET tests in which the effluent is declared to be toxic using TST and different *b* values in relation to whether: (A) the actual mean difference in survival between control and effluent was ≤ or > 25%.

5.0 Evaluation of the TST Approach for 2 Sample-Concentration Test Designs

In ambient and stormwater toxicity testing, a laboratory control and a single concentration (i.e., 100% stormwater or ambient water) are often tested. In these two-concentration WET tests, the objective is to determine if a given sample or site water is toxic, as indicated by a significantly different organism response as compared to the control. In this WET testing design, the determination of pass or fail (i.e., toxic or not toxic) is ascertained using a standard t-test (USEPA 2002c). EPA Regions 9 and 10 recommend that the statistical significance (i.e., pass/fail) of a two-sample test design be determined with either a modified t-test (if homogeneity of variance is not achieved) or a standard t-test (if homogeneity of variance is achieved). In many applications, such as California's SWAMP program, the level of significance used in the t-test (i.e., alpha) is 0.05, indicating that the false positive rate is fixed at no more than 5%. In some States (e.g., North Carolina), a significance level of 0.01 is used in t-tests in NPDES permitting, indicating a false positive rate ≤ 1%. As discussed in Chapter 1 of this document, the t-test, a type of hypothesis test, is not usually designed to minimize the rate of false negatives in the WET program.

The TST approach can be used for two-concentration tests as well as multi-concentration tests. As is evident from the results presented in Chapter 4 of this report, TST is generally more protective than the current hypothesis (t-test) approach particularly when within-test variability is relatively high. To demonstrate the value of TST in ambient toxicity programs, ambient toxicity test data were obtained from California's SWAMP program for 409 tests for *Ceriodaphnia dubia* and 256 chronic tests for *Pimephales promelas* using EPA's 2002 WET test methods (USEPA 2002a). WET data for each WET test method were subjected to the same statistical analyses as described in Section 2.5 of this document.

5.1 *Ceriodaphnia dubia* (water flea) Chronic Ambient Toxicity Tests

Analyses of reproduction WET data from 409 ambient toxicity tests (control and stream sample) indicated that 75% of the tests had an MSD ≤ 0.25 and the 90th percentile MSD was 0.33

(Appendix I), similar to the MSD values observed based on multiple concentration tests in this study (Table 4-1).

Figure 5-1 summarizes findings based on the 409 *C. dubia* ambient toxicity tests analyzed and a $b = 0.68$ identified in Chapter 4 for this test method. Although the majority of the tests examined resulted in the same decision using either TST or the current t-test, approximately 5% of the tests (20 tests) would have been declared non-toxic using the t-test approach with mean effect levels ranging as high as 31%. In addition, 3% of the tests (12 tests) would have been declared toxic at mean effect levels as low as 7%.

Figure 5-2 shows ranges of CV values observed in ambient *C. dubia* toxicity tests for those samples declared toxic using either TST or the standard t-test, but not both approaches. As expected based on results of analyses presented in Chapter 4 of this report, within-test variability was relatively high (higher CVs) for those tests found non-toxic using a t-test but toxic using TST. These results again demonstrate a weakness of standard hypothesis testing when control variability is relatively high. As mentioned above, under these conditions, the t-test did not have the power to detect up to a 31% mean effect. Figure 5-2 also demonstrates that TST is superior to the standard t-test when within-test variability is relatively low and the mean percent effect is well below the risk management level. Under these conditions, the standard t-test declared a sample toxic using this WET test method even when the mean effect was between 7 and 21%. TST, however, declared such samples non-toxic using the *b* value of 0.68 identified in Chapter 4. Thus, TST reduces the number of events classified as toxic when effects are actually well below risk management levels of concern.

5.2 *Pimephales promelas* (fathead minnow) Chronic Ambient Toxicity Tests

Analyses of growth data from 256 *Pimephales promelas* (freshwater vertebrate - fathead minnow) chronic ambient toxicity tests indicated an MSD similar to that observed for multiple concentration results; 75% of the tests achieved an MSD ≤ 0.2 and 90% achieved an MSD ≤ 0.28 (Appendix J).

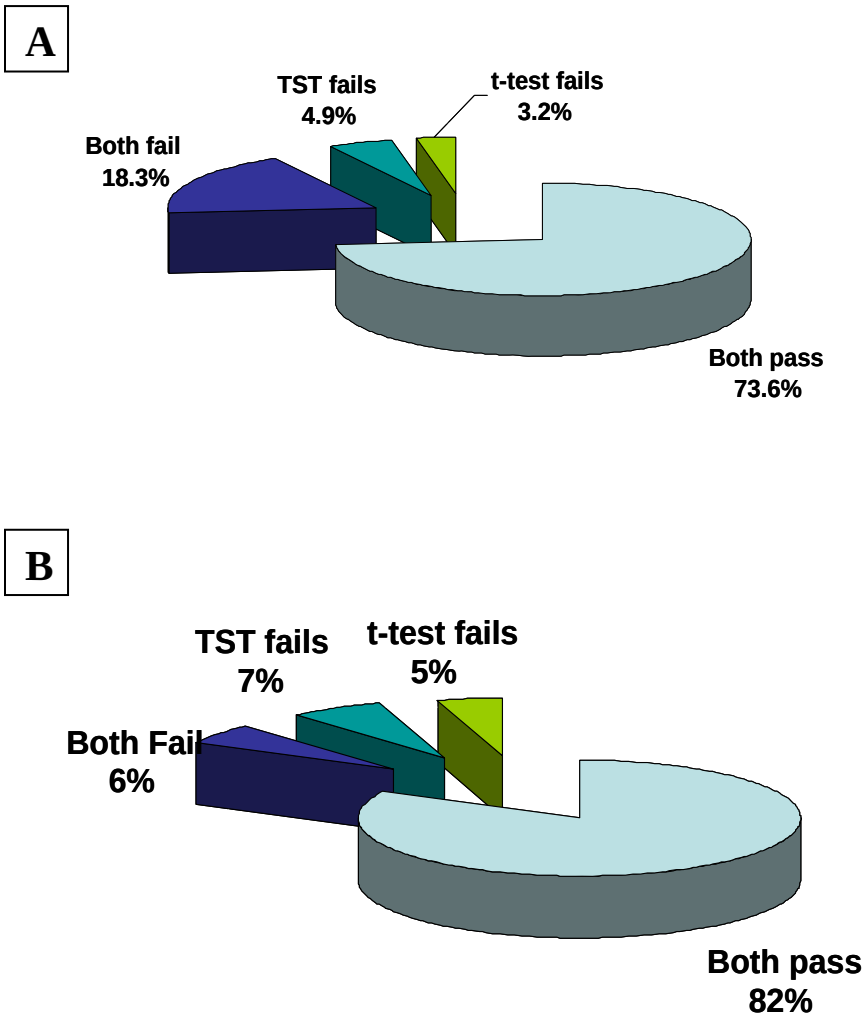


Figure 5-1. Concordance between results of (A) *Ceriodaphnia dubia* (water flea) and (B) *Pimephales promelas (fathead minnow)* EPA chronic ambient toxicity tests using TST and a standard t-test analysis of the data. *b* value = 0.68 for *Ceriodaphnia* and 0.70 for *Pimephales* TST analyses.

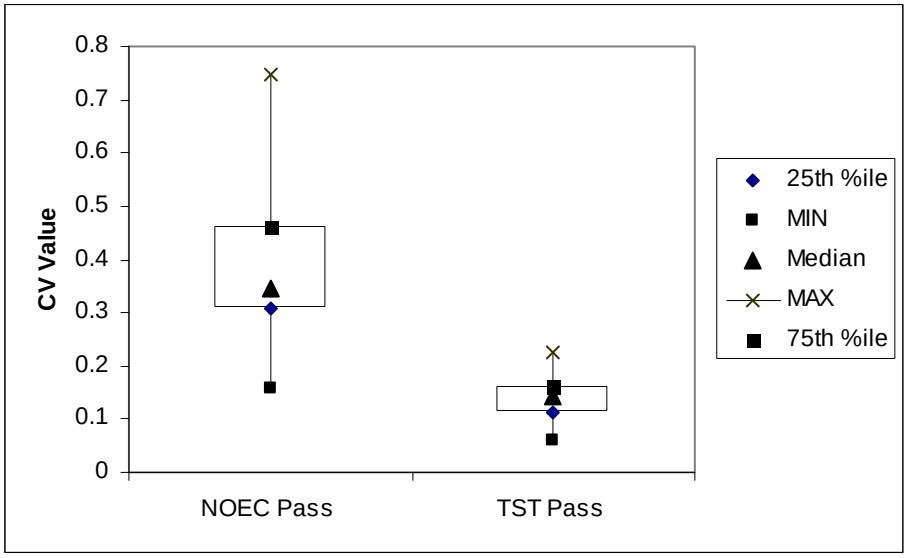


Figure 5-2. Range of coefficient of variation (CV) values observed in ambient chronic *C. dubia* toxicity tests for samples that were found to be non-toxic using the standard t-test but toxic using TST ("NOEC Pass") and for those samples deemed toxic using t-test but not TST ("TST Pass"). Test data supplied by California's Surface Water Ambient Monitoring Program (SWAMP).

Figure 5-1 summarizes test outcomes using either the TST (and $b = 0.70$ based on results in Chapter 4) or the current t-test approach. Similar to the *C. dubia* ambient toxicity tests, most of the *P. promelas* ambient tests yielded the same outcome using either statistical approach. However, 7% of the 256 tests (18 tests) were declared non-toxic using t-test, despite mean effect levels as high as 33%. In addition, 5% (13 tests) were declared toxic using the t-test approach at effect levels as low as 17%. By comparison, TST rarely declared samples as non-toxic at mean effect levels > 20% and rarely declared samples as toxic at mean effect levels < 20%.

Similar to the ambient C. dubia test data, within-test variability was higher in those tests found non-toxic using a t-test but toxic using TST (Figure 5-3). Similarly, those tests deemed non-toxic by TST but toxic using t-test had lower within-test variability and mean effect levels < 25% (Figure 5-3). Thus, as with the ambient *C. dubia* tests, data from ambient *P. promelas* tests demonstrate that TST provides better protection than the standard t-test approach while also identifying those samples that are truly non-toxic from a risk management perspective.

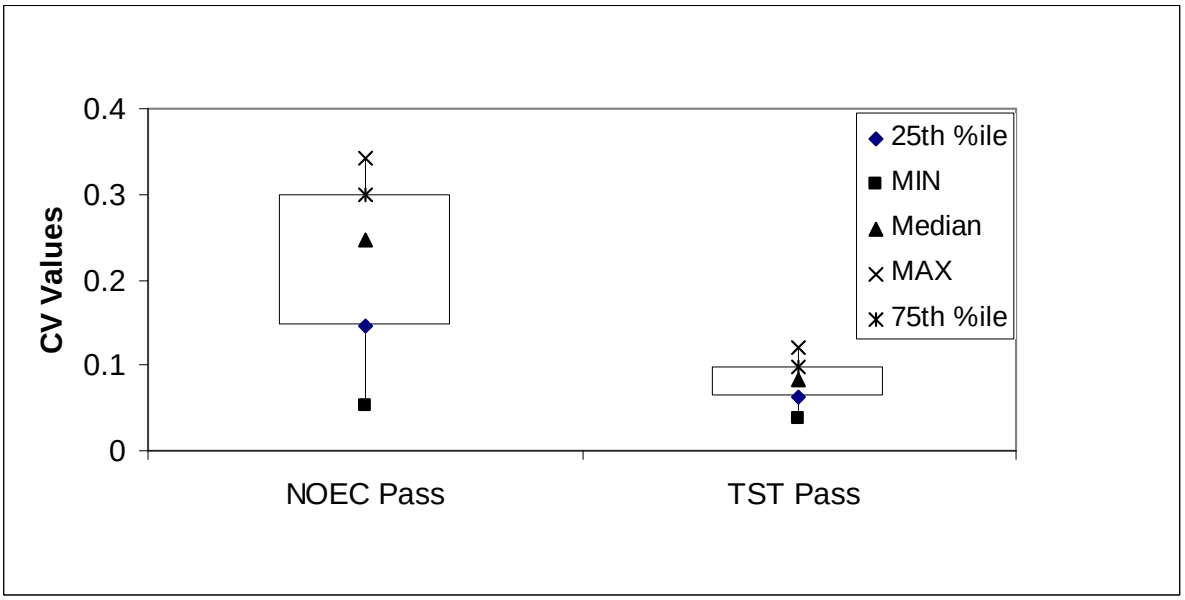


Figure 5-3. Range of coefficient of variation (CV) values observed in ambient chronic *P. promelas* toxicity tests for samples that were declared to be non-toxic using the standard t-test but toxic using TST ("NOEC Pass") and for those samples declared toxic using t-test but not TST ("TST Pass"). Test data supplied by California's Surface Water Ambient Monitoring Program (SWAMP).

6.0 Conclusions and Recommendations

Results of this project indicate that the Test of Significant Toxicity (TST) is a viable alternative approach for analyzing acute and chronic WET test data. Furthermore, given an appropriate test-method specific *b* value, TST is as or more protective as the current EPA recommended hypothesis testing approach while also providing protection for biologically meaningful differences between effluent and control responses (Tables 6-1 and 6-2). Figure 6-1 compares WET test interpretations, derived from simulation analyses, based on using either TST (with the test-method-specific *b* value identified in this project) or the current NOEC approach as a function of within-test variability for the *C. dubia* reproduction test method. A similar pattern would be obtained for the other WET methods. These results demonstrate that the current NOEC approach is less protective than TST for those effluents having a mean percent effect between 15 and 25% when within-test variability is relatively high. Likewise, TST has a lower false positive rate than the current NOEC approach when the mean effect level is well below the risk management level, especially when within-test variability is low (i.e., data quality is high).

This project demonstrates that the use of test-method specific *b* values, that meet desired α and β error rates at critical effect levels, is a technically defensible approach for implementing TST in WET testing. Finally, the *b* values developed in this project are consistent with, and build upon, existing statistical information on WET previously published by EPA, including *Understanding and Accounting for Method Variability in WET Applications Under the NPDES Program* (USEPA 2000a) and *Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing (40 CFR Part 136)* (USEPA 2000b). For most WET test methods examined in this project, a *b* equivalent to 1-the 90th percentile MSD (as discussed in USEPA, 2000a and 2000b) for a given test method is a reasonable approximation of an appropriate *b* value given the α and β error rates and critical effect levels identified as risk management criteria in this project. The degree to which the test-method specific *b* value is approximated by 1-90th percentile MSD is

also dependent on having current, valid, representative WET test data. Using the method-specific *b* values derived in this project, TST:

1. Demonstrated high sensitivity and specificity in terms of detecting a critical effect level (25% for chronic freshwater and East Coast mysid methods and 20% effect for West Coast methods).
2. Demonstrated low α and β error rates at effect levels of interest.
3. Demonstrated that it is generally more protective than the current NOEC or t-test approach, particularly when within-test variability is average or above average.
4. Demonstrates how higher quality WET test data (lower within-test variability) is encouraged.
5. Demonstrates how lower quality WET test data (higher within-test variability) is not rewarded.

If TST declared either an effluent or ambient sample as toxic when the mean effect was less than the risk management effect level, it was because within-test variability was relatively high. As an example, Figure 6-2 shows the range of MSD values observed in *Ceriodaphnia dubia* chronic WET tests that were either declared non-toxic or declared toxic using TST for tests in which there was a 15-20% average effect in the effluent. WET tests having this range of effect in the effluent are typically problematic using current recommended approaches because there is disagreement as to whether the effluent or ambient sample is toxic or not. Using TST those tests that were declared non-toxic had reasonably low within-test variability for this test endpoint (i.e., MSD values were within what 75% of the tests achieved for this test method) (Figure 6-2). Thus, TST fulfills one of the desired objectives: lower within-test variability (higher quality data) is more likely to result in a conclusion of non-toxicity if the mean effect is less than the risk management effect level. Biologically insignificant differences in response between control and effluent or ambient sample are not declared as toxic using TST in this case.

Those tests that were declared to be toxic using TST at a mean effect level between 15 and 20% frequently had relatively high within-test variability (Figure 6-2); i.e., within-test variability was greater than what was observed in 90% of the tests for this test method. Thus, TST fulfills a second desired objective for improving WET data analysis: higher within-test variability (i.e.,

lower quality data) is not rewarded. Such tests are likely to result in declaring the effluent ambient sample toxic.

Table 6-1. Summary of hypothesis rejection rates using Test of Significant Toxicity and the current NOEC approaches for seven EPA WET test methods and Monte Carlo simulation analyses.

EPA WET Test Method	**Number of Tests**	**Current NOEC Approach**		**Draft TST**	
Chronic Freshwater and East Coast Mysid Methods		**Fraction deemed toxic at a 15% mean effect level[1]**	**Fraction deemed non-toxic at a 25% mean effect level[2]**	**Fraction deemed toxic at a 15% mean effect level[1]**	**Fraction deemed non-toxic at a 25% mean effect level[2]**
Ceriodaphnia dubia (water flea) 7-d survival and reproduction	792	0.89	0.20	0.05	0.00
Pimephales promelas (fathead minnow) 7-d survival and growth	472	0.78	0.40	0.14	0.00
Americamysis bahia (mysid shrimp) 7-d survival and growth	210	0.87	0.39	0.04	0.00
Chronic West Coast Marine Methods		**Fraction deemed toxic at a 10% mean effect level[3]**	**Fraction deemed non-toxic at a 20% mean effect level[4]**	**Fraction deemed toxic at a 10% mean effect level[3]**	**Fraction deemed non-toxic at a 20% mean effect level[4]**
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	177	0.85	0.76	0.19	0.00
Haliotis rufescans (red abalone) larval development	136	0.90	0.31	0.16	0.00
Macrocystis pyrifera (giant kelp) germination	135	1.00	0.00	0.00	0.00
germ-tube length	135	0.92	0.07	0.04	0.00
Pimephales promelas (fathead minnow) acute		**Fraction deemed toxic at a 15% mean effect level[5]**	**Fraction deemed non-toxic at a 30% mean effect level[6]**	**Fraction deemed toxic at a 15% mean effect level[5]**	**Fraction deemed non-toxic at a 30% mean effect level[6]**
Pimephales promelas (fathead minnow) acute survival	347	0.58	0.02	0.07	0.21

[1] For tests having $\leq 75^{th}$ percentile within-test CV
[2] For tests having $> 50^{th}$ percentile within-test CV
[3] For tests having $\leq 50^{th}$ percentile within-test CV
[4] For tests having $> 50^{th}$ percentile within-test CV
[5] For tests having $\leq 90^{th}$ percentile within-test CV
[6] For tests having $\geq 90^{th}$ percentile within-test CV

Table 6-2. Summary of test performance characteristics using Test of Significant Toxicity analyses for seven EPA WET test methods examined. *b* values and performance characteristics using actual WET data are based on multi-concentration tests, consistent with current EPA WET test method protocols.

EPA WET Test Method	**Number of Tests**	**Recommended *b* value**		
Chronic Freshwater and East Coast Mysid Methods			**Relative Specificity [1] %**	**Relative Sensitivity[2] %**
Ceriodaphnia dubia (water flea) 7-d survival and reproduction	792	0.68	92	99
Pimephales promelas (fathead minnow) 7-d survival and growth	472	0.70	92	100
Americamysis bahia (mysid shrimp) 7-d survival and growth	210	0.70	96	99
Chronic West Coast Marine Methods			**Relative Specificity [3] %**	**Relative Sensitivity[4] %**
Dendraster excentricus and *Strongylocentrotus purpuratus* (Echinoderm) fertilization	177	0.80	82	100
Haliotis rufescans (red abalone) larval development	136	0.80	83	100
Macrocystis pyrifera (giant kelp) germination	135	0.80	80	100
germ-tube length	347	0.80	80	100
Pimephales promelas (fathead minnow) acute			**Relative Specificity [1] %**	**Relative Sensitivity[2] %**
Pimephales promelas (fathead minnow) acute survival	347	0.70	90	100

1 Percentage of actual WET tests in which TST declared the sample as non-toxic (i.e., bioequivalent) when ≤ 25% difference in mean response was observed between the control and the effluent.
2 Percentage of actual WET tests in which TST declared the sample as toxic (i.e., not bioequivalent) when > 25% difference in mean response was observed between the control and the effluent.
3 Percentage of actual WET tests in which TST declared the sample as non-toxic (i.e., bioequivalent) when ≤ 20% difference in mean response was observed between the control and the effluent.
4 Percentage of actual WET tests in which TST declared the sample as toxic (i.e., not bioequivalent) when > 25% difference in mean response was observed between the control and the effluent.

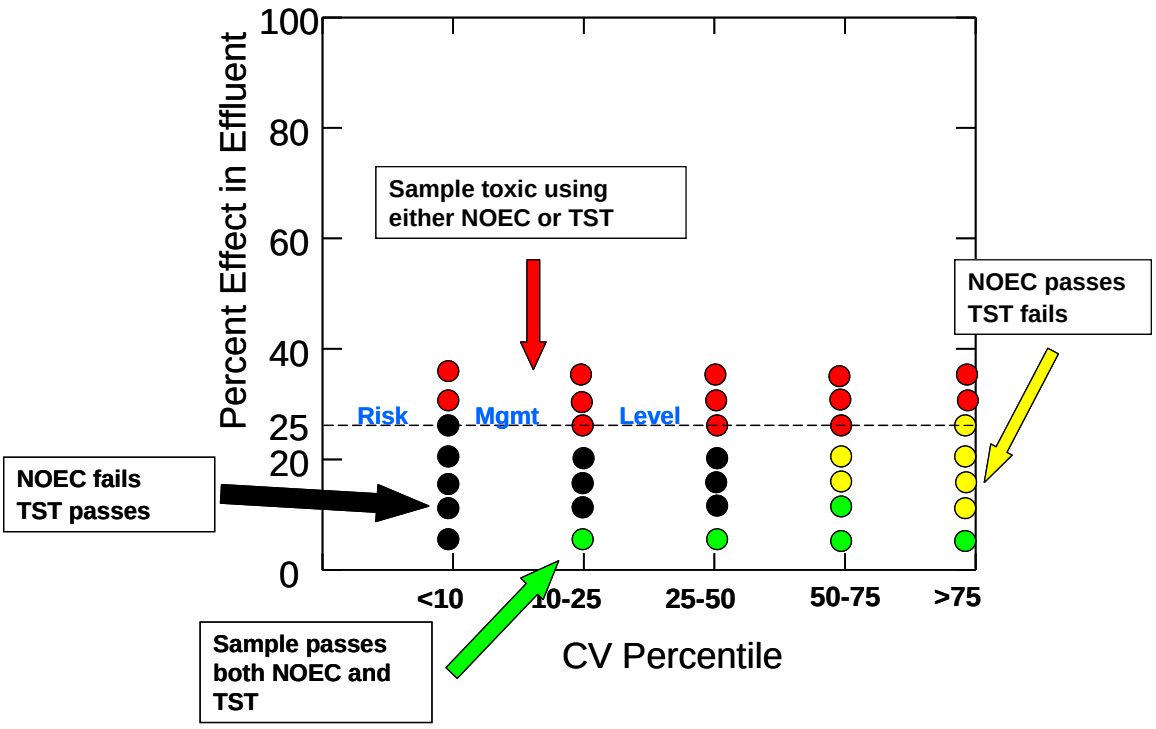


Figure 6-1. Example using *Ceriodaphnia dubia* (freshwater water flea) EPA WET test results illustrating concordance observed between the current NOEC approach and the Test of Significant Toxicity (TST) as a function of coefficient of variation (CV) range observed within a test and percent effect observed in the effluent.

Some of the *b* values generated from this project should apply to other WET test methods that were not examined. For example, the *b* value derived for the *P. promelas* chronic WET test method (USEPA 2002a) should be transferable to other chronic fish test methods that use a similar test design and measure fish growth, such as the sheepshead minnow and inland silverside saltwater chronic fish tests (USEPA 2002b). Similarly, the *b* value derived for the *P. promelas* acute WET test should be transferable to other acute WET tests including daphnid, mysid shrimp, and other fish species tests (USEPA 2002c).

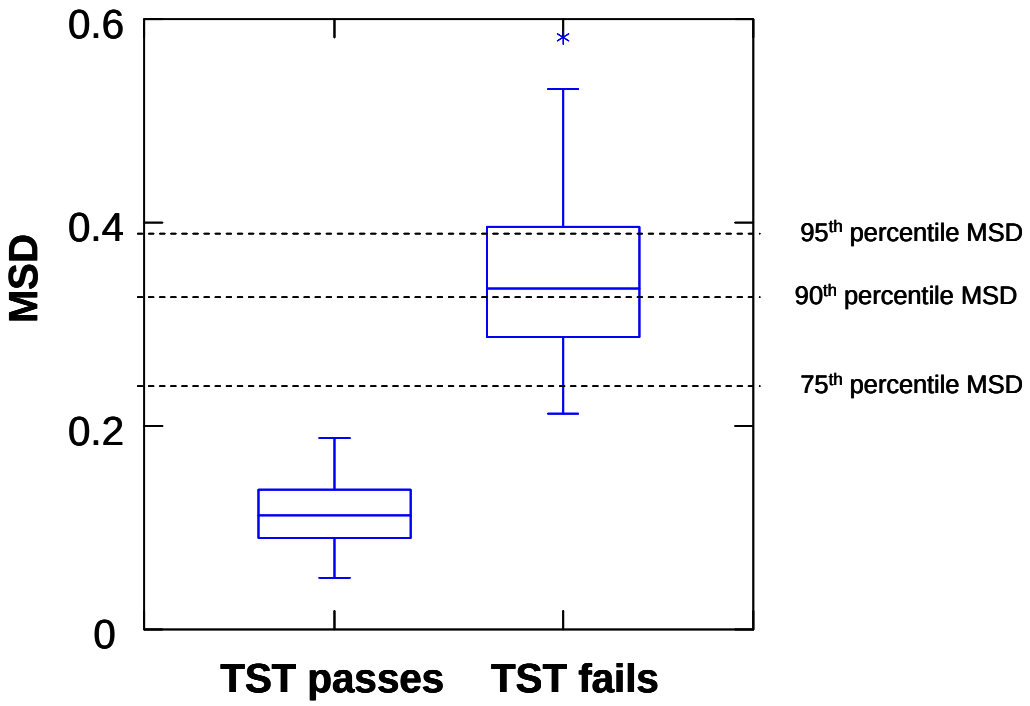


Figure 6-2. Range of variability (MSD values) observed in *Ceriodaphnia dubia* (water flea) EPA WET tests that were declared either non-toxic (TST passes) or toxic (TST fails) using TST and a *b* = 0.70, when the actual effluent effect was between 15 and 20%.

It is important to note that the method specific *b* values identified in this document are dependent on the within-test variability normally achieved by laboratories. As normal within-test variability decreases (i.e., laboratories are able to routinely achieve higher precision or lower MSD values), *b* values should more closely match the critical risk management effect threshold desired. This is particularly an issue for the chronic freshwater and East Coast mysid WET methods. Currently, none of these WET methods are able to support a $b = 0.75$ and achieve the desired α error at 25% mean effect and β error rate at 15% mean effect. Using $b = 0.75$, TST correctly identifies toxicity when it occurs for these WET methods but TST also declares too many tests toxic when they were actually non-toxic. A similar issue has been documented in the bioequivalence literature for other types of data (Berger and Hsu, 1998). Instead, a *b* value near 0.70 was found to meet the risk management error rates. If however, within-test precision for these WET methods should continue to increase in the future (as this project demonstrated, within-test precision has improved for the *C. dubia* chronic methods since 1995), then the test-method specific *b* value should more closely approach 0.75 to achieve the error rates desired at a

risk management effect threshold of 25% effect. Therefore, periodically, EPA will re-evaluate WET test method precision for the chronic freshwater and East Coast methods to ensure that TST continues to provide the level of protection intended.

In summary, analyses of over 2000 EPA WET tests in this project demonstrates that TST incorporates the best features of both the point estimate (i.e., transparency of the effect level, 25%) and hypothesis testing approaches for analyzing WET data. TST incorporates many of the advantages of using an IC_{25} approach for analyzing WET data in that TST is designed to identify an *a priori* risk management effect level with reasonably low α and β error rates. In addition, TST uses a hypothesis testing approach, with re-stated hypotheses, which allows rigorous statistical comparisons similar to the NOEC endpoint, yielding high statistical confidence in the results. Several other specific benefits of using TST in WET analysis are that TST:

- Provides a positive incentive for the permittee to generate high quality WET data
- Incorporates statistical power directly into the NPDES decision process, increasing confidence in the WET test results
- Has the ability to analyze a two-concentration test design (e.g., IWC vs control; stormwater, ambient toxicity test designs) or multi-concentration tests; it is thereby applicable to both NPDES WET permitting and 303(d) watershed assessment programs.

7.0 Literature Cited

Anderson, S. and W.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. 1998. Bioequivalence trials, intersection –union tests and equivalence confidence sets. Statistical Science. 11:283-319.

Denton, DL, Starrett G, Johnson S. 1994. Comparisons of Point Estimate Techniques for West Coast Marine Test Species. Presented at SETAC 15th annual meeting, Denver, CO.

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

Erickson, W.P. and L.L. McDonald. 1995. Tests for non-inferiority of control media and test media in studies of toxicity. Environmental Toxicology and Chemistry 14:1247-1256.

Fairweather, P. 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Marine and Freshwater Research 42:555-567.

Grothe, D.R., K.L. Dickson, and D.K. Reed (eds). 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.

Ng, Tie-Hua. 2001. Choice of delta in equivalence testing. Drug Information Journal 35:1517-1527.

Oris, J. and J. Bailer. 1993. Statistical analysis of the *Ceriodaphnia* toxicity test: sample size determination for reproductive effects. Environ. Toxicol. Chem. 12: 85-90.
Rogers, J., K. Howard, and J. Vessey. 1993. Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin 113:553-565.

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. Can. J. Psychiatry 48(11):756-761.

USEPA. 1991. Technical Support Document for Water Quality-based Toxics Control. EPA/505/2-90-001. United States Environmental Protection Agency, Office of Water, Washington, DC.

USEPA. 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. United States Environmental Protection Agency, National Exposure Research Laboratory – Cincinnati, Office of Research and Development, Washington, DC.

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

USEPA. 2000b. Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing (40 CFR Part 136). EPA/821/B-00-004. United States Environmental Protection Agency, Office of Water, Office of Science and Technology, Washington, DC.

USEPA. 2002a. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms. EPA/821/R-02-013. United States Environmental Protection Agency, Office of Water, Washington, DC.

USEPA. 2002b. Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms, 3rd edition. EPA/821/R-02-14. United States Environmental Protection Agency, Office of Science and Technology, Washington, DC.

USEPA. 2002c. Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms, 5th edition. EPA/821/R-02-012. United States Environmental Protection Agency, Office of Science and Technology, Washington, DC.

APPENDIX A

Data Selection and Processing SOP

Whole Effluent Toxicity (WET) Test Data Selection, Acceptance and Quality Assurance Protocol for Alternate Statistical Approach Analyses

This document describes the data protocol that will be used to obtain, screen and process acceptable Whole Effluent Toxicity (WET) test data for the purpose of comparing results of alternate statistical approaches to determine compliance with WET limits or reasonable potential analysis. These protocols address data sources and procurement, test screening criteria, quality control of data input, and data management for statistical analyses that will be used in a Pilot Study, using North Carolina Ceriodaphnia dubia chronic WET data, and in a larger, more definitive study using data from several WET test methods and data sources. A companion document contains an Analysis Plan that describes proposed analyses for the Pilot Study, which will form the basis of analyses for the more definitive study.

1.0 Data Desired and Data Acquisition

WET test data could be received from several sources including state agencies, EPA regions, and perhaps others. Therefore, it will be important to specify the preferred format of data, the quality of those data, and the representativeness of those data. The following definitions are used in this SOP.

- **WET Test or Test**: An experiment conducted as prescribed in EPA's manuals (1995 or 2002) for WET testing. Also, the set of data resulting from the experiment.
- **Estimate:** The toxicant concentration that is estimated to cause an observable effect on the test organisms. Also known as the Effect Concentration (EC). Estimates of importance in this project include the LC or EC50, which is the concentration that would cause an observed effect in 50 percent of the test organisms on average, and is typically used in acute toxicity tests and sometimes in chronic tests as well. The No Observed Effect Concentration (NOEC) is typically a chronic estimate that refers to the highest toxicant concentration tested at which there is no observed adverse effect on the test organism population. A second chronic estimate that will be examined is the IC_{25}, which is the concentration at which a 25% inhibition or reduction in response is observed in comparison with the response observed in controls.
- **WET Method:** The testing procedure used to conduct the toxicity test (e.g., Method 1000.0 is the fathead minnow larval survival and growth).
- **Endpoint:** The observable effect on test organisms (e.g., inhibition of reproduction, growth, or survival).

- **Estimation Method:** The statistical procedure used to derive point estimates such as Probit, Spearman Karber, or Linear Interpolation.

WET data are desired for six EPA WET test methods listed in Table 1. Preference is given to WET data generated using the 1995 methods for the west coast marine species tests, and the 2002 promulgated methods for other methods. We recognize that much of the WET data currently available may be based on previous versions of EPA methods, particularly with respect to the non-west coast species test methods. Changes made in some of these methods in 2002 could influence our analyses; e.g., the 2002 *C. dubia* chronic test (method 1002.0) requires the use of blocking by known parentage when assigning organisms to test chambers, which has been shown to reduce intra-test variability in many cases. Previous versions of this test method did not require this step. Therefore, the version of the test method used in a given set of WET data will be tracked and subjected to separate statistical analyses initially to determine whether changes in certain methods significantly altered test performance characteristics using a given statistical approach. If no difference in performance is observed with different versions of a given method, then those test data will be used together in statistical analyses.

The first five tests listed in Table 1 are commonly used by regulatory authorities to determine compliance with chronic WET limits or monitoring triggers. The *P. promelas* acute survival test is included as a representative acute WET test method that is commonly required by States and EPA. Test data are desired for each of the six test methods in Table 1 to provide a representative set of test methods that are commonly required of dischargers and that span the types of toxicity endpoints employed. The use of both saltwater and freshwater WET tests also ensures that there is adequate representation of different types of discharge situations.

For each set of test data received, additional metadata information are required including:

- Discharger name and NPDES permit number (coded for anonymity)
- Laboratory name and location (coded for anonymity)
- Design receiving water effluent concentration (expressed as percent effluent upon complete mix) used by the regulatory authority
- Test method version used (cited EPA number)
- Information indicating that all test acceptability criteria were met including those not evident from summary data or data reported at the end of a test (e.g., at least 60% of the control

replicates have at least 3 broods within 8 days in the Ceriodaphnia chronic test – method number 1002.0)

In addition to the above effluent test data and metadata, two other sources of toxicity data will be compiled in this project which will be used to help calculate the range of control organism response by endpoint for each WET test method in Table 1. These data will be instrumental in setting initial *b* values for bioequivalency statistical analysis and other analyses. The first source of data is reference toxicant test control data. Extensive reference toxicant data were previously compiled and analyzed for the EPA document *Understanding and Accounting for Method Variability in Whole Effluent Toxicity Application Under the NPDES Program*, 2000. Therefore, acceptable reference toxicant test data should be readily accessible for this project. Metadata desired in this case would include the laboratory name and location, type of control water used (e.g., reconstituted, dilute mineral water, etc), and method version used.

A second source of control test data that will be compiled in this project is control data generated in ambient toxicity tests by various States. These data could be very useful in supplying information on control responses for a variety of test methods. Many States routinely conduct ambient toxicity tests as part of 305(b) monitoring, TMDLs, and other programs (e.g., California's SWAMP program, Washington Department of Ecology's ambient program, Wisconsin DNR's ambient monitoring program). Metadata desired is the same as that noted above for reference toxicant control data.

WET data already in electronic format, preferably either in ToxCalc® or CETIS®, are desired in this project. Having data in either of these formats will streamline QA/QC regarding test screening, data acceptability and data input, and will provide the most efficient means of analyzing WET data for this project. The next option is receiving data in Excel tables. Hard copy bench sheets are also an option but they will require the most effort in terms of entering the data and the degree of QC needed to ensure quality data for use in this project.

A key aspect of this project and the usefulness of the results obtained will depend on having representative data for each test method. Representativeness is characterized in this project as having data that meet the following:

- Cover a range of facility types, including both industrial and municipal dischargers

- Many facilities are represented for a given test method (i.e., no one facility dominates the data for a given test method)
- Cover a range of target (design) effluent dilutions upon which WET compliance is based, ranging from perhaps 10% to 100% effluent
- Data were generated by several laboratories for a given test method
- Ideally, cover a range of observed effluent toxicity for each method (e.g., NOECs range from < 10% to 100% effluent)

Many tests for each test method in Table 1 will be needed to ensure that the above representativeness characteristics are met and that results of analyses are accurate and can be generalized nationally. The number of tests needed will be defined during the Pilot Study.

1.1 Test Data Format, Selection, and Entry

Processing of raw WET data will begin 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 WET test will be assigned a unique code and each laboratory will also be uniquely coded. Tetra Tech will use a tracking system to help evaluate whether WET data are needed for certain types of WET methods or to help increase representativeness of laboratories, target dilution levels, or types of facilities for a given method. The goal is to have representation of many facilities and types of discharge conditions for each test method in Table 1.

If data are received in either ToxCalc® or CETIS®, they will be imported directly into Tetra Tech's CETIS® database dedicated to this project. Unless otherwise directed (or shown to be redundant because data were already QC'd by the data source organization), WET data in ten percent of the tests received, selected at random, will be checked against the raw data bench sheets to document correct data entry. If any reporting errors are found in this random check that could impact analyses, a decision will be made with the EPA workgroup as to whether the entire dataset is checked and corrected or whether the dataset is not used in the project.

If data are received in Excel or other spreadsheet format, it may not be possible to directly import them into CETIS®, depending on the input format used. In cases where data have not yet been entered

electronically by the source organization and they are using Excel, Tetra Tech will supply them with a template that will help make data transfer to CETIS® fairly easy and help minimize transfer errors. If data are received in an alternate Excel format, data manipulation and greater QA will be required to not only document that data were entered correctly in the spreadsheet received, but also that they were processed correctly. Data entry will be checked on 10% of the tests received against the raw data bench sheets as described above. If the data pass this QC test, then data will be processed into CETIS®. Once in CETIS®, the data will be checked against the spreadsheets for accuracy.

If data are received as hard copies or PDF copies of bench sheets, then data will first be checked to ensure that all method TAC are met, as well as several other requirements discussed in the next section and summarized in Table 1 to minimize extraneous data input effort and expense. Those tests meeting all requirements will be input into the CETIS® database directly using the double entry mode and a comparison of entries to ensure accuracy of data input.

2.0 Data Screening

Prior to conducting any statistical analyses of WET data in this project, all data will be screened to determine whether they meet the quality requirements as summarized in Table 1. As described below, several quality requirements must be met for each set of test data in order to be included in the analyses. These are:

- Test meets the specific toxicity test method's test acceptability criteria (TAC)
- Minimum number of test concentrations were used in the test
- Data are reported for all endpoints relevant to a given method
- Minimum number of replicates are used as prescribed by the method

It is also desirable to have tests spanning a range of observed toxicity for a given method (i.e., non-toxic as well as moderate and highly toxicity observed in tests). To this end, more test data may be collected for certain methods after reviewing the range of toxicity represented in the dataset.

2.1 Test Conditions and Acceptability Criteria

All data must achieve the toxicity test method's test acceptance criteria before they can be used in analyses. A formal QA process will be used to document that test conditions and acceptance criteria listed in the WET methods manuals are met (see Table 1).

2.2 Multiple Endpoints

Data must be reported on the bench sheets and in electronic files for each endpoint in the method. If information is missing for a given endpoint or is incomplete for a test, that test will be removed from the database.

2.3 Number of Test Concentrations in the Dilution Series

For Methods 1000.0, and 1002.0, all tests with less than six concentrations (including the control) will be removed from the database. For other methods, all tests with less than five concentrations (including the control) will be removed from the database.

2.4 Number of Replicates

Tests having less than the minimum number of replicates for a given test method in Table 1 will be eliminated from the database and not used. Any test that used more replicates than the minimum number of replicates as indicated in Table 1 will be flagged as such and subjected to separate statistical analyses initially because number of replicates is expected to influence WET test performance characteristics. For example, tests having more than 10 replicates for the *C. dubia* chronic test method (method 1002.0) will be flagged and initially used in separate statistical analyses. All data will be reviewed to verify that tests meet TAC, including the minimum number of replicates.

2.5 Laboratory Test Selection

For any given facility, data generated by only one laboratory for a particular test method will be compiled in the CETIS® database. Thus, for a given WET test method, each facility will correspond to

one laboratory. Although unlikely, depending on the facility and the test methods, one facility may use different labs for different types of WET tests and such data would be acceptable in this project. For any test method, only the 20 most recent tests will be used if more than 20 tests are available for a given laboratory and facility. As mentioned previously, this project seeks to obtain adequate representation of facilities and laboratories. Therefore, efforts will be made to ensure that a number of laboratories are represented for each test method. This may mean excluding data for a facility in some cases if it used the same testing laboratory as another facility for a given test method.

3.0 Test Estimate Calculation

All test estimate values will be produced using CETIS®. For each method, the results for various endpoints will be calculated according to EPA flowcharts in the WET manuals as well as several alternative statistical approaches as described in the Analysis Plan. Test results will be reviewed to ensure that (1) data were properly reported and (2) proper estimates were generated according to the statistical method employed. For example, verification that Fisher's exact test, rather than the Dunnett or the Wilcoxon test, is used to derive the NOEC and LOEC for the Ceriodaphnia survival endpoint using the EPA flowchart. Also, no estimate from the Probit model will be accepted if the Chi-square test of homogeneity indicated lack of fit at $P<0.05$.

3.1 Quality Assurance

Test estimate values produced using CETIS®, along with related test information, will be subjected to a variety of data QA procedures to ensure that the data are within-range, properly imported and exported, and that the frequencies of tests and laboratories (and toxicants if reference toxicant test data are used) agree between initial and final data sets. Furthermore, all statistical analyses will be passed through various stages of QA to ensure that endpoints derived, as well as statistical properties measured for a test, are accurate given the statistical approach used. Furthermore, QA checks will be used to ensure that compilations (i.e., meta-analysis) of test statistics (e.g., power, confidence) for a given statistical approach are complete and accurate.

TABLE A-1: Summary of Test Condition Requirements and Test Acceptance Criteria for Each EPA WET Method Desired in Alternative Statistics 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)
1000.0	Fathead Minnow *(Pimephales promelas)*	Survival and growth (larval)	Chronic	10	4	6	7 days	$\geq$ 80% survival in controls; average dry weight per surviving organism in control chambers equals or exceeds 0.25 mg
1002.0	Daphnia *(Ceriodaphnia dubia)*	Survival and reproduction	Chronic	1	10	6	Until 60% of surviving control organisms have 3 broods (6 - 8 days)	$\geq$ 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
1003.0	Green Alga *(Selenastrum capricornutum)*	Growth	Chronic	10,000 cells/ml (initial density)	4	6	96 h.	10^6 cells/ml with EDTA or $2*10^5$ cells/ml without EDTA in the controls: variability of controls $\leq$ 20%
1007.0	Mysid *(Mysidopsis bahia)*	Survival, growth	Chronic	5	8	6	7 days	$\geq$ 80% survival; average dry weight $\geq$ 0.20 mg in controls
1016.0	Purple Urchin *(Strongylocentrotus purpuratus) or* Sand dollar *(Dendraster excentricus)*	Fertilization	Chronic	About 1,120 eggs and $\leq$ 3,360,000 sperm per test tube	4	5	40 min (20 min plus 20 min)	$\geq$ 70% egg fertilization in controls; %MSD of <25%; and appropriate sperm counts
2000.0	Fathead Minnow (*Pimephales promelas)*	Mortality	Acute	10	2	4	48 and 96 hrs[1]	90% or greater survival in controls

1: Fathead minnow acute test data may need to segregated initially by test duration and renewal schedule if any, to minimize extraneous sources of variability in statistical analyses.

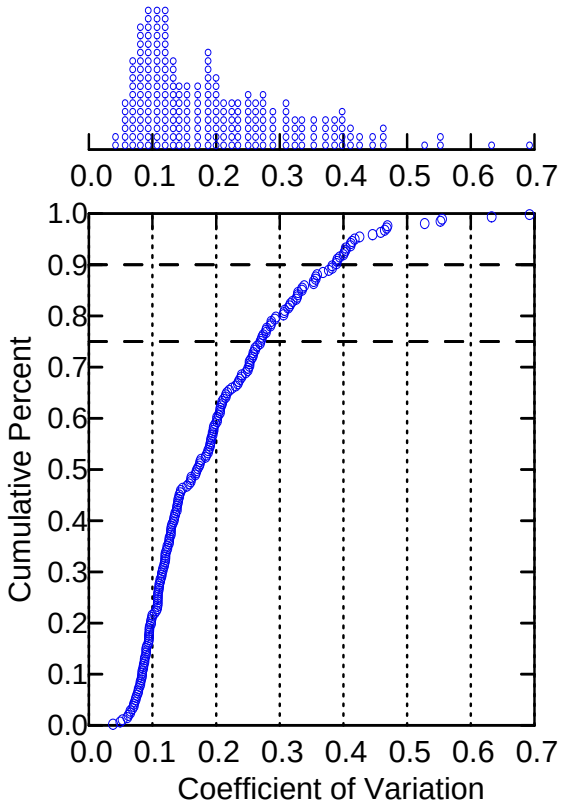
APPENDIX B
Ceriodaphnia dubia Reproduction

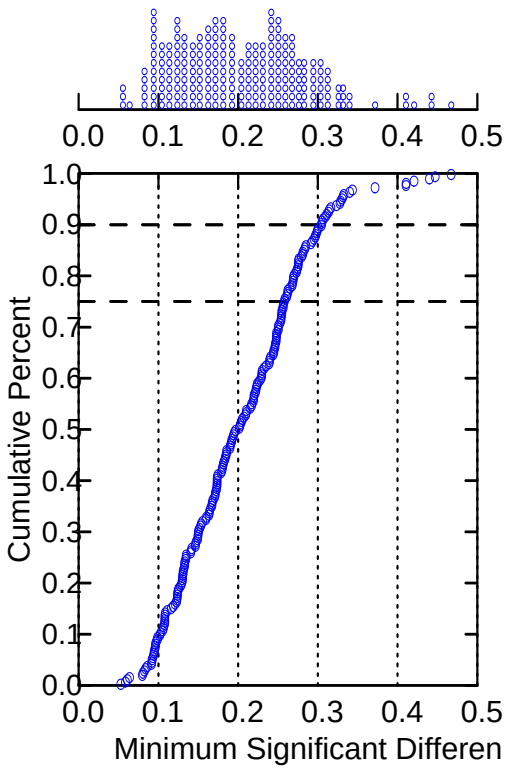
The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75^{th}, 90^{th}) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

Table B-1. Monte Carlo simulation results for chronic *Ceriodaphnia* reproduction in the EPA multi-concentration WET test method indicating the percentage of tests deemed toxic (this page) or non-toxic (next page) given different b values as a function of mean effect levels of the effluent (5% [0.5], 10% [0.10], 15% [0.15], 20% [0.20], 25% [0.25], 30% [0.30], and 35% [0.35] and within-test variability at 0~10th, 10~25th, 25~50th, 50~75th, 75~85th, and 85~95th percentile of s observed in controls and effluent concentrations in actual tests). Control mean varied between 10~90th percentile of observed offspring/female. The within-test variance was determined using a ratio between 25th to 75th percentile of observed variance ratio between control and effluent concentrations.

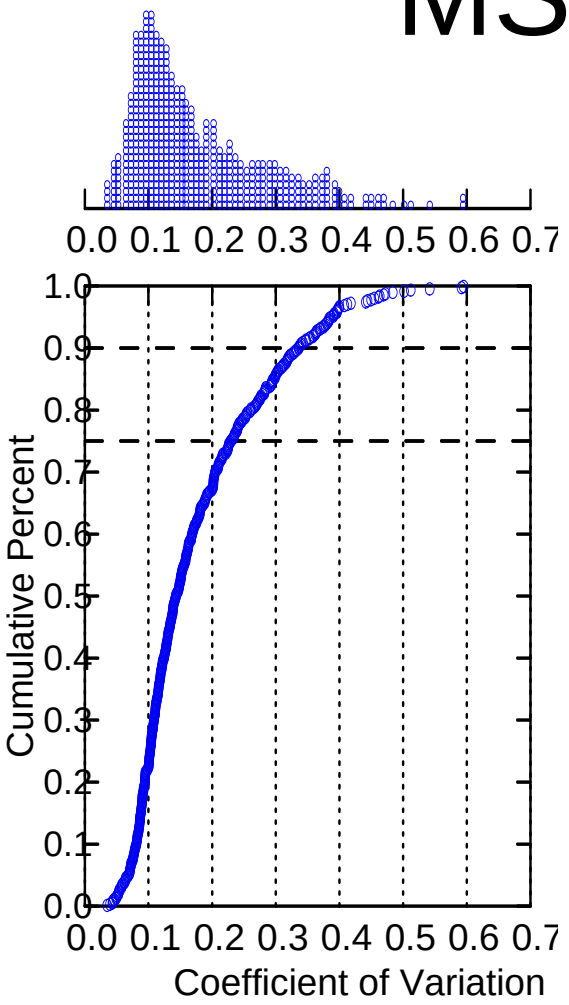
Mean Difference	C.V. Range	Result	NOEC	b=0.75	b=0.7	b=0.68	b=0.63	C.V. percentile
0.05	(3.5~8.2%)	Toxic	53.9%	0.0%	0.0%	0.0%	0.0%	0~10th
0.05	(8.2~10.4%)	Toxic	0.2%	0.0%	0.0%	0.0%	0.0%	10~25th
0.05	(10.4~14.8%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	25~50th
0.05	(14.8~24.3%)	Toxic	0.0%	1.1%	0.0%	0.0%	0.0%	50~75th
0.05	(24.3~30%)	Toxic	0.0%	25.6%	0.0%	0.0%	0.0%	75~85th
0.05	(30~40%)	Toxic	0.0%	64.4%	26.2%	14.6%	0.0%	85~95th
0.1	(3.5~8.2%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.1	(8.2~10.4%)	Toxic	94.6%	0.0%	0.0%	0.0%	0.0%	10~25th
0.1	(10.4~14.8%)	Toxic	57.5%	0.0%	0.0%	0.0%	0.0%	25~50th
0.1	(14.8~24.3%)	Toxic	5.5%	23.4%	0.4%	0.0%	0.0%	50~75th
0.1	(24.3~30%)	Toxic	0.0%	75.2%	25.6%	10.4%	0.0%	75~85th
0.1	(30~40%)	Toxic	0.0%	1.0%	61.0%	44.5%	12.5%	85~95th
0.15	(3.5~8.2%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.15	(8.2~10.4%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	10~25th
0.15	(10.4~14.8%)	Toxic	99.7%	22.6%	0.0%	0.0%	0.0%	25~50th
0.15	(14.8~24.3%)	Toxic	58.8%	83.2%	19.4%	8.2%	0.0%	50~75th
0.15	(24.3~30%)	Toxic	3.4%	100.0%	68.7%	45.9%	6.6%	75~85th
0.15	(30~40%)	Toxic	0.0%	100.0%	100.0%	8.9%	39.2%	85~95th
0.2	(3.5~8.2%)	Toxic	100.0%	29.8%	0.0%	0.0%	0.0%	0~10th
0.2	(8.2~10.4%)	Toxic	100.0%	87.9%	0.0%	0.0%	0.0%	10~25th
0.2	(10.4~14.8%)	Toxic	100.0%	100.0%	20.5%	2.0%	0.0%	25~50th
0.2	(14.8~24.3%)	Toxic	91.9%	100.0%	81.6%	0.499	6.7%	50~75th
0.2	(24.3~30%)	Toxic	56.1%	100.0%	100.0%	100.0%	41.9%	75~85th
0.2	(30~40%)	Toxic	11.3%	100.0%	100.0%	100.0%	85.8%	85~95th
0.25	(3.5~8.2%)	Toxic	100.0%	100.0%	28.1%	2.0%	0.0%	0~10th
0.25	(8.2~10.4%)	Toxic	100.0%	100.0%	84.6%	31.1%	0.0%	10~25th
0.25	(10.4~14.8%)	Toxic	100.0%	100.0%	100.0%	100.0%	1.5%	25~50th
0.25	(14.8~24.3%)	Toxic	100.0%	100.0%	100.0%	100.0%	45.8%	50~75th
0.25	(24.3~30%)	Toxic	87.6%	100.0%	100.0%	100.0%	99.7%	75~85th
0.25	(30~40%)	Toxic	52.6%	100.0%	100.0%	100.0%	100.0%	85~95th
0.3	(3.5~8.2%)	Toxic	100.0%	100.0%	100.0%	100.0%	0.9%	0~10th
0.3	(8.2~10.4%)	Toxic	100.0%	100.0%	100.0%	100.0%	28.8%	10~25th
0.3	(10.4~14.8%)	Toxic	100.0%	100.0%	100.0%	100.0%	70.4%	25~50th
0.3	(14.8~24.3%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	50~75th
0.3	(24.3~30%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	75~85th
0.3	(30~40%)	Toxic	77.9%	100.0%	100.0%	100.0%	100.0%	85~95th
0.35	(3.5~8.2%)	Toxic	100.0%	100.0%	100.0%	100.0%	99.8%	0~10th
0.35	(8.2~10.4%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	10~25th
0.35	(10.4~14.8%)	Toxic	100.0%	100.0%	100.0 %	100.0%	100.0 %	25~50th
0.35	(14.8~24.3%)	Toxic	100.0%	100.0%	100.0 %	100.0%	100.0 %	50~75th
0.35	(24.3~30%)	Toxic	100.0%	100.0%	100.0 %	100.0%	100.0 %	75~85th
0.35	(30~40%)	Toxic	95.1%	100.0%	100.0 %	100.0%	100.0 %	85~95th

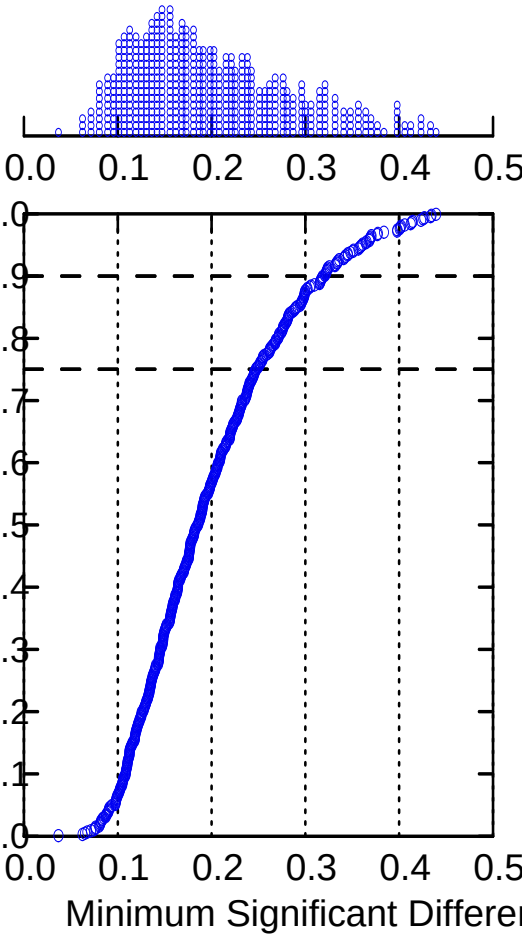
Ref Tox CV and MSD

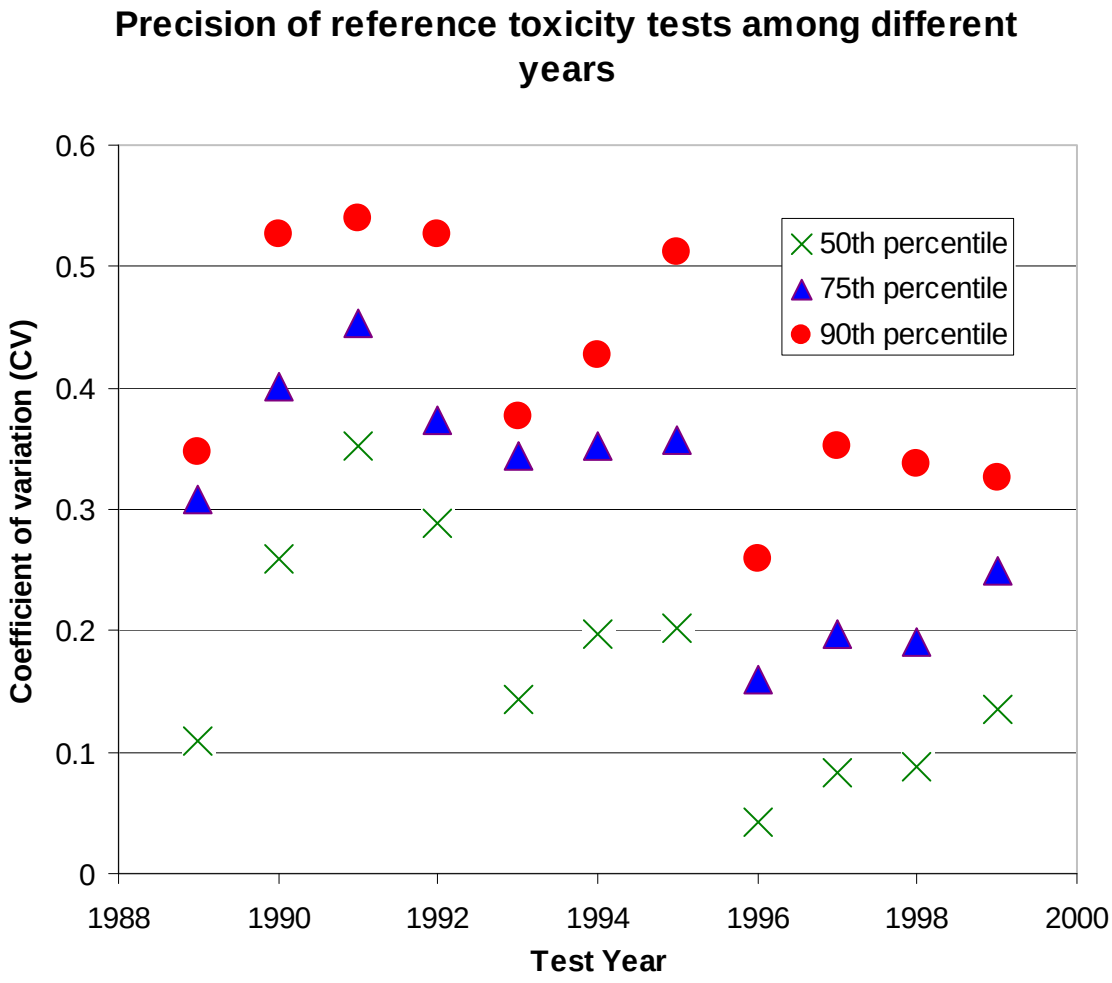




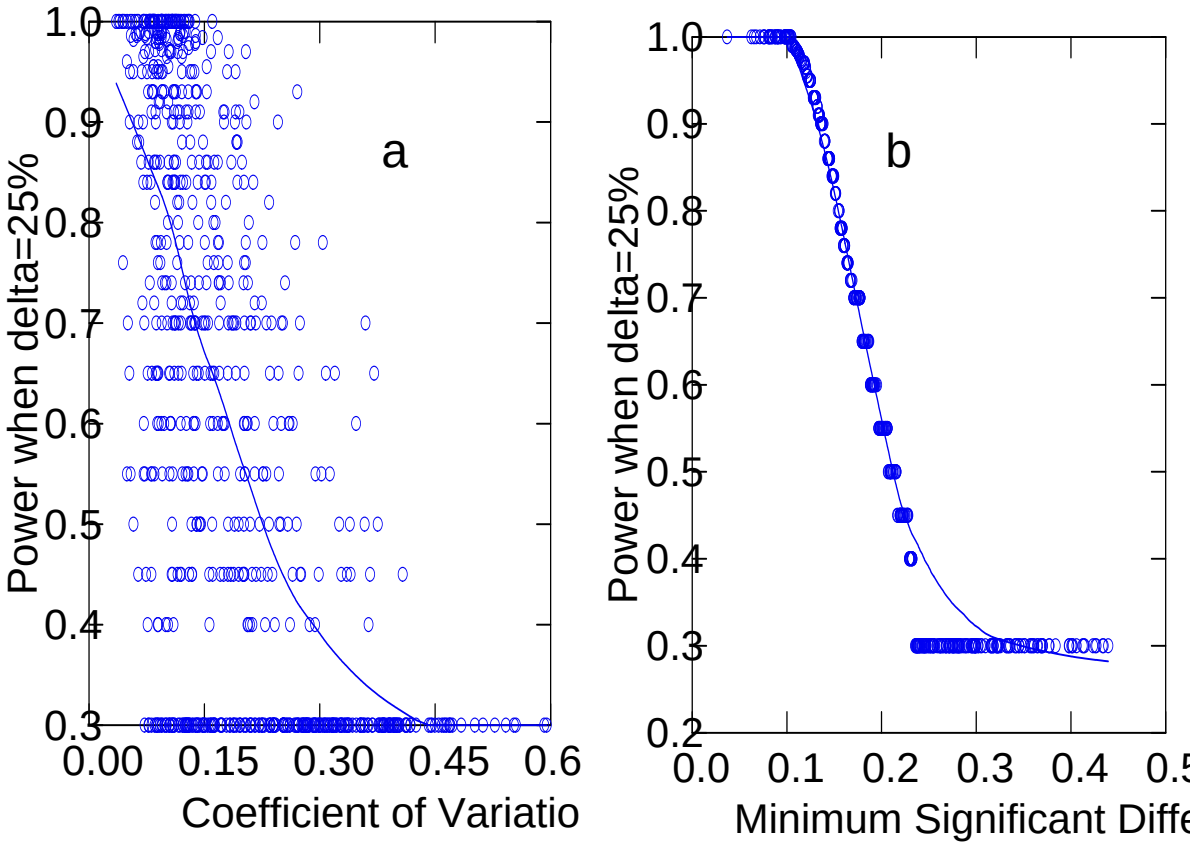
Effluent Test CV and MSD





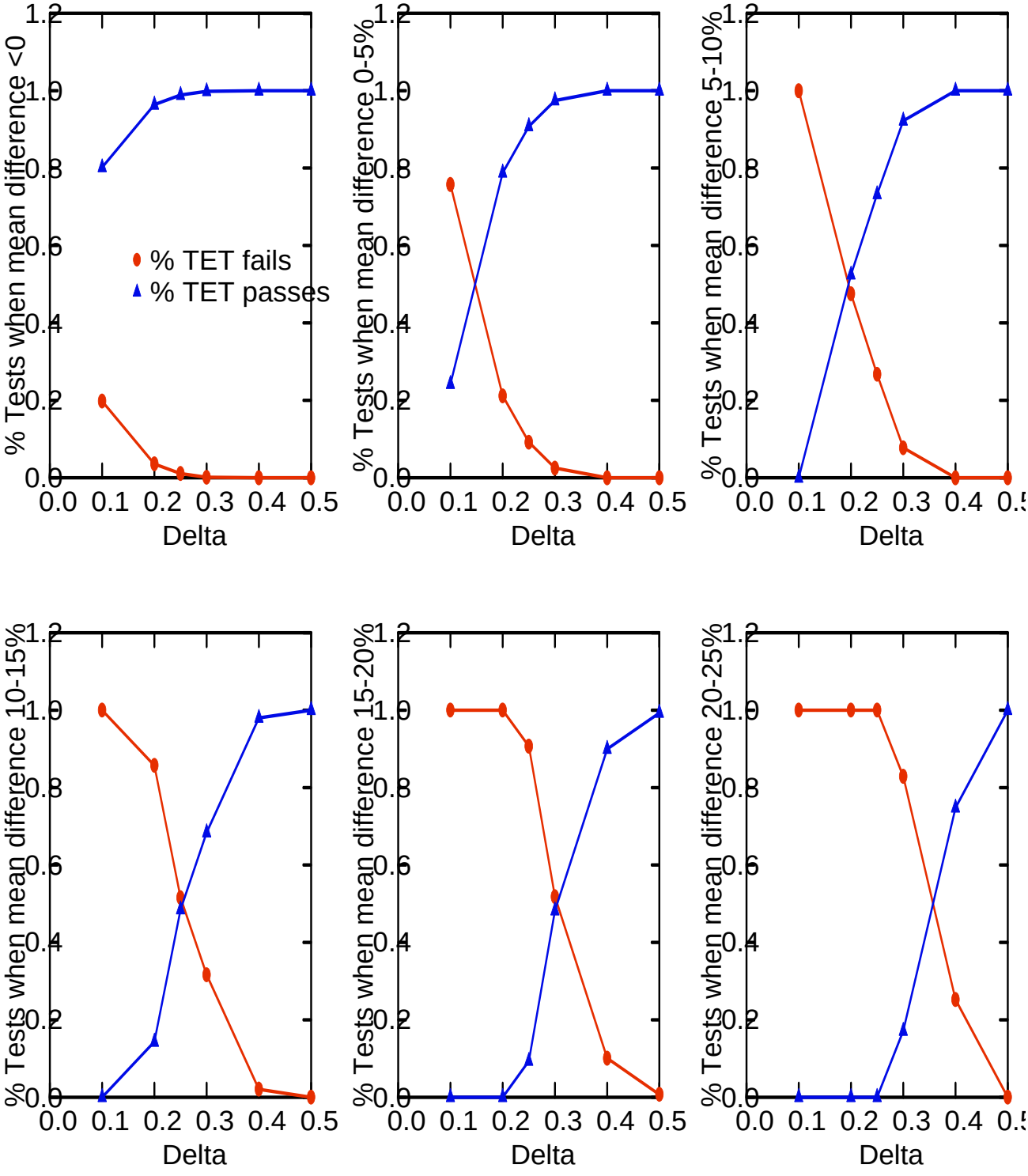

Precision of reference toxicity tests among different years
Coefficient of variation (CV)
0.6
0.5
0.4
0.3
0.2
0.1
0
1988
1990
1992
1994
1996
1998
2000
Test Year
50th percentile
75th percentile
90th percentile

Relationship between Power and either CV or MSD



Detailed Analysis

Delta = 1-*b* where *b* is the proportional value used in TST analysis



APPENDIX C

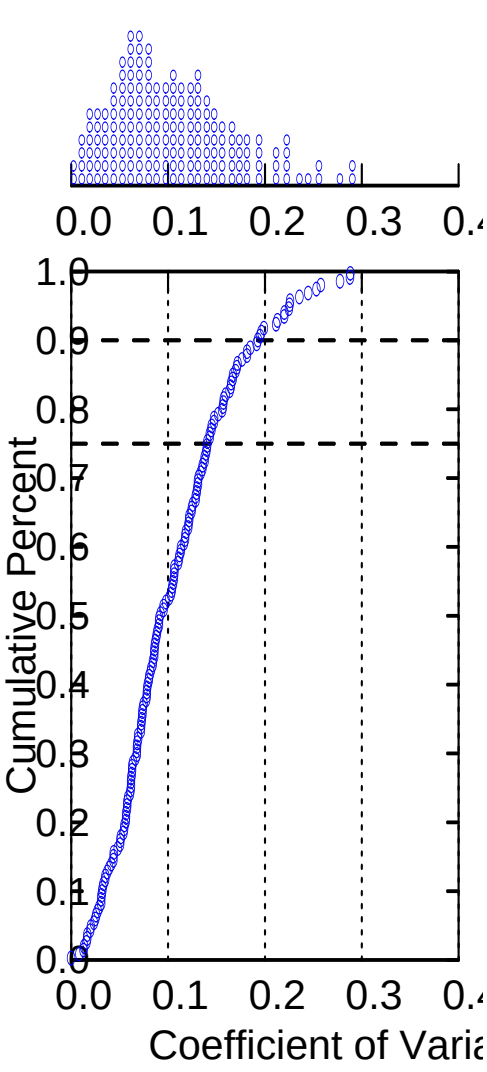
Pimephales promelas Chronic Growth Analyses

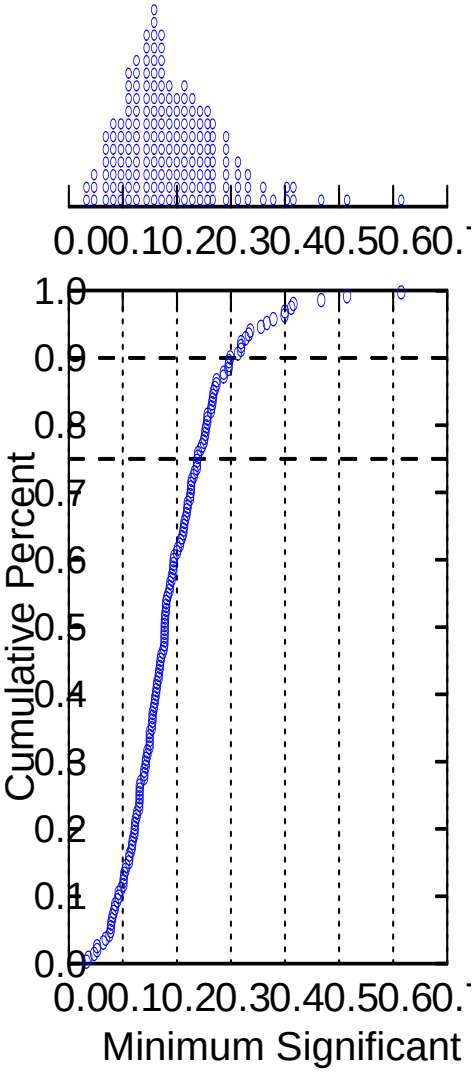
The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75^{th}, 90^{th}) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

Table C-1. Simulation results for chronic *P. promelas* growth endpoint for tests deemed toxic. See Table B-1 for simulation method information.

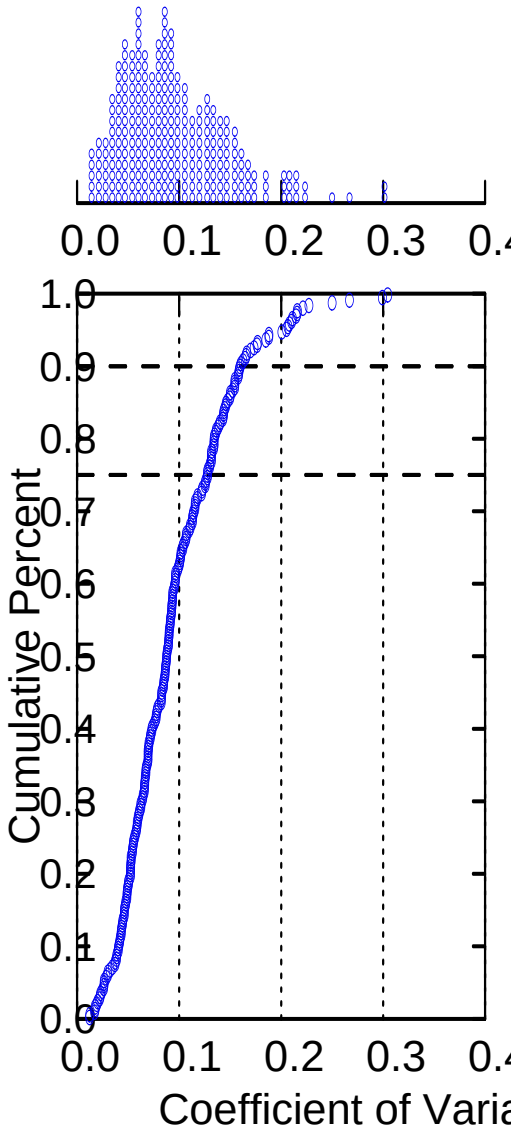
Mean Difference	C.V. Range	Result	NOEC	b=0.75	b=0.7	b=0.65	b=0.6	C.V. percentile
0.05	(2.6~3.9%)	Toxic	47.4%	0.0%	0.0%	0.0%	0.0%	0~10th
0.05	(3.9~5.8%)	Toxic	2.9%	0.0%	0.0%	0.0%	0.0%	10~25th
0.05	(5.8~8.9%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	25~50th
0.05	(8.9~13.2%)	Toxic	0.0%	15.9%	0.0%	0.0%	0.0%	50~75th
0.05	(13.2~15%)	Toxic	0.0%	50.8%	12.4%	0.0%	0.0%	75~85th
0.05	(15~25%)	Toxic	0.0%	88.7%	62.1%	33.8%	12.1%	85~95th
0.1	(2.6~3.9%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.1	(3.9~5.8%)	Toxic	84.9%	0.0%	0.0%	0.0%	0.0%	10~25th
0.1	(5.8~8.9%)	Toxic	29.6%	4.9%	0.0%	0.0%	0.0%	25~50th
0.1	(8.9~13.2%)	Toxic	0.0%	53.3%	11.2%	0.0%	0.0%	50~75th
0.1	(13.2~15%)	Toxic	0.0%	89.7%	46.7%	8.9%	0.0%	75~85th
0.1	(15~25%)	Toxic	0.0%	100.0%	87.1%	59.4%	29.9%	85~95th
0.15	(2.6~3.9%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.15	(3.9~5.8%)	Toxic	100.0%	3.6%	0.0%	0.0%	0.0%	10~25th
0.15	(5.8~8.9%)	Toxic	85.1%	53.4%	2.8%	0.0%	0.0%	25~50th
0.15	(8.9~13.2%)	Toxic	27.1%	97.0%	52.3%	10.2%	0.0%	50~75th
0.15	(13.2~15%)	Toxic	0.3%	100.0%	82.8%	41.8%	6.6%	75~85th
0.15	(15~25%)	Toxic	0.0%	100.0%	100.0%	85.1%	53.6%	85~95th
0.2	(2.6~3.9%)	Toxic	100.0%	36.5%	0.0%	0.0%	0.0%	0~10th
0.2	(3.9~5.8%)	Toxic	100.0%	87.7%	3.8%	0.0%	0.0%	10~25th
0.2	(5.8~8.9%)	Toxic	100.0%	100.0%	51.5%	2.2%	0.0%	25~50th
0.2	(8.9~13.2%)	Toxic	68.3%	100.0%	96.3%	48.3%	9.2%	50~75th
0.2	(13.2~15%)	Toxic	32.6%	100.0%	100.0%	84.1%	40.2%	75~85th
0.2	(15~25%)	Toxic	3.2%	100.0%	100.0%	99.7%	84.4%	85~95th
0.25	(2.6~3.9%)	Toxic	100.0%	100.0%	32.2%	0.0%	0.0%	0~10th
0.25	(3.9~5.8%)	Toxic	100.0%	100.0%	85.7%	1.2%	0.0%	10~25th
0.25	(5.8~8.9%)	Toxic	100.0%	100.0%	100.0%	47.5%	1.3%	25~50th
0.25	(8.9~13.2%)	Toxic	94.8%	100.0%	100.0%	95.1%	40.1%	50~75th
0.25	(13.2~15%)	Toxic	67.3%	100.0%	100.0%	100.0%	74.9%	75~85th
0.25	(15~25%)	Toxic	18.7%	100.0%	100.0%	100.0%	98.8%	85~95th
0.3	(2.6~3.9%)	Toxic	100.0%	100.0%	100.0%	28.9%	0.0%	0~10th
0.3	(3.9~5.8%)	Toxic	100.0%	100.0%	100.0%	83.7%	0.4%	10~25th
0.3	(5.8~8.9%)	Toxic	100.0%	100.0%	100.0%	100.0%	44.7%	25~50th
0.3	(8.9~13.2%)	Toxic	100.0%	100.0%	100.0%	100.0%	94.5%	50~75th
0.3	(13.2~15%)	Toxic	93.1%	100.0%	100.0%	100.0%	100.0%	75~85th
0.3	(15~25%)	Toxic	41.1%	100.0%	100.0%	100.0%	100.0%	85~95th
0.35	(2.6~3.9%)	Toxic	100.0%	100.0%	100.0%	100.0%	23.0%	0~10th
0.35	(3.9~5.8%)	Toxic	100.0%	100.0%	100.0%	100.0%	84.2%	10~25th
0.35	(5.8~8.9%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	25~50th
0.35	(8.9~13.2%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	50~75th
0.35	(13.2~15%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	75~85th
0.35	(15~25%)	Toxic	66.2%	100.0%	100.0%	100.0%	100.0%	85~95th

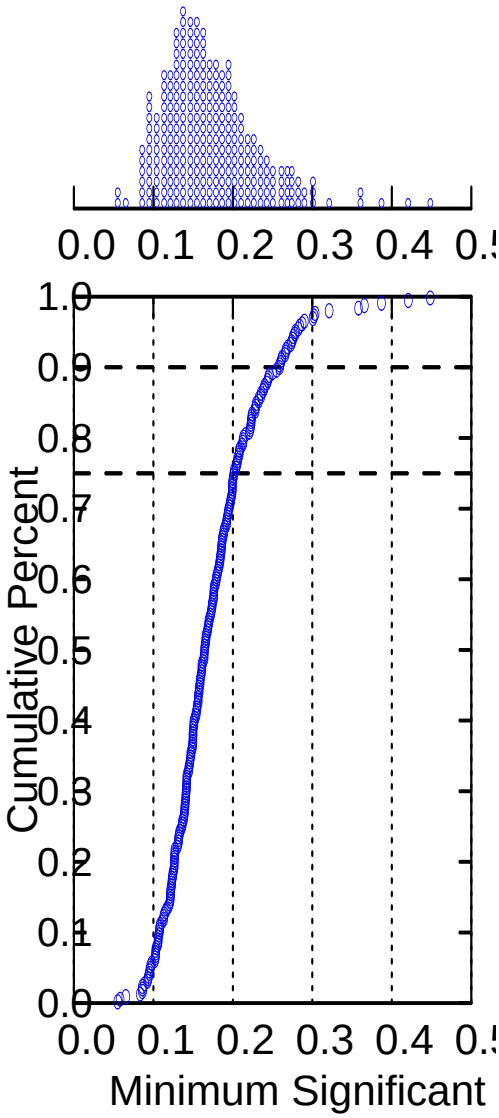
Reference Toxicity tests



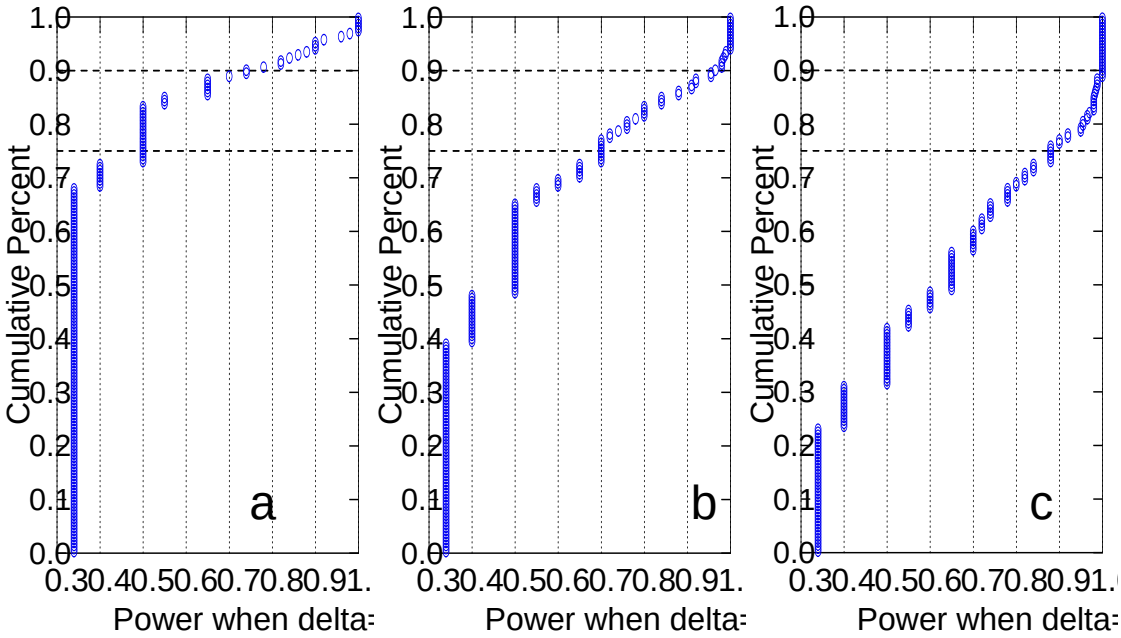


Effluent tests

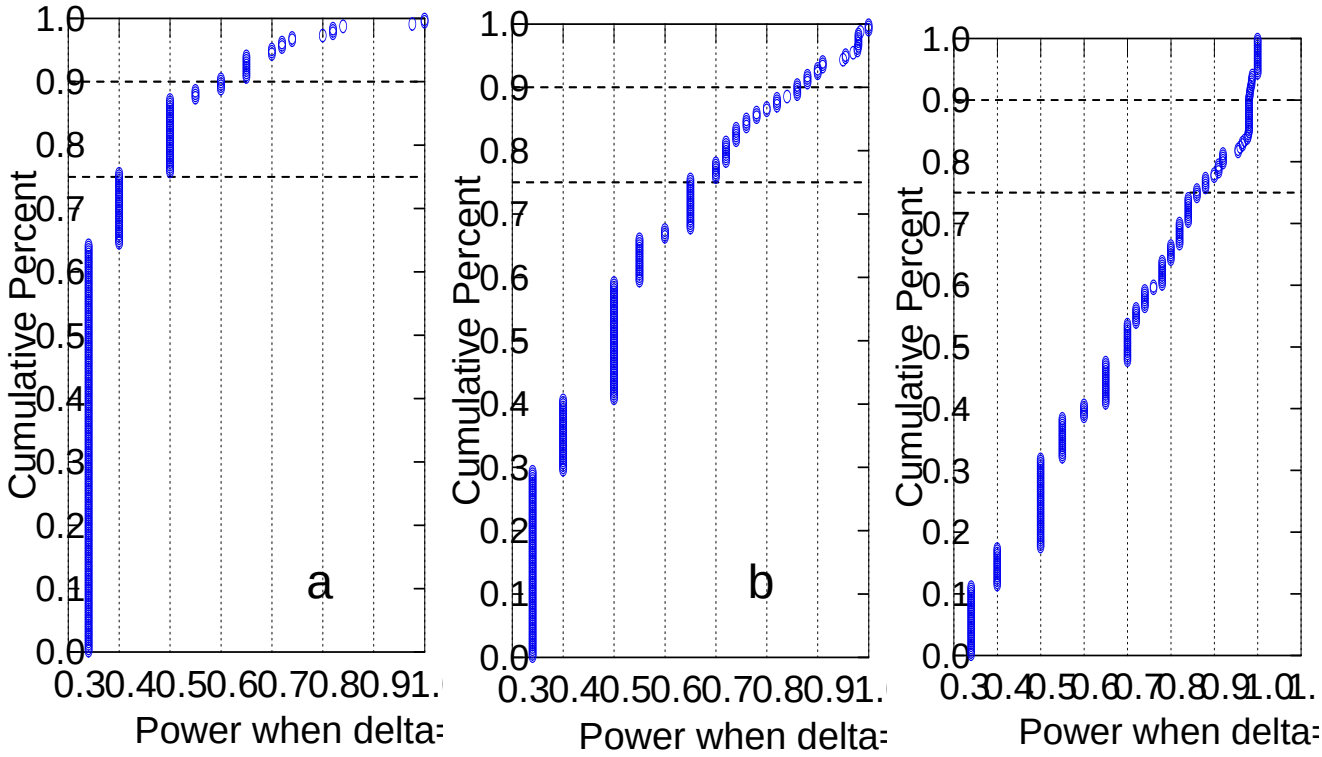




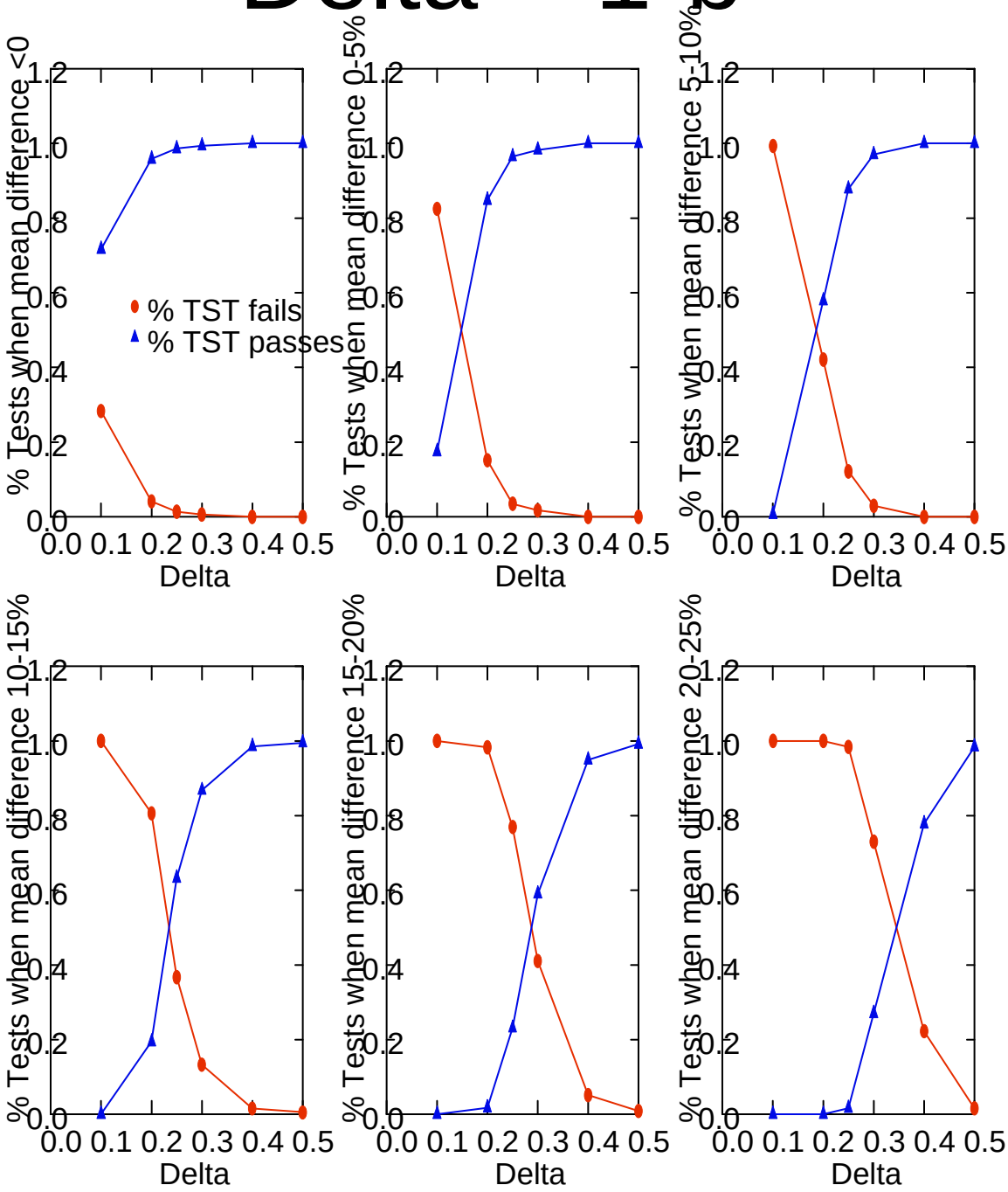
Reference toxicity Tests: power as a function of delta = 1-b



Effluent toxicity tests: power as a function of delta = 1-b



Detailed Analyses; Delta = 1-b



APPENDIX D

Americamysis bahia Chronic Growth Analyses

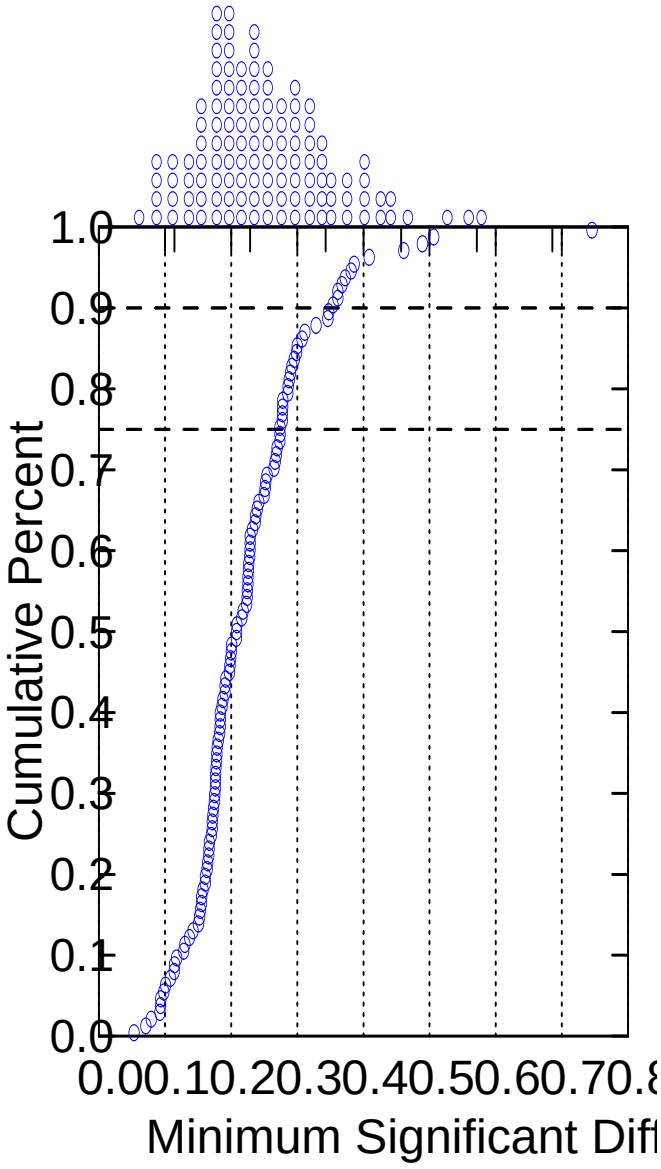
The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75^{th}, 90^{th}) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

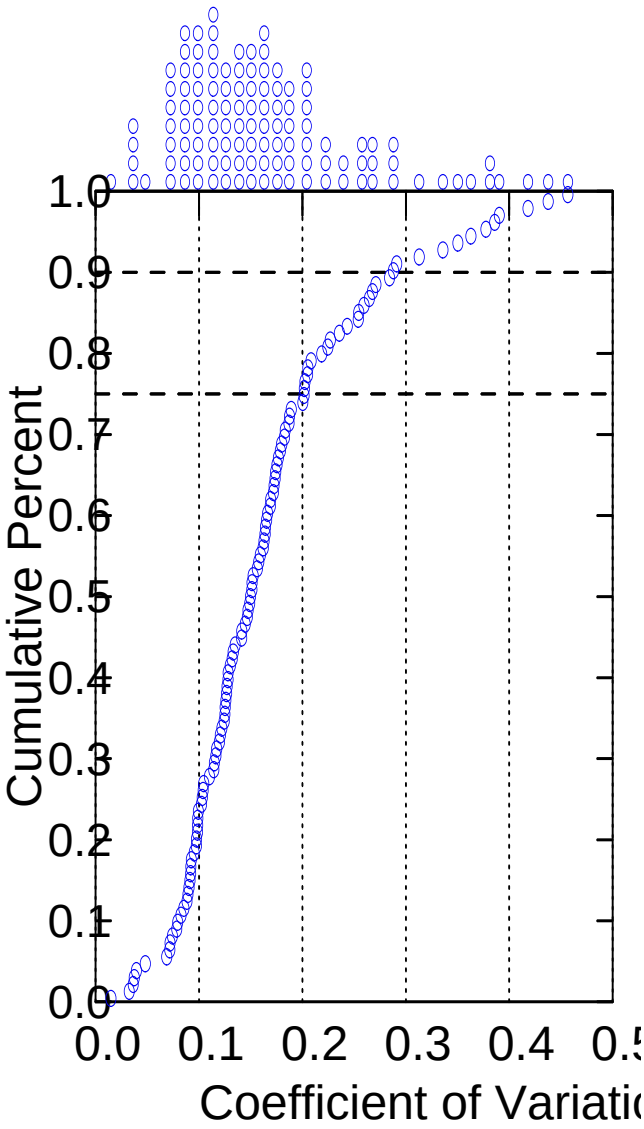
Table D-1. Simulation results for chronic mysid growth endpoints for tests deemed toxic. See Table A-1 for simulation method information.

Mean Difference	C.V. Range	Result	NOEC	b=0.75	b=0.7	b=0.65	b=0.6	C.V. percentile
0.05	(7~8%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.05	(8~10%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	10~25th
0.05	(10~14%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	25~50th
0.05	(14~18%)	Toxic	0.0%	0.0%	0.0%	0.0%	0.0%	50~75th
0.05	(18~25%)	Toxic	0.0%	25.0%	0.3%	0.0%	0.0%	75~85th
0.05	(25~35%)	Toxic	0.0%	87.8%	43.0%	7.7%	0.0%	85~95th
0.1	(7~8%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.1	(8~10%)	Toxic	85.1%	0.0%	0.0%	0.0%	0.0%	10~25th
0.1	(10~14%)	Toxic	25.2%	0.0%	0.0%	0.0%	0.0%	25~50th
0.1	(14~18%)	Toxic	0.0%	21.4%	0.0%	0.0%	0.0%	50~75th
0.1	(18~25%)	Toxic	0.0%	80.6%	19.5%	0.0%	0.0%	75~85th
0.1	(25~35%)	Toxic	0.0%	100.0%	83.8%	39.8%	6.9%	85~95th
0.15	(7~8%)	Toxic	100.0%	0.0%	0.0%	0.0%	0.0%	0~10th
0.15	(8~10%)	Toxic	100.0%	1.3%	0.0%	0.0%	0.0%	10~25th
0.15	(10~14%)	Toxic	96.4%	46.7%	0.0%	0.0%	0.0%	25~50th
0.15	(14~18%)	Toxic	51.1%	96.6%	17.7%	0.0%	0.0%	50~75th
0.15	(18~25%)	Toxic	4.6%	100.0%	78.5%	17.0%	0.0%	75~85th
0.15	(25~35%)	Toxic	0.0%	100.0%	100.0%	83.3%	35.2%	85~95th
0.2	(7~8%)	Toxic	100.0%	93.9%	0.0%	0.0%	0.0%	0~10th
0.2	(8~10%)	Toxic	100.0%	100.0%	0.1%	0.0%	0.0%	10~25th
0.2	(10~14%)	Toxic	100.0%	100.0%	40.9%	0.0%	0.0%	25~50th
0.2	(14~18%)	Toxic	96.5%	100.0%	95.0%	16.3%	0.0%	50~75th
0.2	(18~25%)	Toxic	49.9%	100.0%	100.0%	73.1%	15.7%	75~85th
0.2	(25~35%)	Toxic	1.6%	100.0%	100.0%	100.0%	79.3%	85~95th
0.25	(7~8%)	Toxic	100.0%	100.0%	90.4%	0.0%	0.0%	0~10th
0.25	(8~10%)	Toxic	100.0%	100.0%	100.0%	0.0%	0.0%	10~25th
0.25	(10~14%)	Toxic	100.0%	100.0%	100.0%	37.3%	0.0%	25~50th
0.25	(14~18%)	Toxic	100.0%	100.0%	100.0%	92.1%	11.0%	50~75th
0.25	(18~25%)	Toxic	87.9%	100.0%	100.0%	100.0%	70.7%	75~85th
0.25	(25~35%)	Toxic	25.4%	100.0%	100.0%	100.0%	100.0%	85~95th
0.3	(7~8%)	Toxic	100.0%	100.0%	100.0%	84.0%	0.0%	0~10th
0.3	(8~10%)	Toxic	100.0%	100.0%	100.0%	100.0%	0.0%	10~25th
0.3	(10~14%)	Toxic	100.0%	100.0%	100.0%	100.0%	34.5%	25~50th
0.3	(14~18%)	Toxic	100.0%	100.0%	100.0%	100.0%	88.5%	50~75th
0.3	(18~25%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	75~85th
0.3	(25~35%)	Toxic	66.3%	100.0%	100.0%	100.0%	100.0%	85~95th
0.35	(7~8%)	Toxic	100.0%	100.0%	100.0%	100.0%	79.2%	0~10th
0.35	(8~10%)	Toxic	100.0%	100.0%	100.0%	100.0%	99.7%	10~25th
0.35	(10~14%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	25~50th
0.35	(14~18%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	50~75th
0.35	(18~25%)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	75~85th
0.35	(25~35%)	Toxic	89.0%	100.0%	100.0%	100.0%	100.0%	85~95th

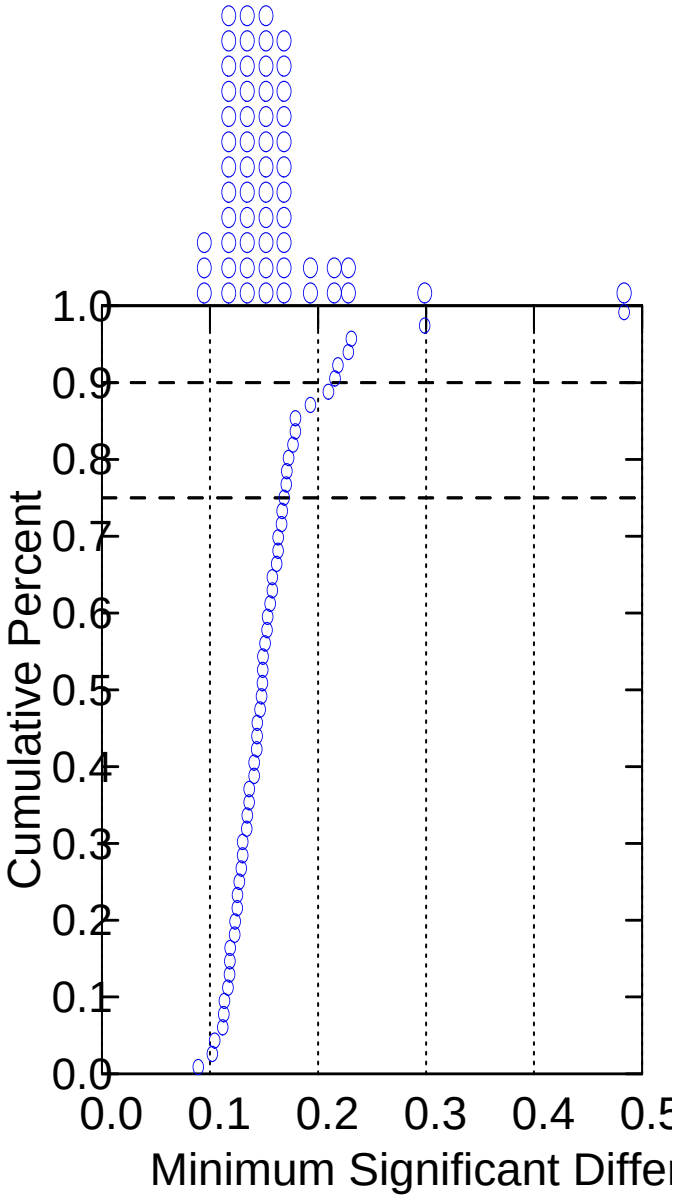
000107

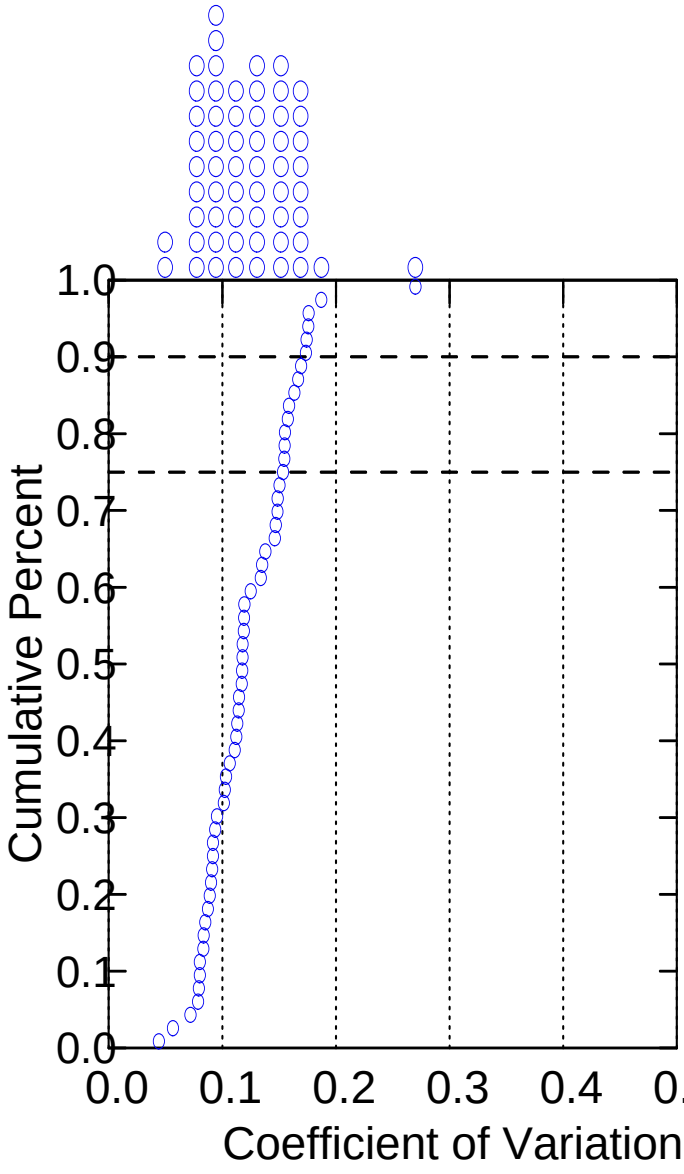
MSD and CV for Ref Tox (111 tests)



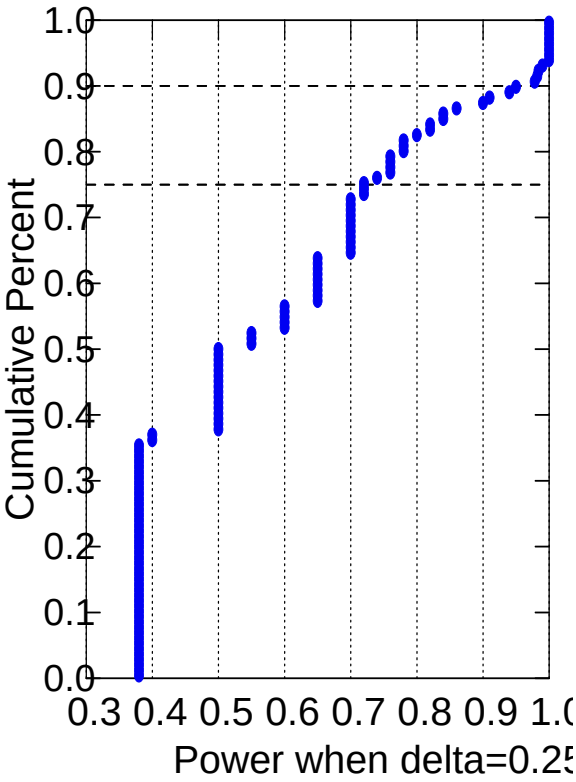


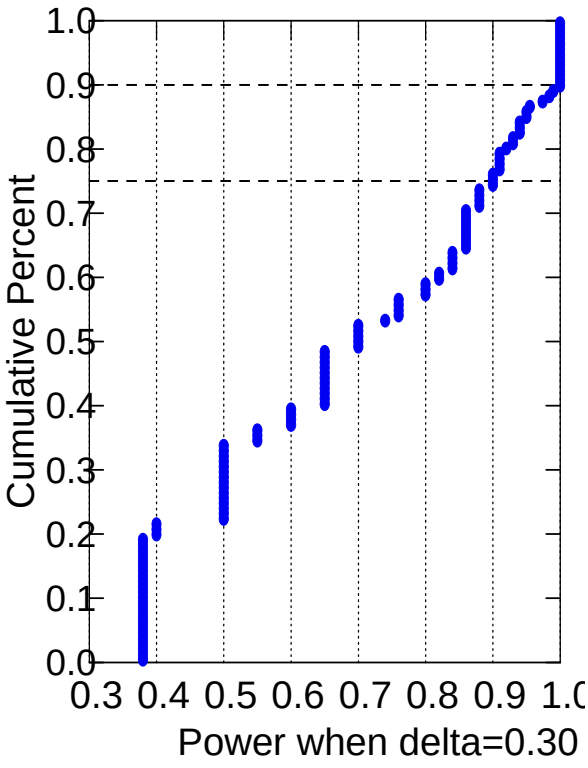
Effluent MSD and CV (58 Tests)



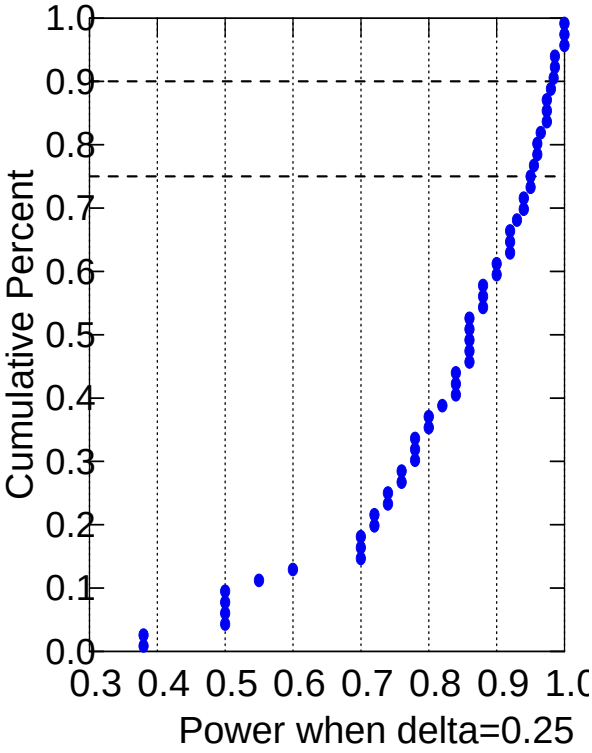


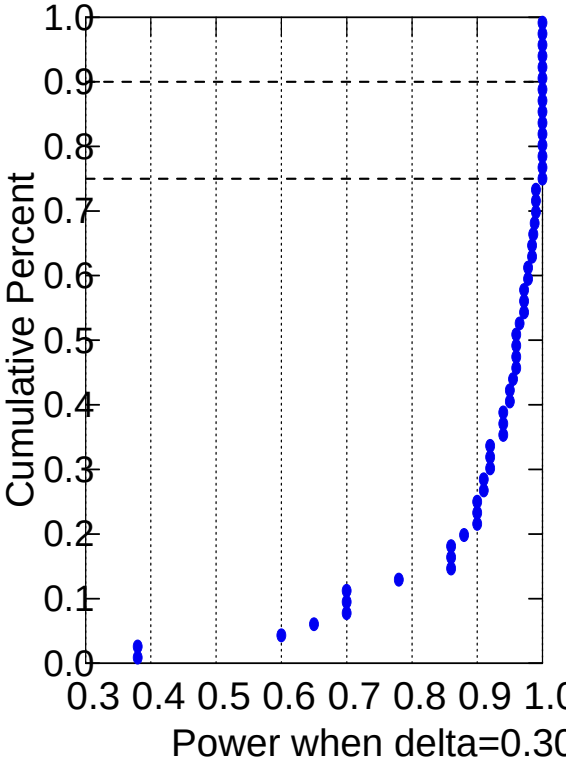
Ref Tox Power as function of delta = 1-b



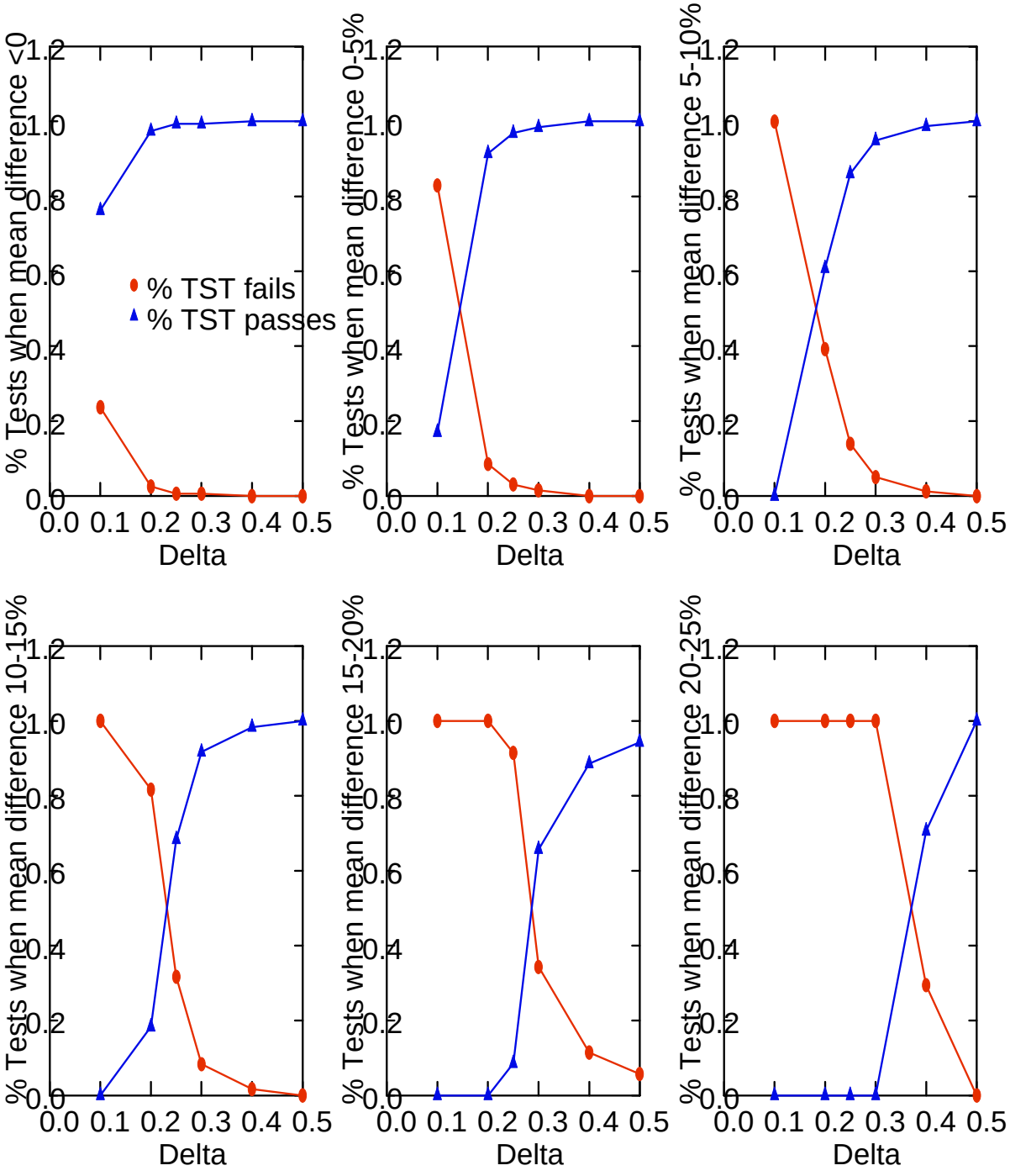


Effluent Power as a function of delta = 1-b





Detailed Analysis
Delta = 1-b



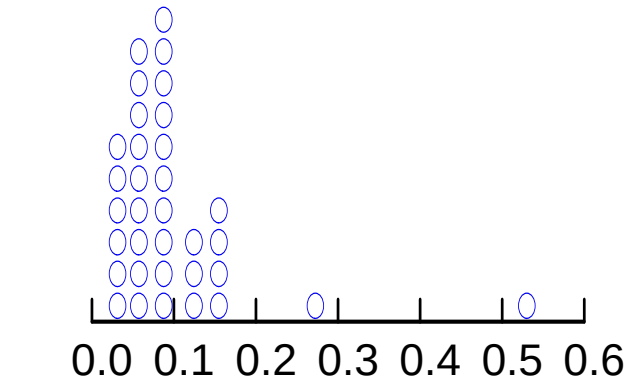
Appendix E Echinoderm Chronic WET Analysis

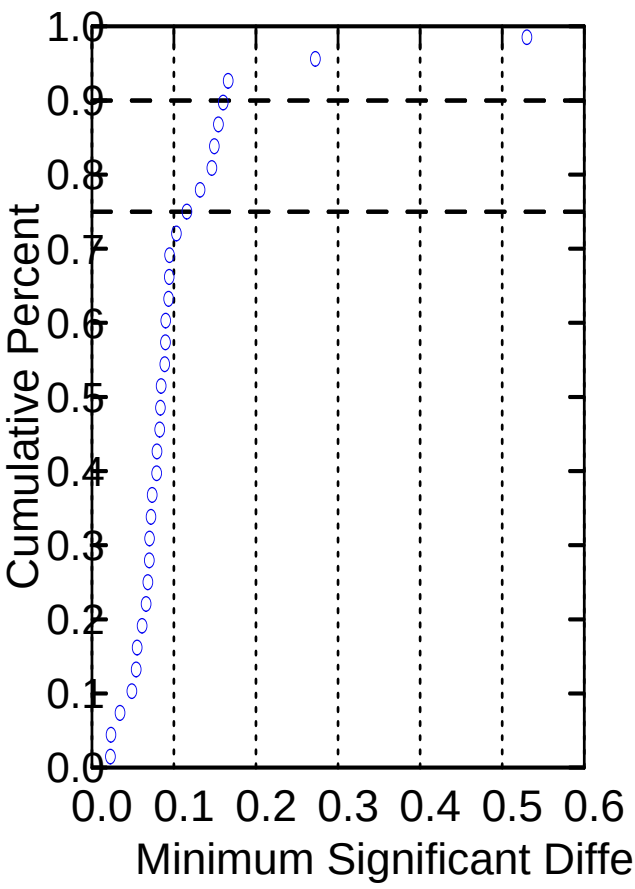
The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75^{th}, 90^{th}) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

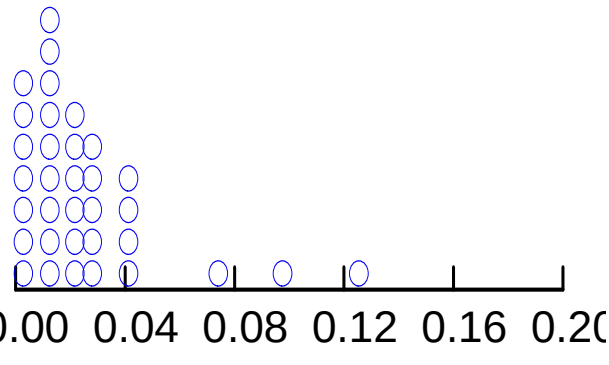
Table E-1. Simulation results for Sea Urchin fertilization endpoint for tests deemed non-toxic. See Table B-1 for simulation method information

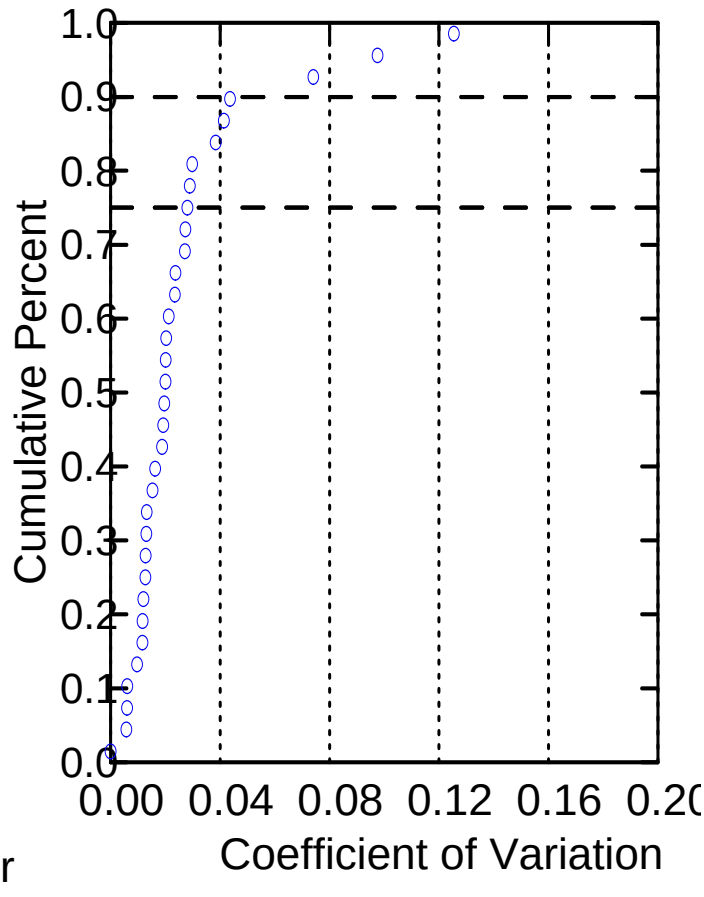
Mean Difference	C.V. Range	Result	NOEC	b=0.85	b=0.8	b=0.75	b=0.7	C.V. percentile
0.05	(0~0.5%)	Non-Toxic	0.00%	100.00%	100.00%	100.00%	100.00%	0~10th
0.05	(0.5~1.2%)	Non-Toxic	28.60%	93.40%	99.90%	100.00%	100.00%	10~25th
0.05	(1.2~2.7%)	Non-Toxic	67.50%	50.30%	80.20%	95.60%	99.80%	25~50th
0.05	(2.7~6.5%)	Non-Toxic	94.10%	16.20%	30.10%	43.70%	59.70%	50~75th
0.05	(6.5~10%)	Non-Toxic	99.90%	2.80%	14.60%	24.60%	32.40%	75~85th
0.05	(10~13.5%)	Non-Toxic	100.00%	0.00%	3.70%	12.70%	20.10%	85~95th
0.1	(0~0.5%)	Non-Toxic	0.00%	84.30%	100.00%	100.00%	100.00%	0~10th
0.1	(0.5~1.2%)	Non-Toxic	2.60%	48.00%	93.60%	100.00%	100.00%	10~25th
0.1	(1.2~2.7%)	Non-Toxic	41.10%	20.10%	48.60%	79.00%	96.60%	25~50th
0.1	(2.7~6.5%)	Non-Toxic	79.10%	2.40%	19.00%	34.70%	46.80%	50~75th
0.1	(6.5~10%)	Non-Toxic	95.10%	0.00%	2.50%	13.90%	23.60%	75~85th
0.1	(10~13.5%)	Non-Toxic	99.90%	0.00%	0.00%	5.00%	14.20%	85~95th
0.15	(0~0.5%)	Non-Toxic	0.00%	26.70%	87.90%	100.00%	100.00%	0~10th
0.15	(0.5~1.2%)	Non-Toxic	0.00%	5.40%	55.70%	94.50%	99.90%	10~25th
0.15	(1.2~2.7%)	Non-Toxic	24.70%	0.30%	21.00%	51.20%	80.10%	25~50th
0.15	(2.7~6.5%)	Non-Toxic	66.90%	0.00%	3.70%	20.30%	34.00%	50~75th
0.15	(6.5~10%)	Non-Toxic	87.10%	0.00%	0.00%	4.90%	15.00%	75~85th
0.15	(10~13.5%)	Non-Toxic	96.30%	0.00%	0.00%	0.80%	6.70%	85~95th
0.2	(0~0.5%)	Non-Toxic	0.00%	0.00%	41.70%	90.50%	100.00%	0~10th
0.2	(0.5~1.2%)	Non-Toxic	0.00%	0.00%	12.60%	65.30%	95.90%	10~25th
0.2	(1.2~2.7%)	Non-Toxic	8.40%	0.00%	2.40%	27.80%	60.60%	25~50th
0.2	(2.7~6.5%)	Non-Toxic	55.80%	0.00%	0.00%	7.20%	24.60%	50~75th
0.2	(6.5~10%)	Non-Toxic	80.40%	0.00%	0.00%	0.00%	5.40%	75~85th
0.2	(10~13.5%)	Non-Toxic	91.80%	0.00%	0.00%	0.00%	0.30%	85~95th
0.25	(0~0.5%)	Non-Toxic	0.00%	0.00%	0.00%	55.80%	96.20%	0~10th
0.25	(0.5~1.2%)	Non-Toxic	0.00%	0.00%	0.00%	22.80%	77.10%	10~25th
0.25	(1.2~2.7%)	Non-Toxic	2.00%	0.00%	0.00%	7.00%	31.70%	25~50th
0.25	(2.7~6.5%)	Non-Toxic	50.20%	0.00%	0.00%	0.60%	11.20%	50~75th
0.25	(6.5~10%)	Non-Toxic	72.00%	0.00%	0.00%	0.00%	0.60%	75~85th
0.25	(10~13.5%)	Non-Toxic	86.30%	0.00%	0.00%	0.00%	0.00%	85~95th
0.3	(0~0.5%)	Non-Toxic	0.00%	0.00%	0.00%	2.40%	72.00%	0~10th
0.3	(0.5~1.2%)	Non-Toxic	0.00%	0.00%	0.00%	0.10%	37.20%	10~25th
0.3	(1.2~2.7%)	Non-Toxic	0.30%	0.00%	0.00%	0.00%	15.10%	25~50th
0.3	(2.7~6.5%)	Non-Toxic	39.60%	0.00%	0.00%	0.00%	2.20%	50~75th
0.3	(6.5~10%)	Non-Toxic	67.40%	0.00%	0.00%	0.00%	0.00%	75~85th
0.3	(10~13.5%)	Non-Toxic	79.00%	0.00%	0.00%	0.00%	0.00%	85~95th
0.35	(0~0.5%)	Non-Toxic	0.00%	0.00%	0.00%	0.00%	18.20%	0~10th
0.35	(0.5~1.2%)	Non-Toxic	0.00%	0.00%	0.00%	0.00%	4.60%	10~25th
0.35	(1.2~2.7%)	Non-Toxic	0.00%	0.00%	0.00%	0.00%	0.00%	25~50th
0.35	(2.7~6.5%)	Non-Toxic	32.20%	0.00%	0.00%	0.00%	0.00%	50~75th
0.35	(6.5~10%)	Non-Toxic	66.40%	0.00%	0.00%	0.00%	0.00%	75~85th
0.35	(10~13.5%)	Non-Toxic	75.60%	0.00%	0.00%	0.00%	0.00%	85~95th

MSD and CV for Ref Tox (94 tests)

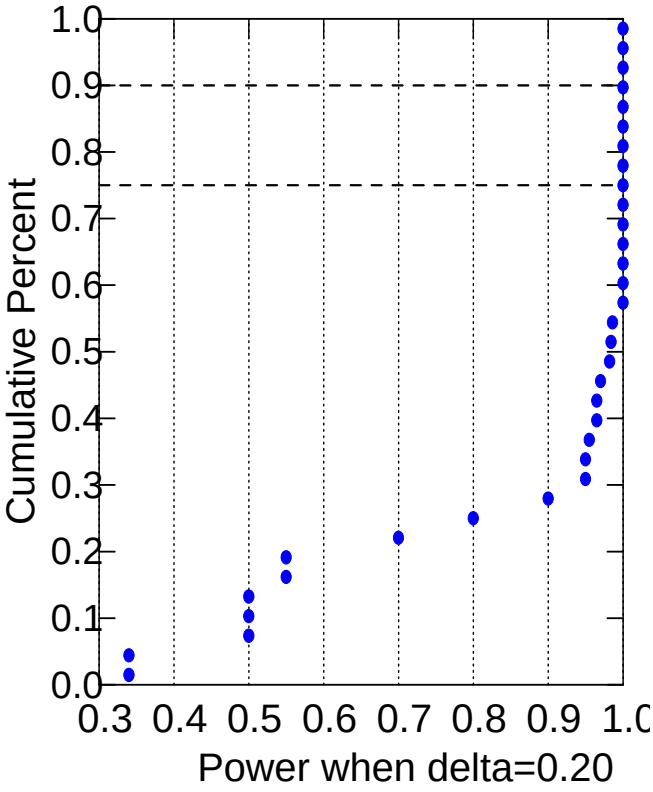


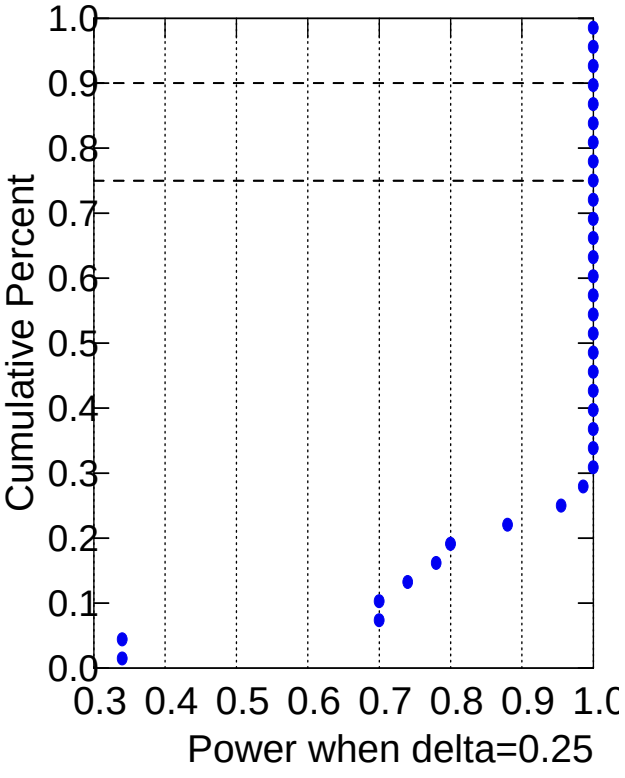






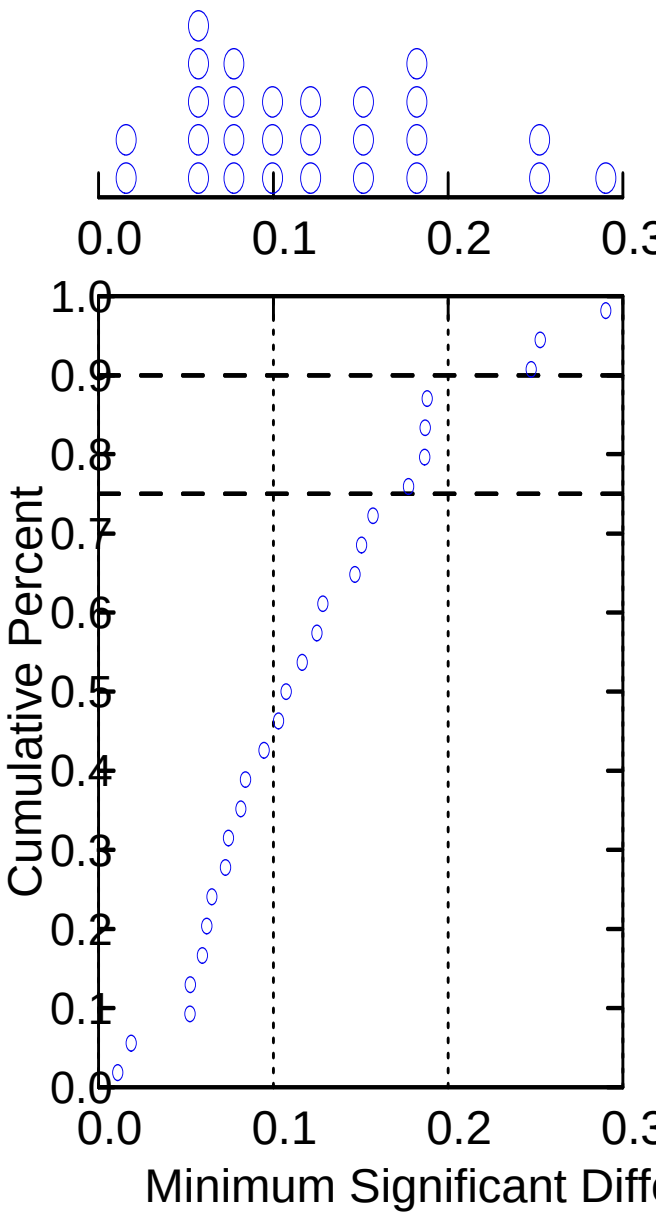
Ref Tox Test Power as function of delta = 1-b

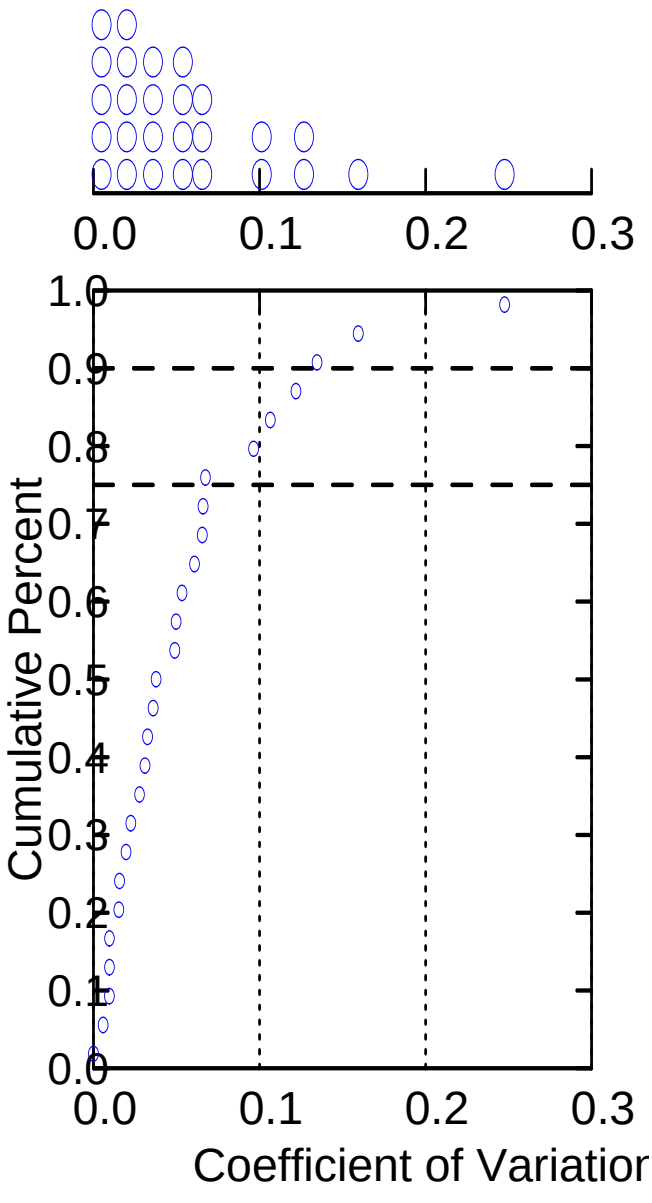




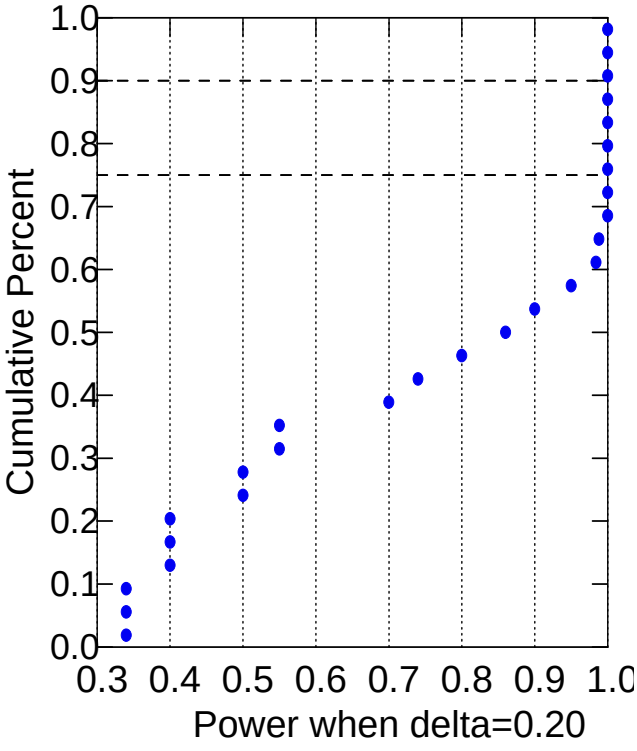
000116

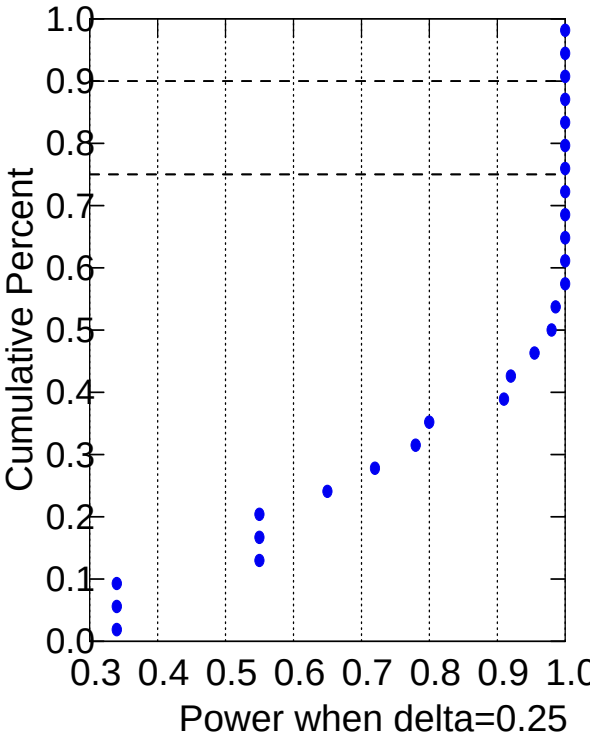
Effluent MSD and CV (27 tests)



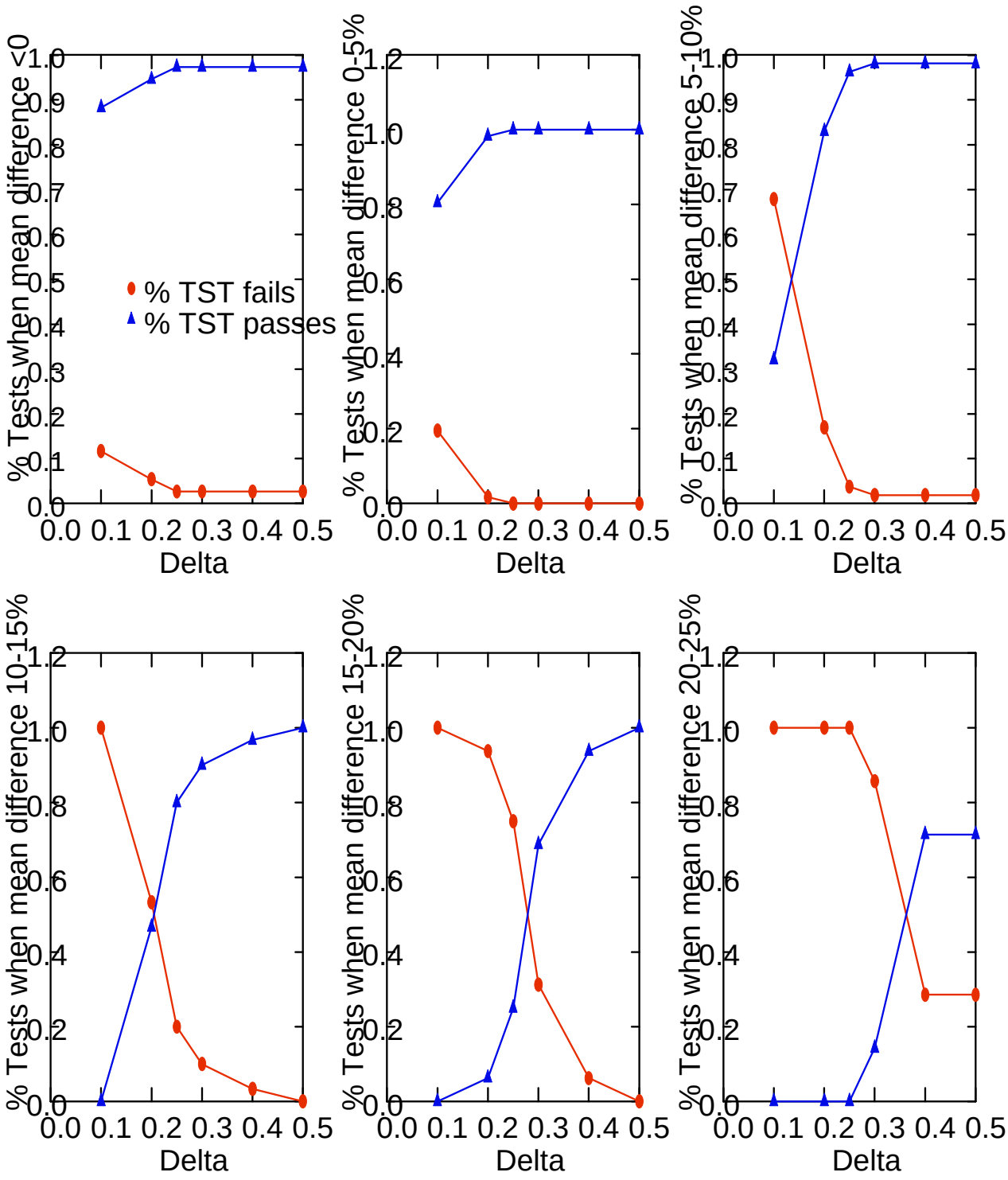


Effluent Test Power as function of delta = 1-b





Detailed Analysis
Delta = 1-b



APPENDIX F
Red Abalone Chronic Analysis

The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75^{th}, 90^{th}) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

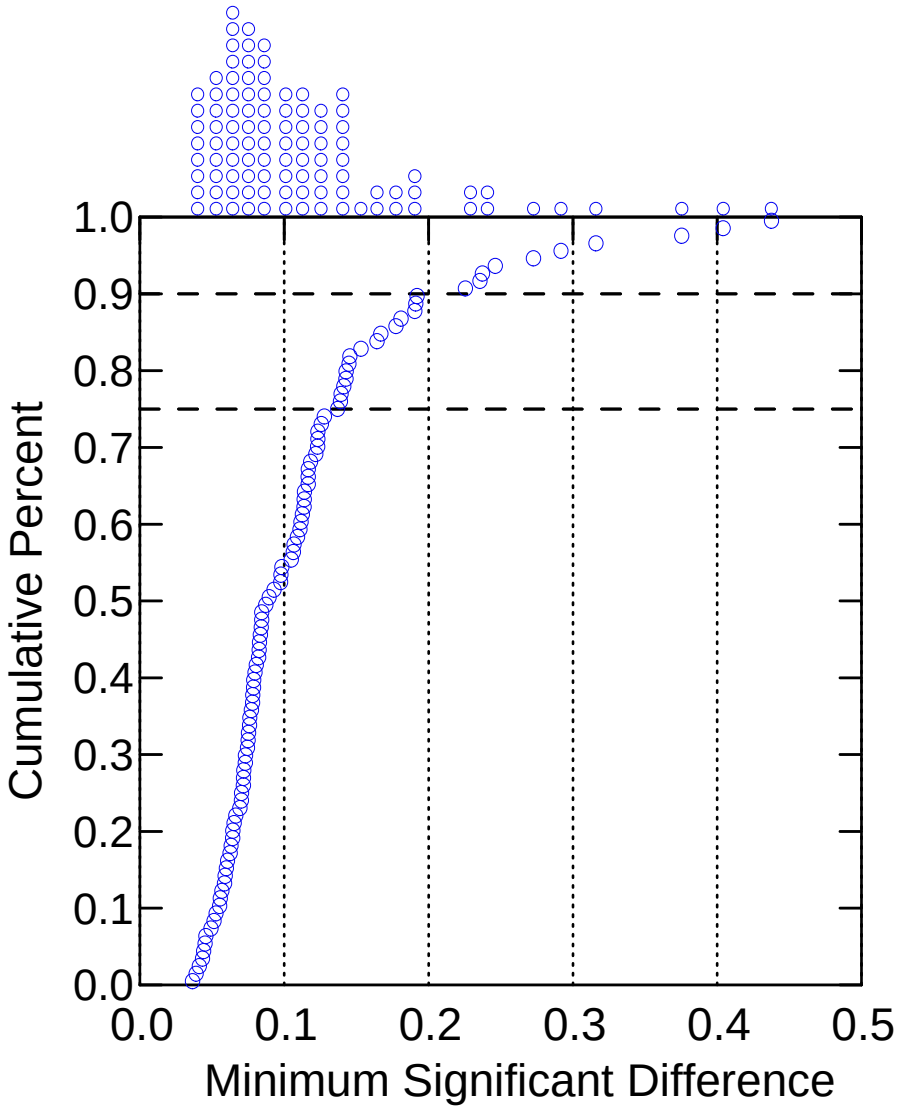
Analysis Methods

- MSD and Power were based on transformed data and then transformed back to original
- Detailed analysis of data based on non-transformed and transformed t-tests
- Comparison of NOEC and TST methods based on transformed data

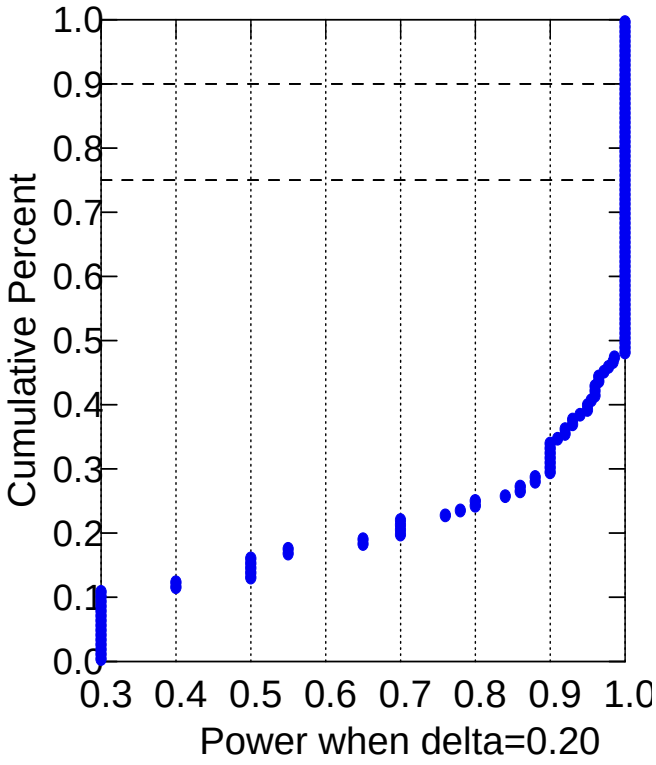
Table F-1. Simulation results for Red abalone development endpoint for tests deemed toxic. See Table B-1 for simulation method information.

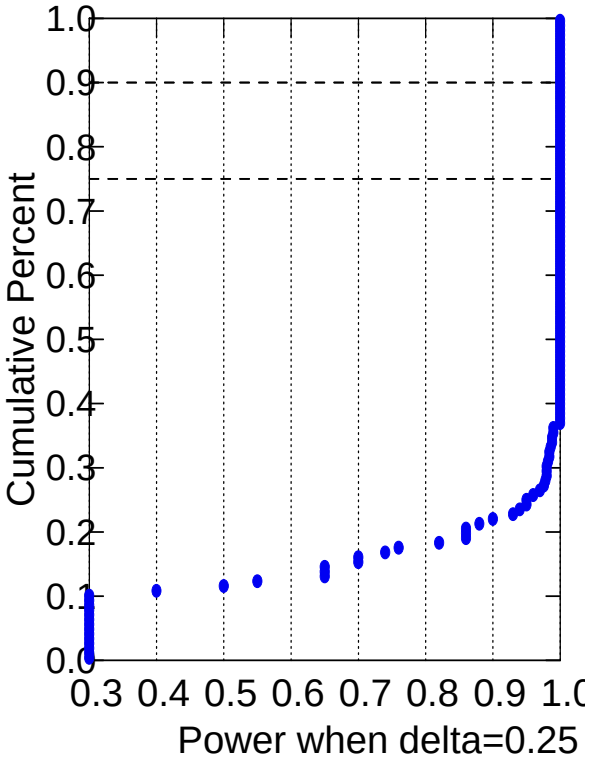
Mean Difference	C.V. Range	Result	NOEC	b=0.85	b=0.8	b=0.75	b=0.7	C.V. percentile
0.05	(0.5~1.5%)	Toxic	90.80%	0.00%	0.00%	0.00%	0.00%	0~10th
0.05	(1.5~2.1%)	Toxic	55.70%	9.20%	0.00%	0.00%	0.00%	10~25th
0.05	(2.1~3.1%)	Toxic	33.70%	38.20%	3.00%	0.00%	0.00%	25~50th
0.05	(3.1~4.5%)	Toxic	15.90%	64.60%	31.10%	4.60%	0.00%	50~75th
0.05	(4.5~5.3%)	Toxic	5.50%	73.30%	50.70%	23.80%	1.90%	75~85th
0.05	(5.3~6.9%)	Toxic	1.70%	84.10%	65.20%	43.30%	18.90%	85~95th
0.1	(0.5~1.5%)	Toxic	100.00%	35.30%	0.90%	0.00%	0.00%	0~10th
0.1	(1.5~2.1%)	Toxic	97.10%	65.60%	10.90%	0.00%	0.00%	10~25th
0.1	(2.1~3.1%)	Toxic	73.00%	80.30%	35.40%	3.20%	0.00%	25~50th
0.1	(3.1~4.5%)	Toxic	43.80%	93.70%	62.70%	27.20%	2.80%	50~75th
0.1	(4.5~5.3%)	Toxic	31.50%	98.70%	74.60%	50.80%	22.20%	75~85th
0.1	(5.3~6.9%)	Toxic	23.40%	100.00%	82.90%	63.00%	40.60%	85~95th
0.15	(0.5~1.5%)	Toxic	100.00%	92.10%	25.60%	0.90%	0.00%	0~10th
0.15	(1.5~2.1%)	Toxic	100.00%	99.90%	59.10%	8.40%	0.00%	10~25th
0.15	(2.1~3.1%)	Toxic	96.50%	100.00%	76.40%	28.20%	2.70%	25~50th
0.15	(3.1~4.5%)	Toxic	70.20%	100.00%	89.80%	58.30%	22.10%	50~75th
0.15	(4.5~5.3%)	Toxic	46.60%	100.00%	96.70%	74.20%	49.80%	75~85th
0.15	(5.3~6.9%)	Toxic	35.60%	100.00%	99.00%	82.40%	60.70%	85~95th
0.2	(0.5~1.5%)	Toxic	100.00%	100.00%	83.10%	15.90%	0.30%	0~10th
0.2	(1.5~2.1%)	Toxic	100.00%	100.00%	95.80%	47.90%	4.50%	10~25th
0.2	(2.1~3.1%)	Toxic	100.00%	100.00%	99.30%	66.10%	21.20%	25~50th
0.2	(3.1~4.5%)	Toxic	89.50%	100.00%	100.00%	83.60%	50.90%	50~75th
0.2	(4.5~5.3%)	Toxic	68.60%	100.00%	100.00%	92.10%	68.80%	75~85th
0.2	(5.3~6.9%)	Toxic	48.40%	100.00%	100.00%	96.90%	76.10%	85~95th
0.25	(0.5~1.5%)	Toxic	100.00%	100.00%	100.00%	65.20%	10.40%	0~10th
0.25	(1.5~2.1%)	Toxic	100.00%	100.00%	100.00%	87.80%	31.10%	10~25th
0.25	(2.1~3.1%)	Toxic	100.00%	100.00%	100.00%	95.10%	58.70%	25~50th
0.25	(3.1~4.5%)	Toxic	98.70%	100.00%	100.00%	99.40%	75.50%	50~75th
0.25	(4.5~5.3%)	Toxic	85.90%	100.00%	100.00%	100.00%	85.20%	75~85th
0.25	(5.3~6.9%)	Toxic	64.70%	100.00%	100.00%	100.00%	91.90%	85~95th
0.3	(0.5~1.5%)	Toxic	100.00%	100.00%	100.00%	100.00%	47.30%	0~10th
0.3	(1.5~2.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	76.30%	10~25th
0.3	(2.1~3.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	87.50%	25~50th
0.3	(3.1~4.5%)	Toxic	100.00%	100.00%	100.00%	100.00%	93.90%	50~75th
0.3	(4.5~5.3%)	Toxic	97.60%	100.00%	100.00%	100.00%	99.30%	75~85th
0.3	(5.3~6.9%)	Toxic	80.50%	100.00%	100.00%	100.00%	100.00%	85~95th
0.35	(0.5~1.5%)	Toxic	100.00%	100.00%	100.00%	100.00%	93.80%	0~10th
0.35	(1.5~2.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	99.70%	10~25th
0.35	(2.1~3.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	25~50th
0.35	(3.1~4.5%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	50~75th
0.35	(4.5~5.3%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	75~85th
0.35	(5.3~6.9%)	Toxic	91.10%	100.00%	100.00%	100.00%	100.00%	85~95th

MSD and CV for Ref Tox (N =135)

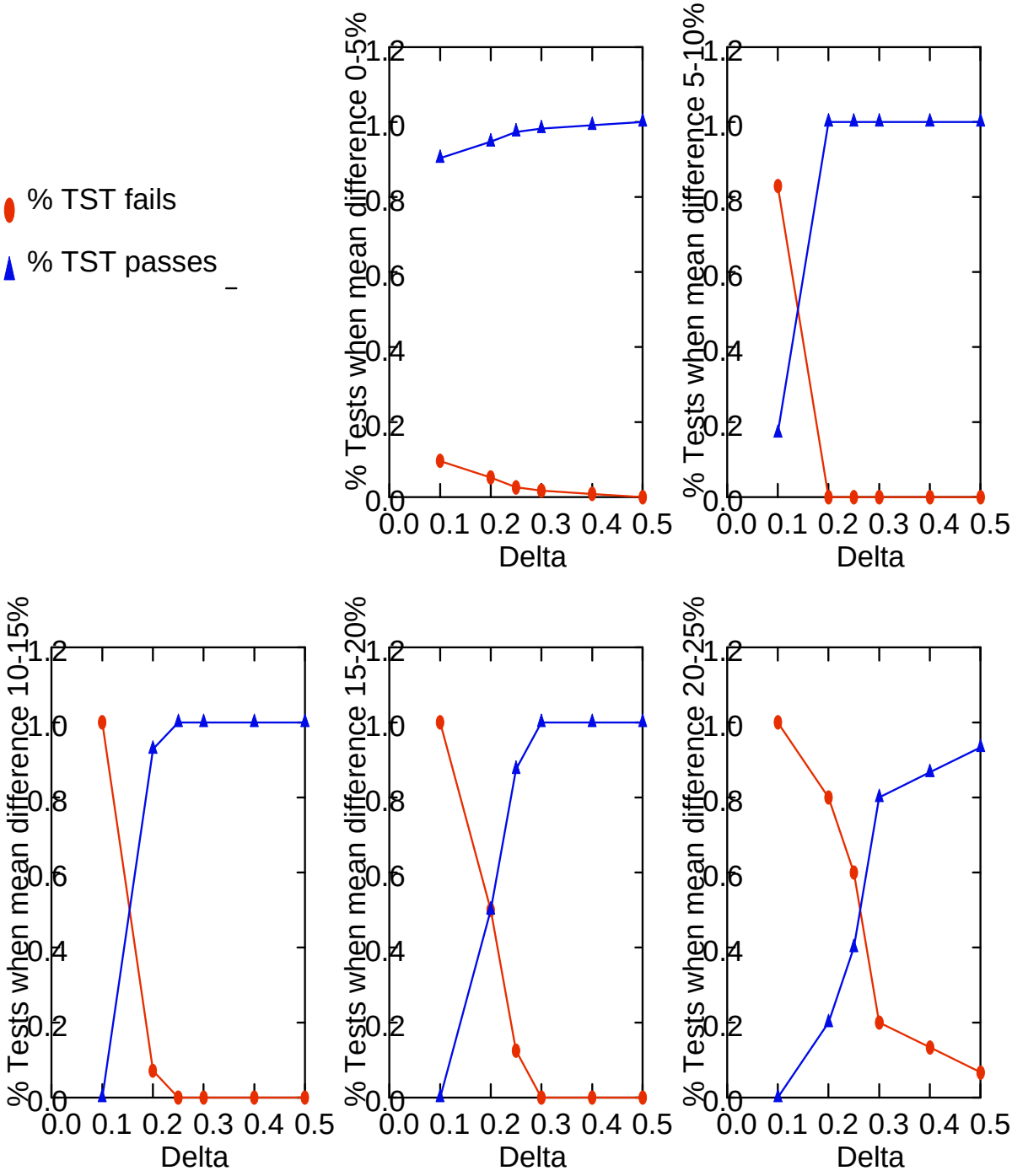


Power as a function of delta = 1-b





Detailed Analysis
Delta = 1-b



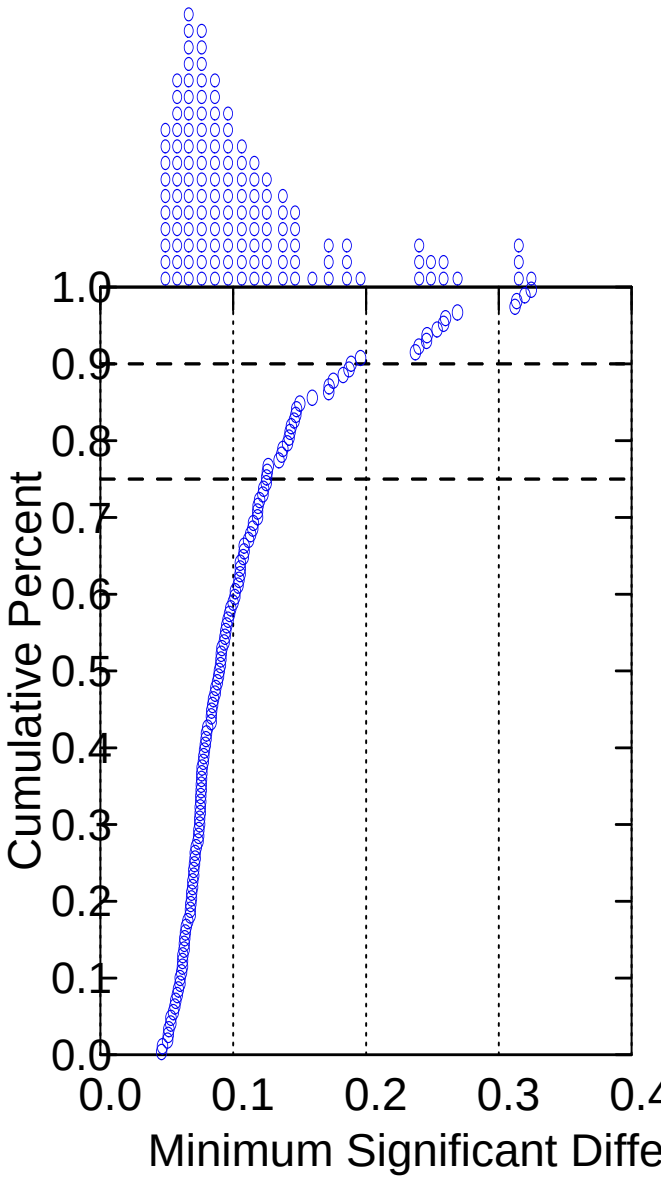
APPENDIX G
Giant Kelp Chronic Analysis

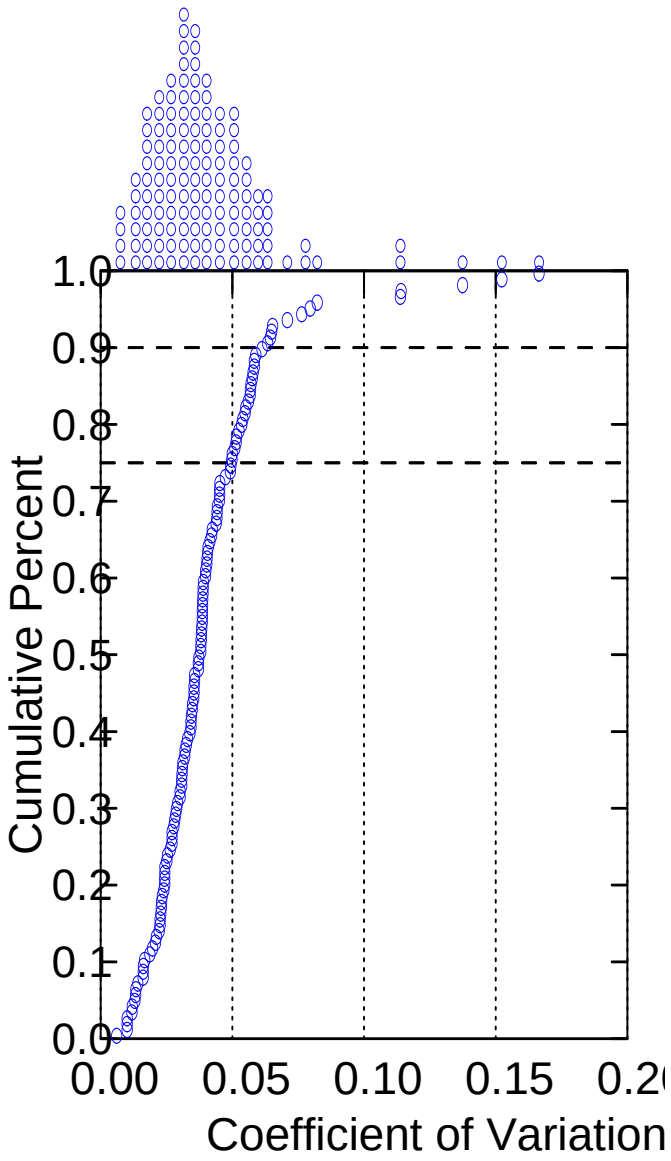
The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75th, 90th) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

Table G-1. Simulation results for Kelp tube length endpoint for tests deemed toxic. See Table B-1 for simulation method information.

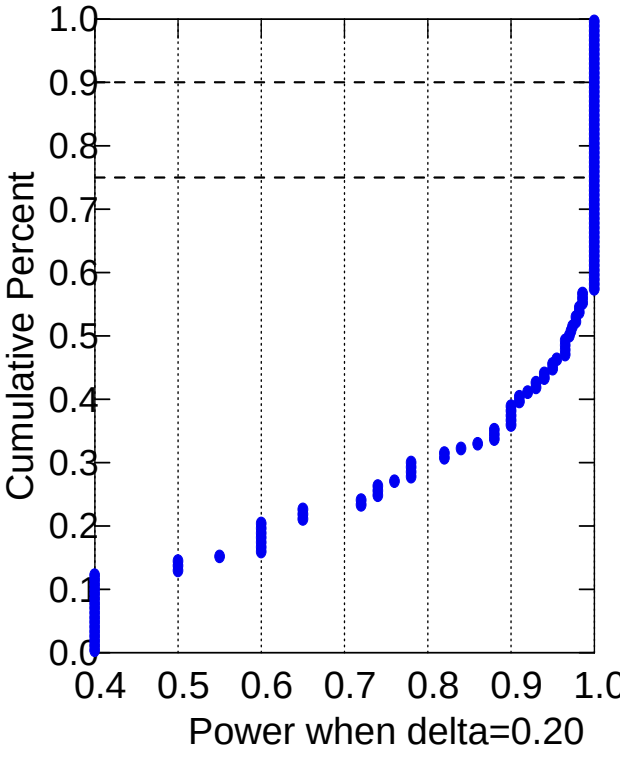
Mean	Variance	H0	NOEC	b=0.85	b=0.8	b=0.75	b=0.7	C.V. percentile
0.05	(1~3%)	Toxic	99.30%	0.00%	0.00%	0.00%	0.00%	0~10th
0.05	(3~5.2%)	Toxic	30.00%	0.00%	0.00%	0.00%	0.00%	10~25th
0.05	(5.2~7.3%)	Toxic	0.00%	11.50%	0.00%	0.00%	0.00%	25~50th
0.05	(7.3~9.1%)	Toxic	0.00%	57.20%	0.00%	0.00%	0.00%	50~75th
0.05	(9.1~11%)	Toxic	0.00%	93.80%	16.60%	0.00%	0.00%	75~85th
0.05	(11~14%)	Toxic	0.00%	100.00%	56.30%	4.70%	0.00%	85~95th
0.1	(1~3%)	Toxic	100.00%	0.00%	0.00%	0.00%	0.00%	0~10th
0.1	(3~5.2%)	Toxic	100.00%	57.40%	0.00%	0.00%	0.00%	10~25th
0.1	(5.2~7.3%)	Toxic	75.80%	100.00%	10.50%	0.00%	0.00%	25~50th
0.1	(7.3~9.1%)	Toxic	26.80%	100.00%	53.40%	0.00%	0.00%	50~75th
0.1	(9.1~11%)	Toxic	0.60%	100.00%	92.30%	13.00%	0.00%	75~85th
0.1	(11~14%)	Toxic	0.00%	100.00%	100.00%	54.10%	3.60%	85~95th
0.15	(1~3%)	Toxic	100.00%	100.00%	0.00%	0.00%	0.00%	0~10th
0.15	(3~5.2%)	Toxic	100.00%	100.00%	52.70%	0.00%	0.00%	10~25th
0.15	(5.2~7.3%)	Toxic	100.00%	100.00%	100.00%	6.90%	0.00%	25~50th
0.15	(7.3~9.1%)	Toxic	96.00%	100.00%	100.00%	49.20%	0.00%	50~75th
0.15	(9.1~11%)	Toxic	62.00%	100.00%	100.00%	87.90%	10.20%	75~85th
0.15	(11~14%)	Toxic	24.90%	100.00%	100.00%	100.00%	51.90%	85~95th
0.2	(1~3%)	Toxic	100.00%	100.00%	100.00%	0.00%	0.00%	0~10th
0.2	(3~5.2%)	Toxic	100.00%	100.00%	100.00%	53.00%	0.00%	10~25th
0.2	(5.2~7.3%)	Toxic	100.00%	100.00%	100.00%	99.90%	5.80%	25~50th
0.2	(7.3~9.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	43.90%	50~75th
0.2	(9.1~11%)	Toxic	100.00%	100.00%	100.00%	100.00%	85.20%	75~85th
0.2	(11~14%)	Toxic	78.00%	100.00%	100.00%	100.00%	100.00%	85~95th
0.25	(1~3%)	Toxic	100.00%	100.00%	100.00%	100.00%	0.00%	0~10th
0.25	(3~5.2%)	Toxic	100.00%	100.00%	100.00%	100.00%	43.10%	10~25th
0.25	(5.2~7.3%)	Toxic	100.00%	100.00%	100.00%	100.00%	99.70%	25~50th
0.25	(7.3~9.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	50~75th
0.25	(9.1~11%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	75~85th
0.25	(11~14%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	85~95th
0.3	(1~3%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	0~10th
0.3	(3~5.2%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	10~25th
0.3	(5.2~7.3%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	25~50th
0.3	(7.3~9.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	50~75th
0.3	(9.1~11%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	75~85th
0.3	(11~14%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	85~95th
0.35	(1~3%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	0~10th
0.35	(3~5.2%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	10~25th
0.35	(5.2~7.3%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	25~50th
0.35	(7.3~9.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	50~75th
0.35	(9.1~11%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	75~85th
0.35	(11~14%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	85~95th
0.05	(1~3%)	Non-Toxic	0.70%	100.00%	100.00%	100.00%	100.00%	0~10th

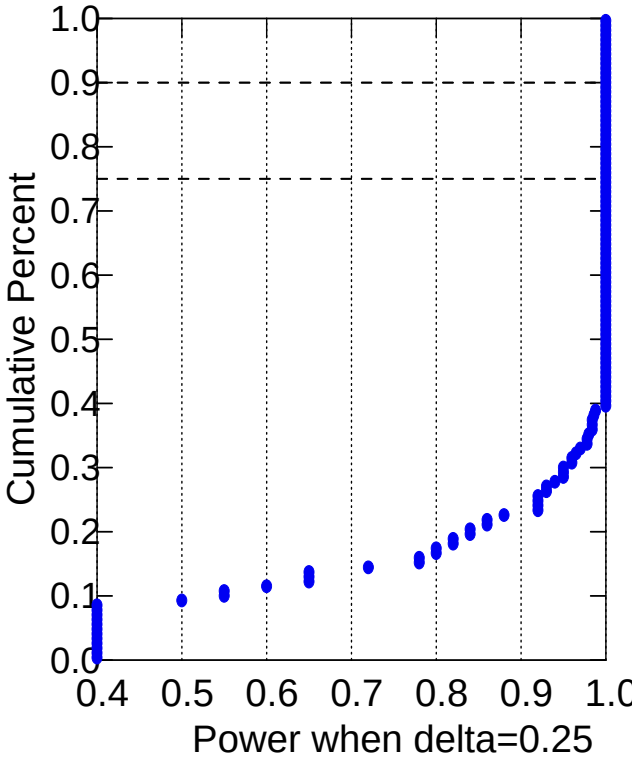
CV and MSD: Tube Length





Power: Tube Length as function of delta = 1-b





Detailed Delta Analysis: tube length delta = 1-b

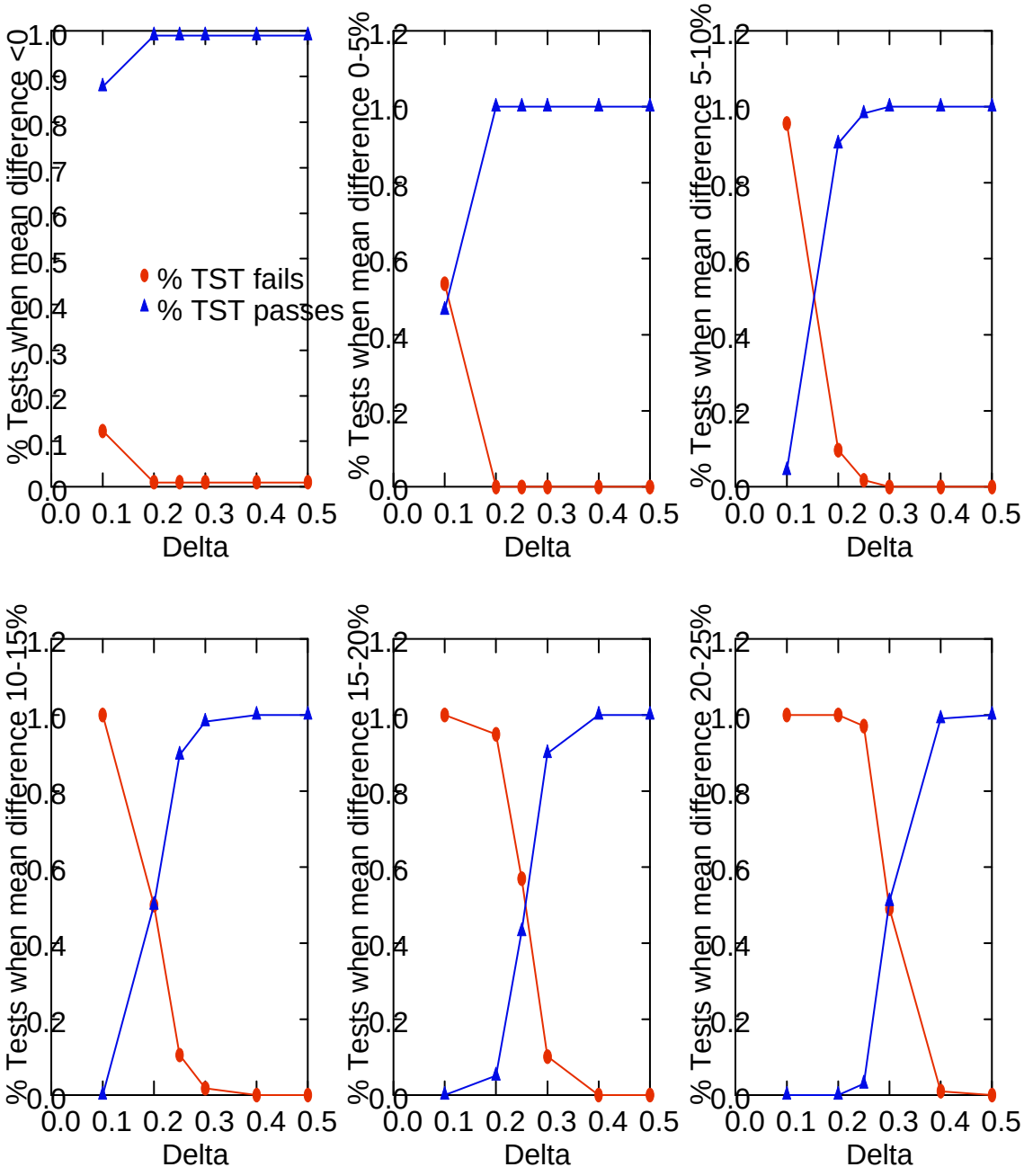
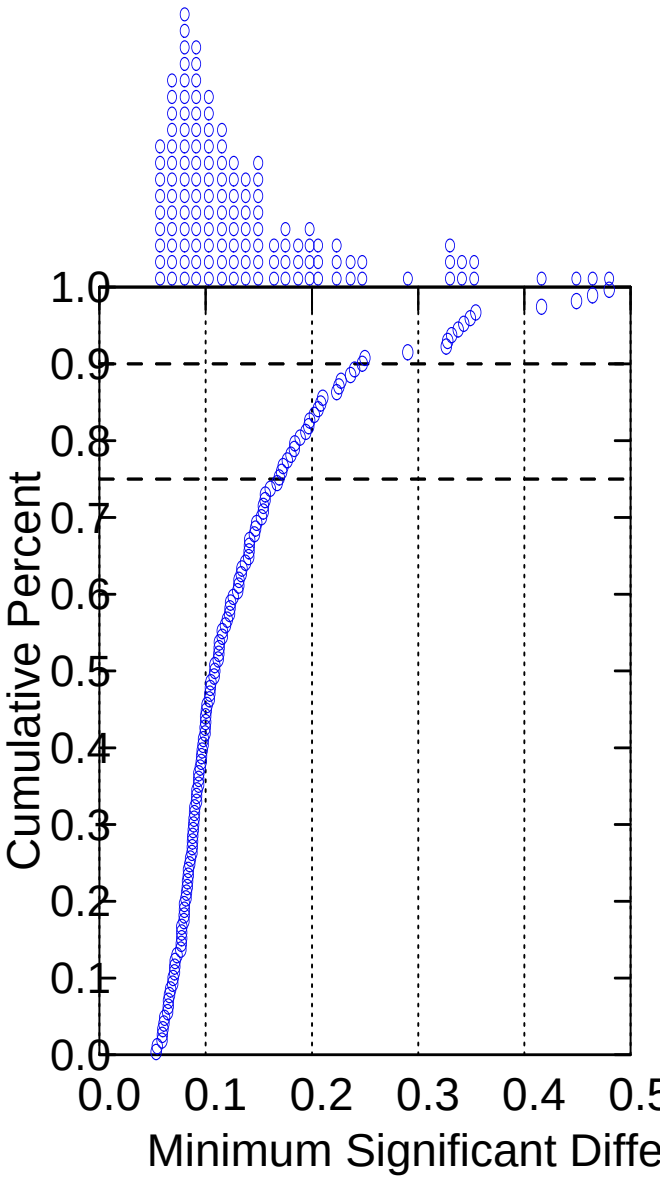
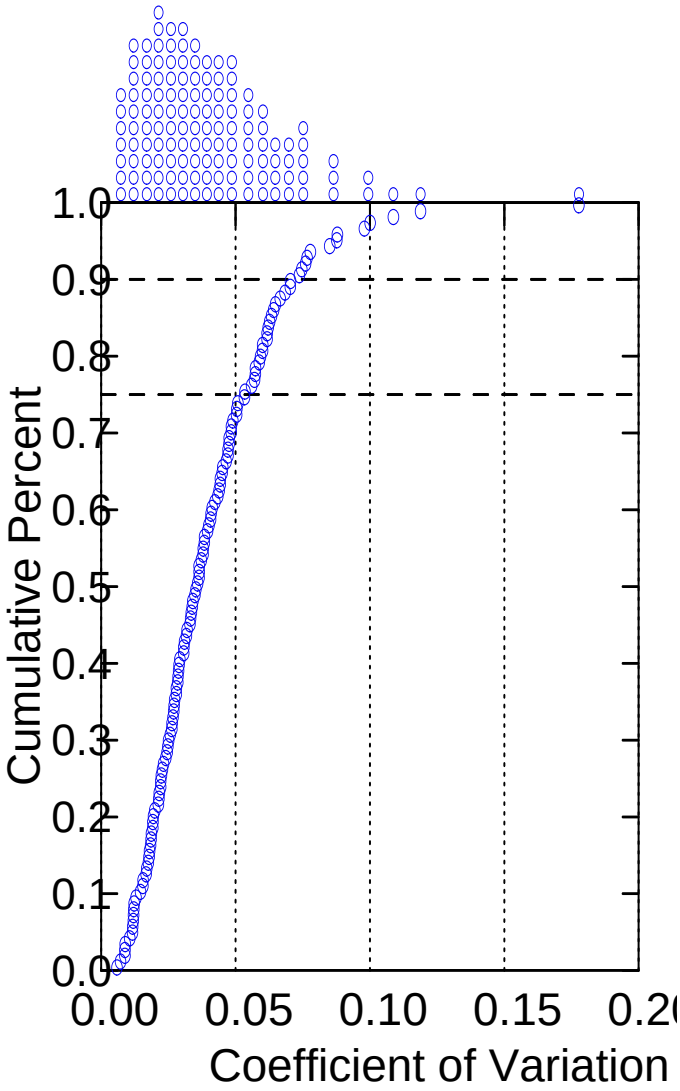


Table G-2. Simulation results for Kelp germination endpoint for tests deemed toxic. See Table B-1 for simulation method information.

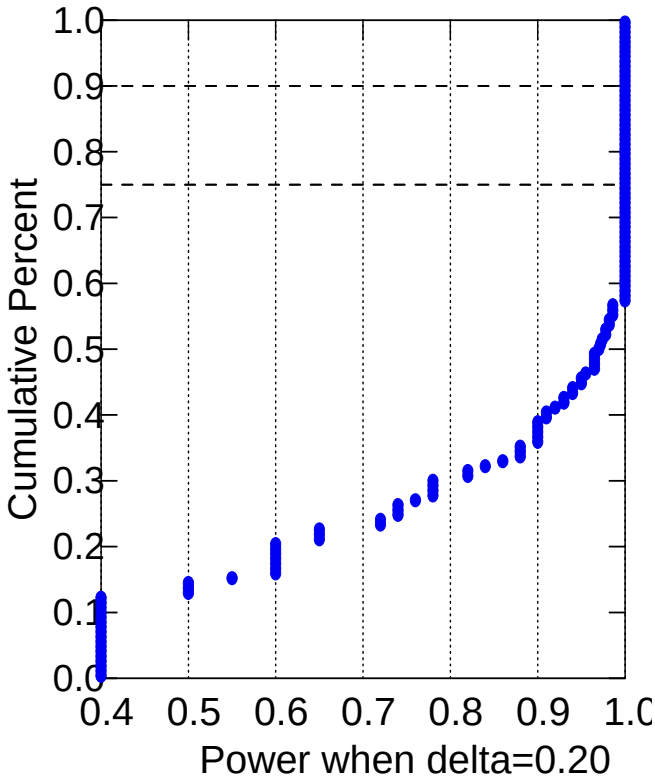
Mean	Variance	H0	NOEC	b=0.85	b=0.8	b=0.75	b=0.7	percentile
0.05	(1~1.8%)	Toxic	100.00%	0.00%	0.00%	0.00%	0.00%	0~10th
0.05	(1.8~2.7%)	Toxic	99.70%	0.00%	0.00%	0.00%	0.00%	10~25th
0.05	(2.7~3.8%)	Toxic	71.70%	0.00%	0.00%	0.00%	0.00%	25~50th
0.05	(3.8~5.1%)	Toxic	24.10%	0.20%	0.00%	0.00%	0.00%	50~75th
0.05	(5.1~6%)	Toxic	5.00%	5.90%	0.00%	0.00%	0.00%	75~85th
0.05	(6~10.1%)	Toxic	0.00%	57.50%	3.00%	0.00%	0.00%	85~95th
0.1	(1~1.8%)	Toxic	100.00%	1.80%	0.00%	0.00%	0.00%	0~10th
0.1	(1.8~2.7%)	Toxic	100.00%	12.90%	0.00%	0.00%	0.00%	10~25th
0.1	(2.7~3.8%)	Toxic	100.00%	29.30%	0.00%	0.00%	0.00%	25~50th
0.1	(3.8~5.1%)	Toxic	99.70%	66.20%	0.70%	0.00%	0.00%	50~75th
0.1	(5.1~6%)	Toxic	91.40%	88.30%	6.30%	0.00%	0.00%	75~85th
0.1	(6~10.1%)	Toxic	33.10%	99.70%	50.30%	2.70%	0.00%	85~95th
0.15	(1~1.8%)	Toxic	100.00%	69.60%	0.90%	0.00%	0.00%	0~10th
0.15	(1.8~2.7%)	Toxic	100.00%	93.90%	6.80%	0.00%	0.00%	10~25th
0.15	(2.7~3.8%)	Toxic	100.00%	99.90%	19.90%	0.00%	0.00%	25~50th
0.15	(3.8~5.1%)	Toxic	100.00%	100.00%	47.40%	0.30%	0.00%	50~75th
0.15	(5.1~6%)	Toxic	100.00%	100.00%	73.60%	2.90%	0.00%	75~85th
0.15	(6~10.1%)	Toxic	83.90%	100.00%	95.80%	36.00%	2.30%	85~95th
0.2	(1~1.8%)	Toxic	100.00%	100.00%	45.60%	0.00%	0.00%	0~10th
0.2	(1.8~2.7%)	Toxic	100.00%	100.00%	67.30%	3.00%	0.00%	10~25th
0.2	(2.7~3.8%)	Toxic	100.00%	100.00%	92.40%	11.40%	0.00%	25~50th
0.2	(3.8~5.1%)	Toxic	100.00%	100.00%	100.00%	26.20%	0.00%	50~75th
0.2	(5.1~6%)	Toxic	100.00%	100.00%	100.00%	51.90%	1.30%	75~85th
0.2	(6~10.1%)	Toxic	99.20%	100.00%	100.00%	89.80%	25.30%	85~95th
0.25	(1~1.8%)	Toxic	100.00%	100.00%	100.00%	30.40%	0.00%	0~10th
0.25	(1.8~2.7%)	Toxic	100.00%	100.00%	100.00%	46.50%	0.00%	10~25th
0.25	(2.7~3.8%)	Toxic	100.00%	100.00%	100.00%	68.10%	3.90%	25~50th
0.25	(3.8~5.1%)	Toxic	100.00%	100.00%	100.00%	94.30%	16.20%	50~75th
0.25	(5.1~6%)	Toxic	100.00%	100.00%	100.00%	99.70%	29.10%	75~85th
0.25	(6~10.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	72.10%	85~95th
0.3	(1~1.8%)	Toxic	100.00%	100.00%	100.00%	96.40%	15.90%	0~10th
0.3	(1.8~2.7%)	Toxic	100.00%	100.00%	100.00%	100.00%	26.30%	10~25th
0.3	(2.7~3.8%)	Toxic	100.00%	100.00%	100.00%	100.00%	44.90%	25~50th
0.3	(3.8~5.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	72.20%	50~75th
0.3	(5.1~6%)	Toxic	100.00%	100.00%	100.00%	100.00%	89.00%	75~85th
0.3	(6~10.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	99.40%	85~95th
0.35	(1~1.8%)	Toxic	100.00%	100.00%	100.00%	100.00%	70.10%	0~10th
0.35	(1.8~2.7%)	Toxic	100.00%	100.00%	100.00%	100.00%	89.90%	10~25th
0.35	(2.7~3.8%)	Toxic	100.00%	100.00%	100.00%	100.00%	99.50%	25~50th
0.35	(3.8~5.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	50~75th
0.35	(5.1~6%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	75~85th
0.35	(6~10.1%)	Toxic	100.00%	100.00%	100.00%	100.00%	100.00%	85~95th

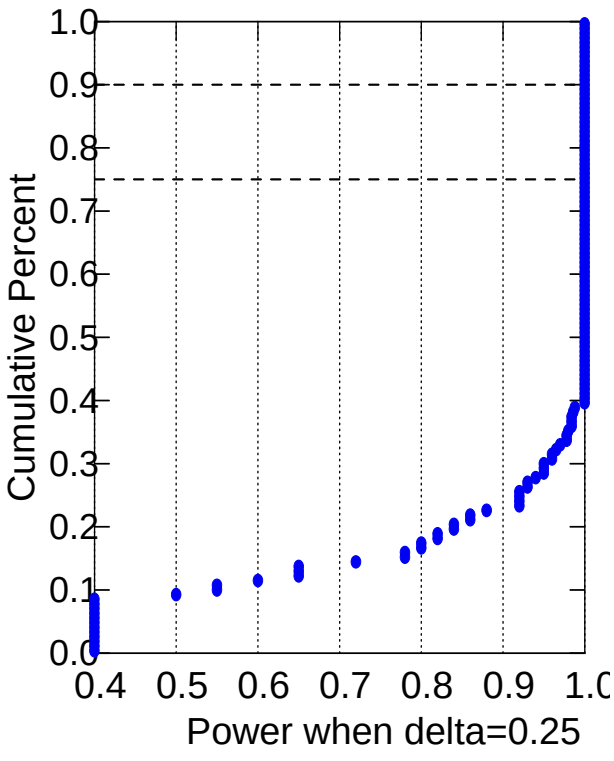
CV and MSD: Germination (transformed)



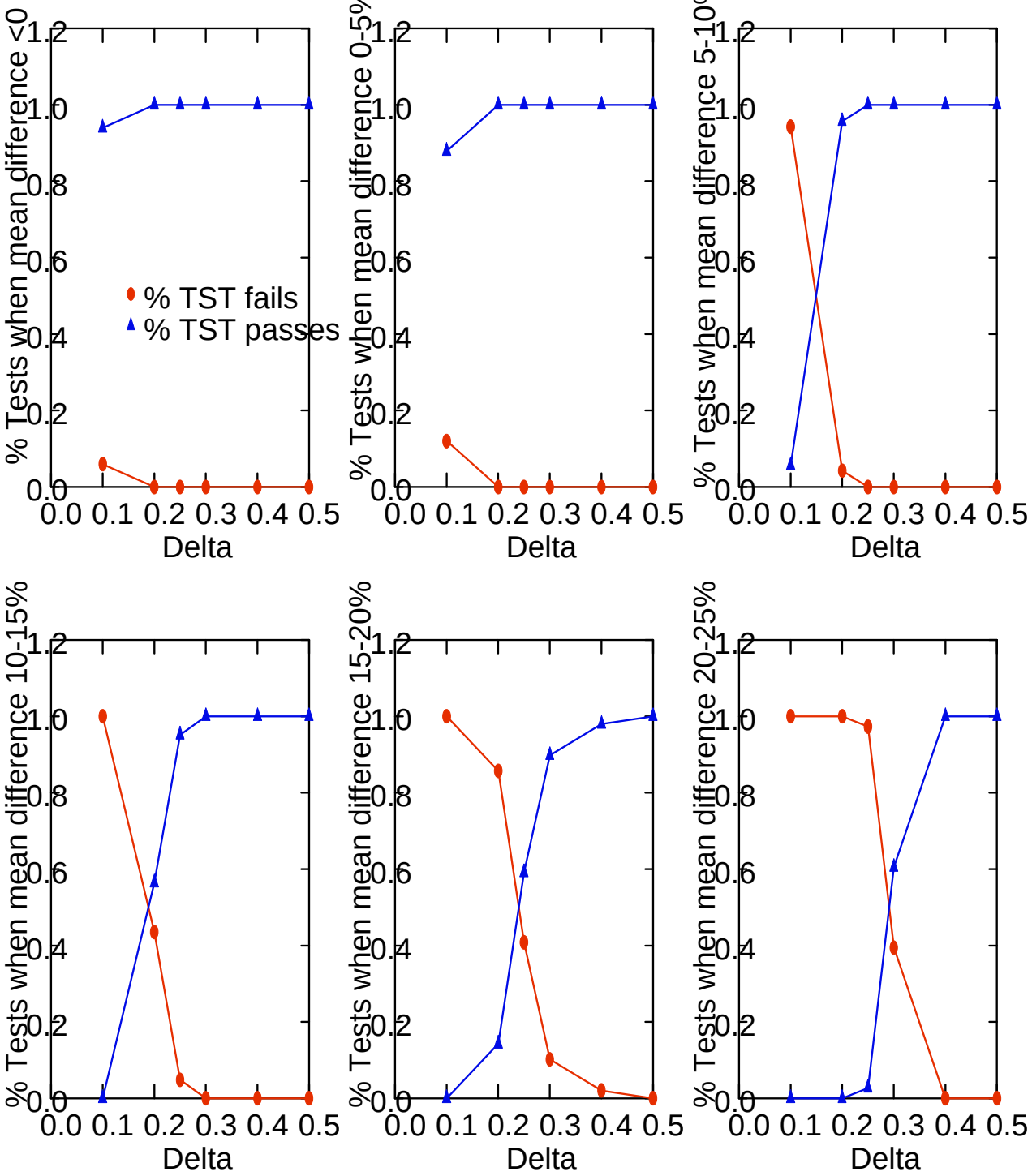


Power: Germination (after transformation) as function of delta = 1-b





Detailed Delta Analysis: germination delta = 1-b



APPENDIX H

Pimephales promelas Acute Survival Results

The following Appendix includes a summary of simulation analyses using both TST and the current NOEC approach as well as figures summarizing within-test variability (as MSD and CV) observed in effluent and reference toxicant tests obtained from many laboratories. Cumulative distributions of each are presented such that percentiles of interest (e.g., 75th, 90th) can be readily identified. These cumulative distribution graphs are followed by graphs summarizing power of the test to identify a given percent effect as toxic as a function of within-test variability (expressed either as CV or MSD). This appendix also includes graphs showing the fraction of WET tests examined that would show toxicity as a function of percent effect range (e.g., 10-15%) and what is referred to as delta. Delta = 1-b, where b is the proportion of the control mean used in TST analysis.

Data

- One laboratory conducted approximately 75% of the 347 tests (259)
- To account for lab variability, analyses of MSD and power were run using the entire data set (347 tests) and using a dataset without the one dominant lab (88 tests). Insignificant differences were observed between the two datasets. Therefore, parameters reported are for all 347 tests.

Table H-1. Simulation results for *P. promelas* Acute Survival. See legend for Table B-1, Appendix A for general information on simulation parameters. Mean difference or effect levels ranged between 0.2 (20%) and 0.4 (40%) for this simulation.

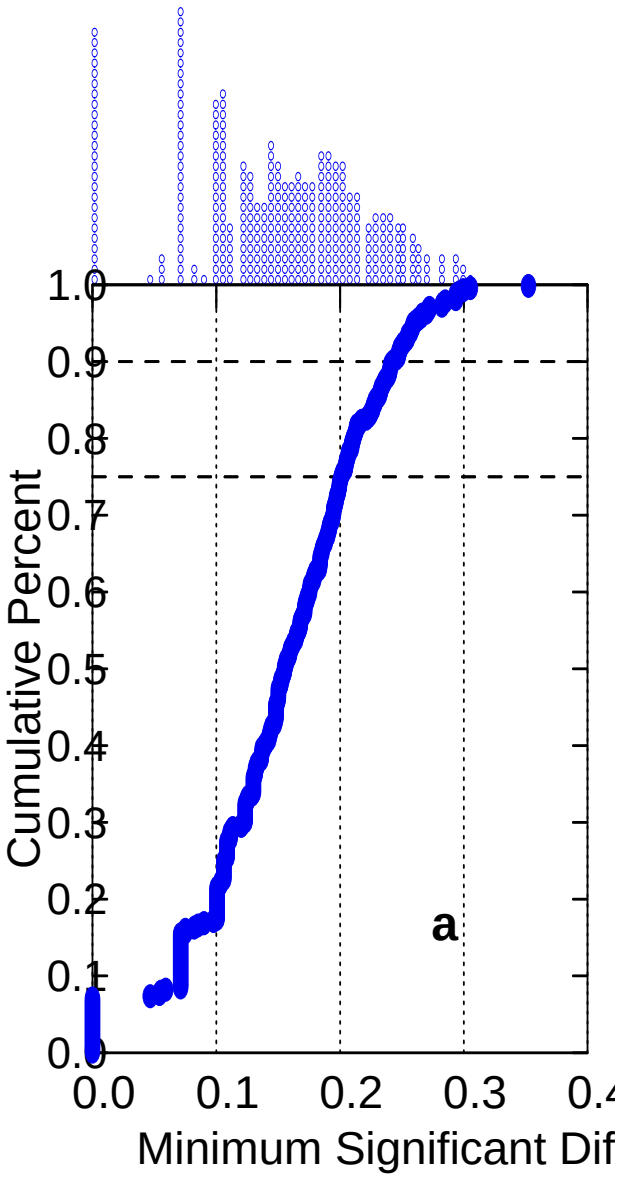
Mean Difference	C.V. Range	Result	NOEC	b=0.75	b=0.70	b=0.65	b=0.60	C.V. percentile
0.15	(<10 %)	Toxic	58.0%	45.6 %	0.07 %	0.0%	0.0%	< 90th
0.15	(>10 %)	Toxic	1.1%	99.0%	83.5%	51.5%	21.6%	> 90th
0.2	(<10 %)	Toxic	87.8%	91.6%	33.9%	2.60%	0.00%	< 90th
0.2	(>10 %)	Toxic	20.5%	100.0%	97.0%	79.7%	13.9%	> 90th
0.25	(<10 %)	Toxic	98.0%	100.0%	78.6.0%	22.1%	0.3%	< 90th
0.25	(>10 %)	Toxic	50.0%	100.0%	100.0%	93.0%	64.1%	> 90th
0.3	(<10 %)	Toxic	100.0%	100.0%	99.9%	58.9%	11.0%	< 90th
0.3	(>10 %)	Toxic	73.1%	100.0%	100.0%	100.0%	64.1%	> 90th
0.4	(<10 %)	Toxic	100.0%	100.0%	100.0%	100.0%	100.0%	< 90th
0.4	(>10 %)	Toxic	98.4%	100.0%	100.0%	100.0%	100.0%	> 90th

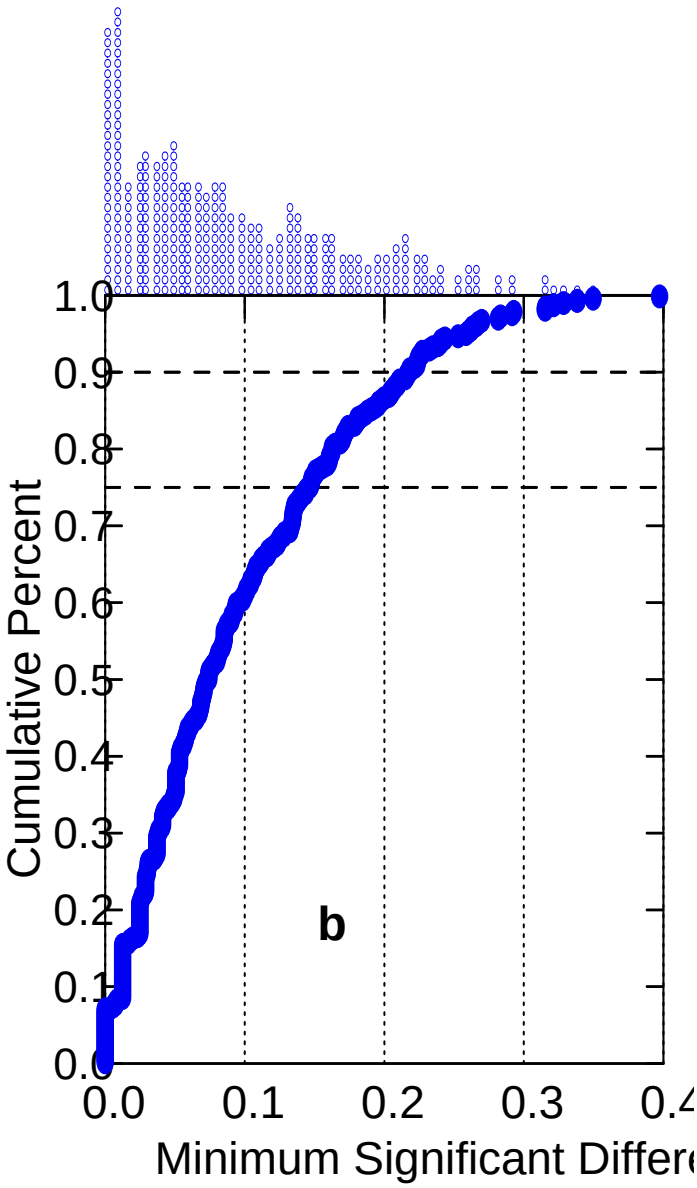
000137

MSD Analyses

ALL (347) Tests

(a) transformed data (b) untransformed data



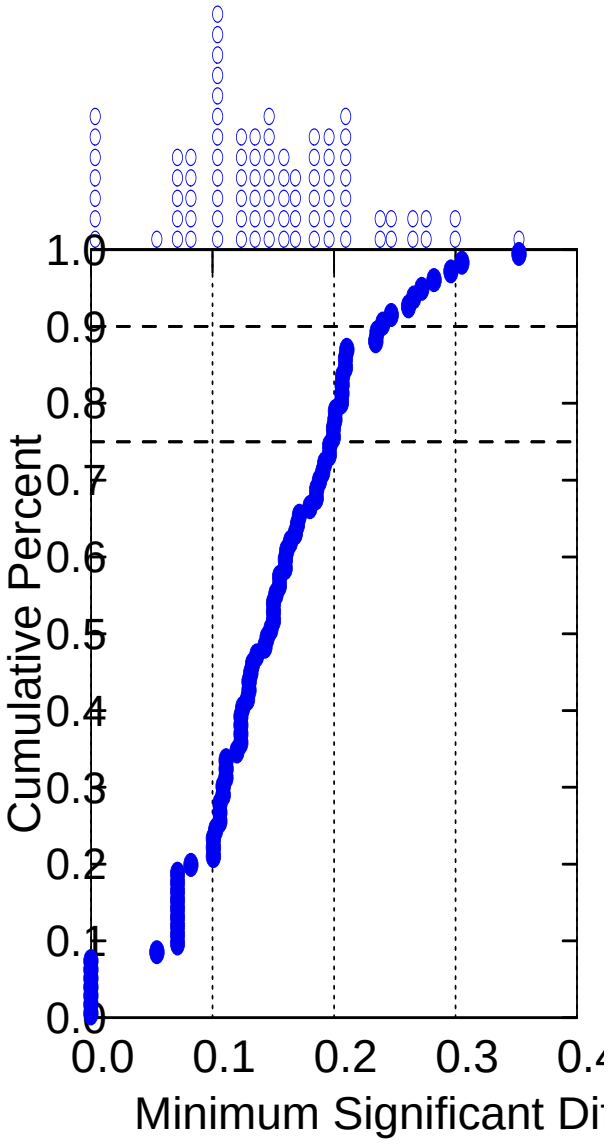


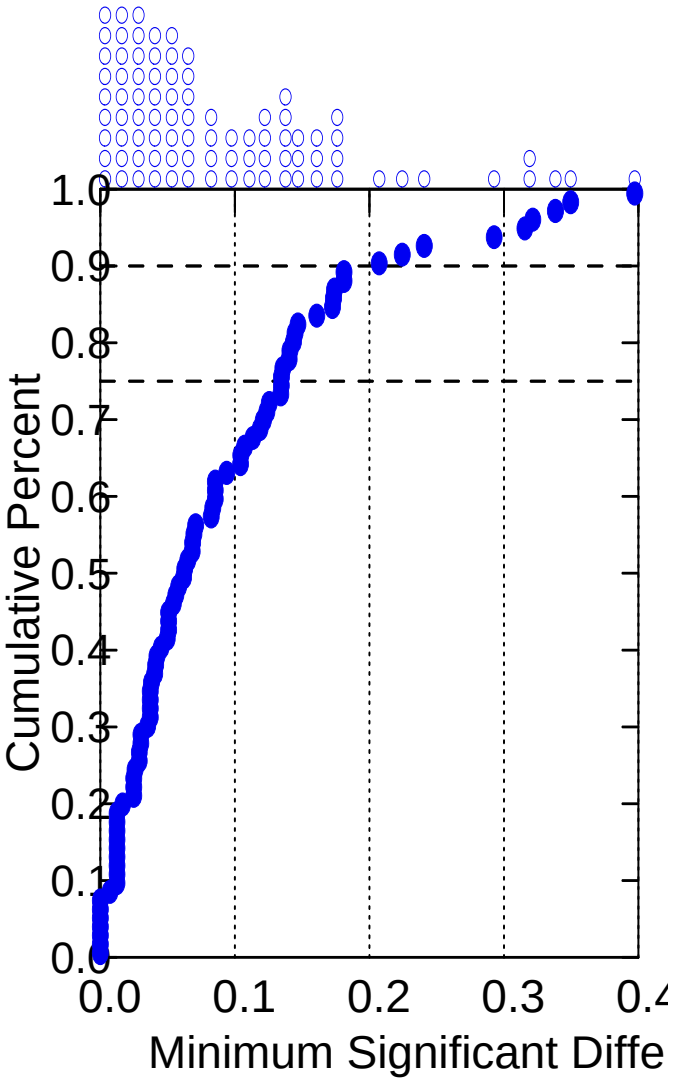
000138

MSD Analyses

Selected cases (88 tests)

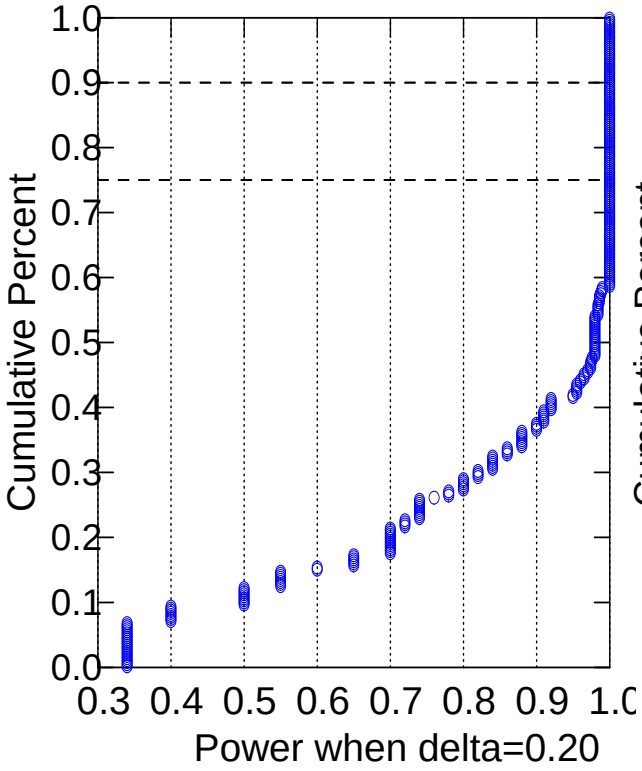
(a) transformed data (b) untransformed data

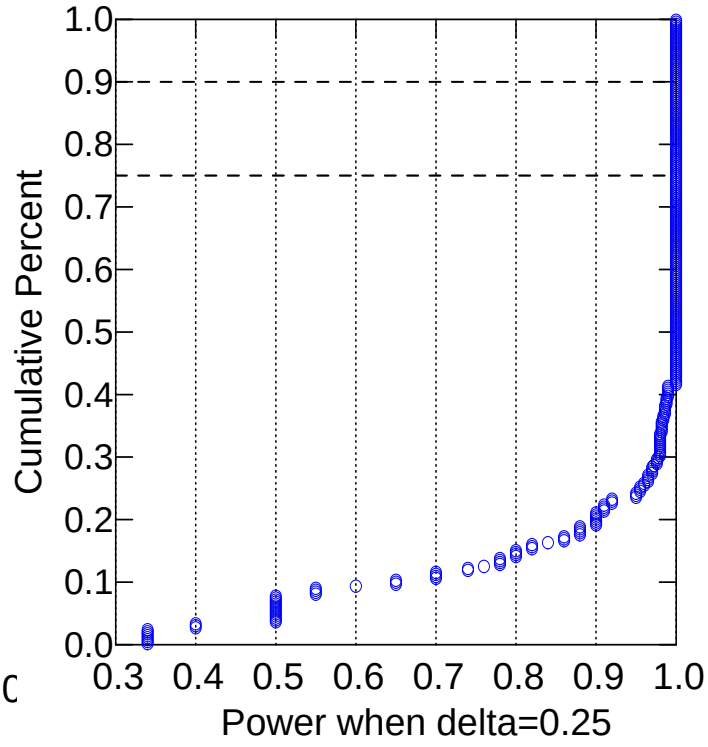


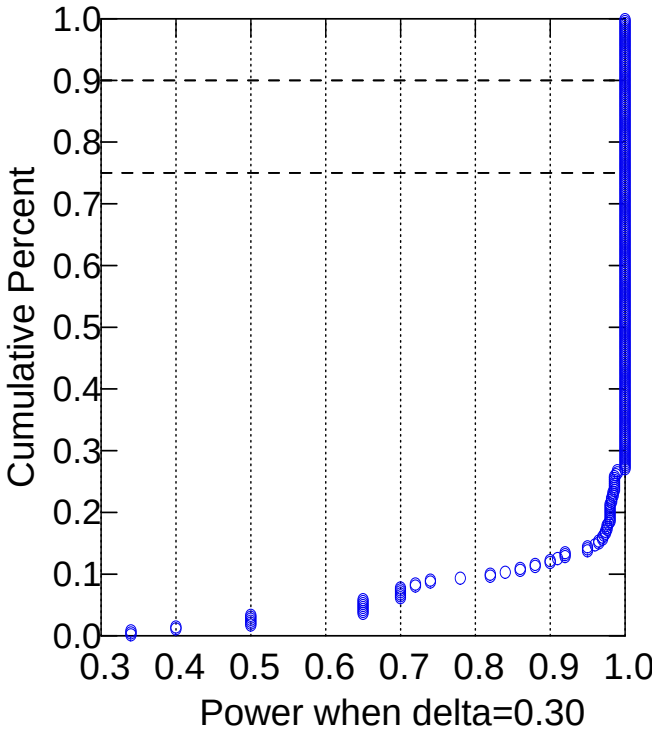


Power Analyses (All tests)

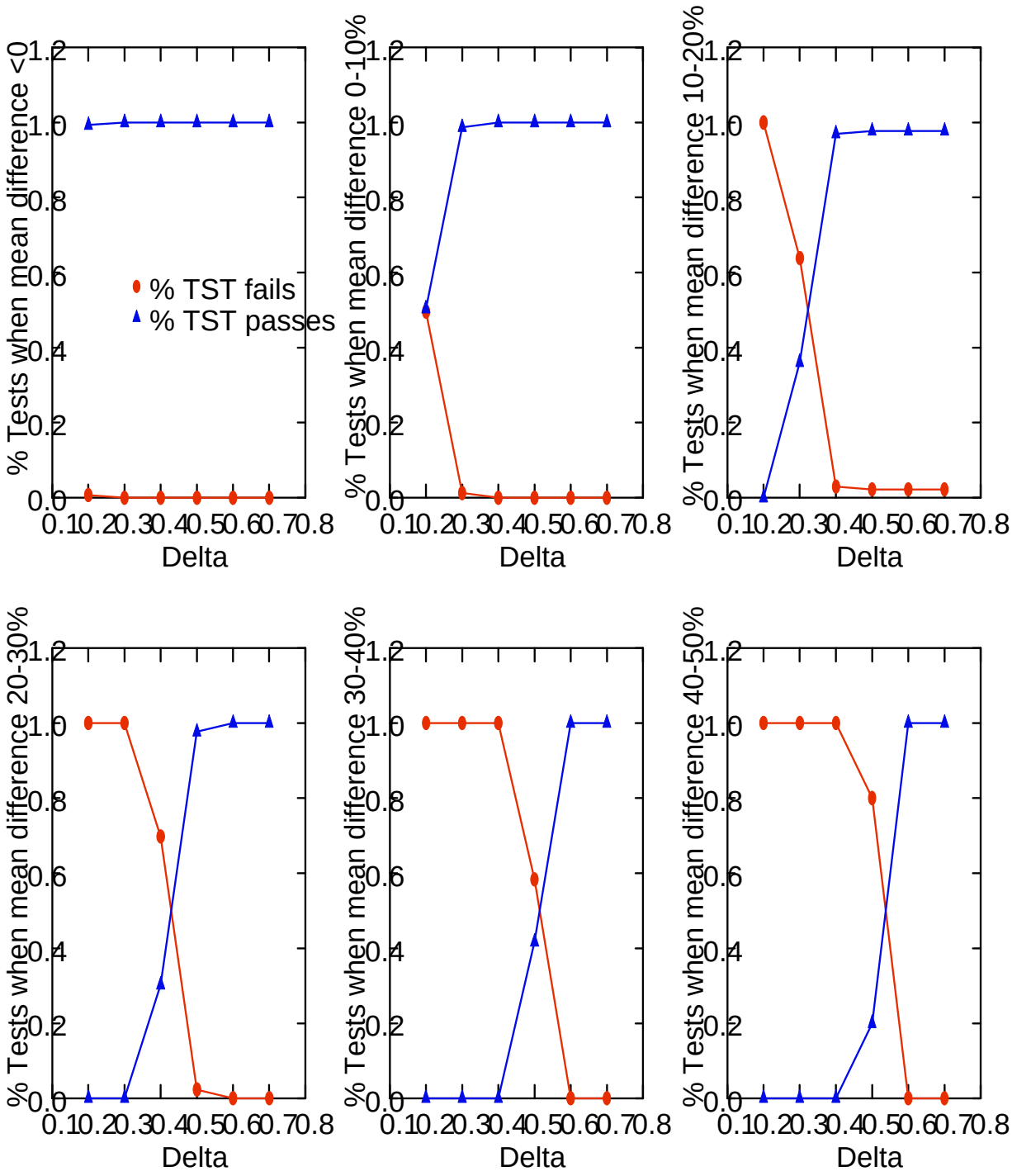
delta = 1-b





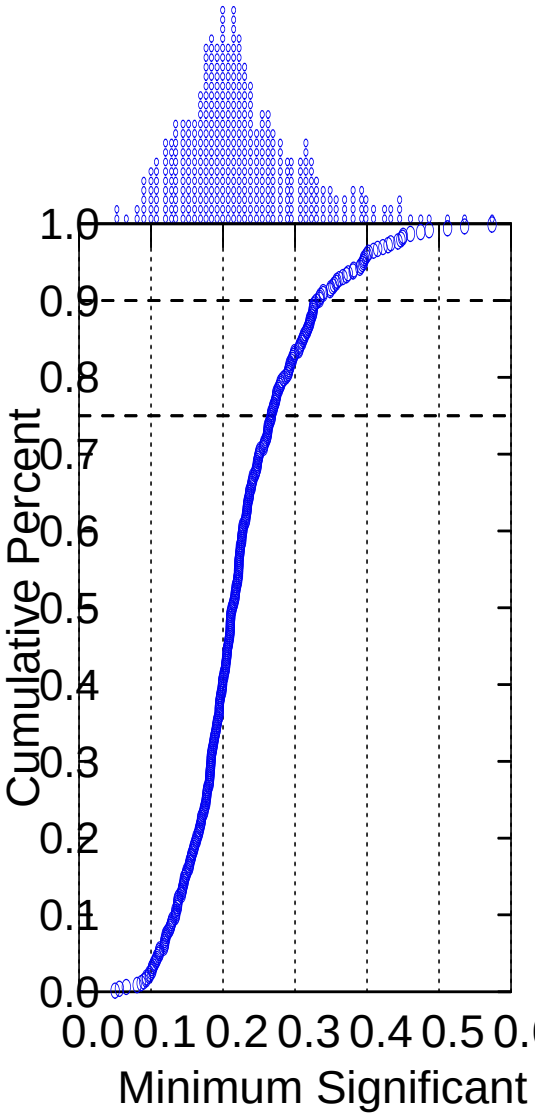


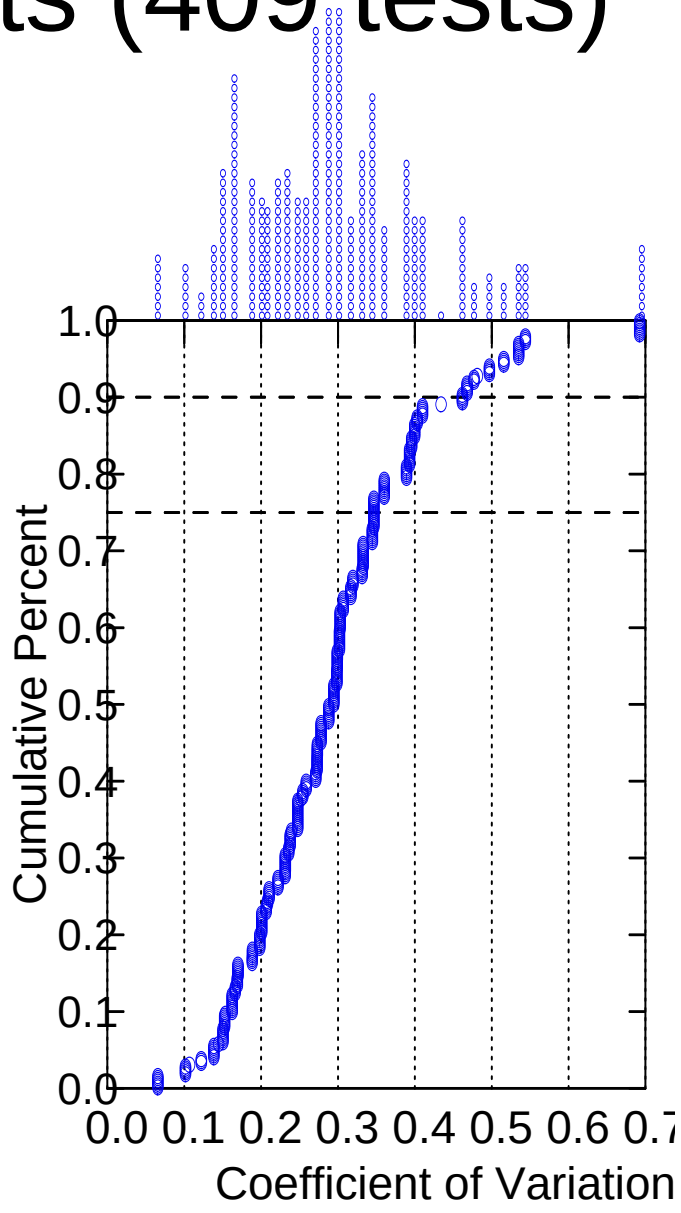
Detailed Analyses
delta = 1-b



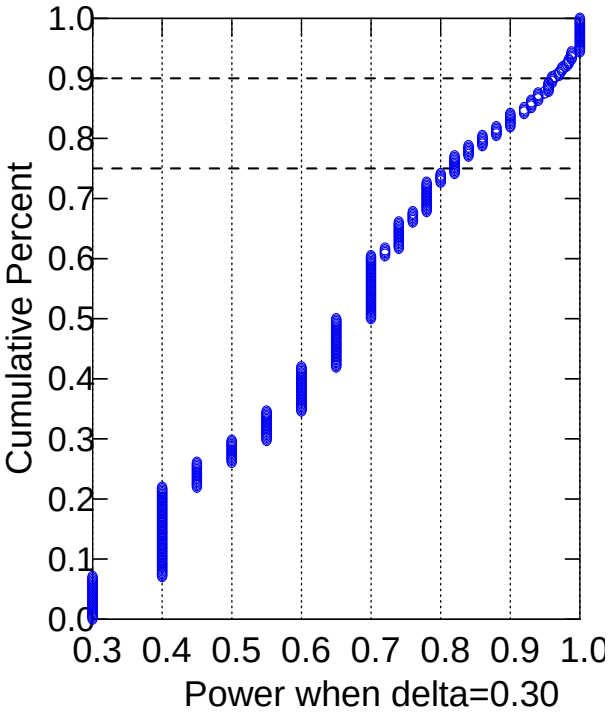
APPENDIX I
Ceriodaphnia Ambient Toxicity Data

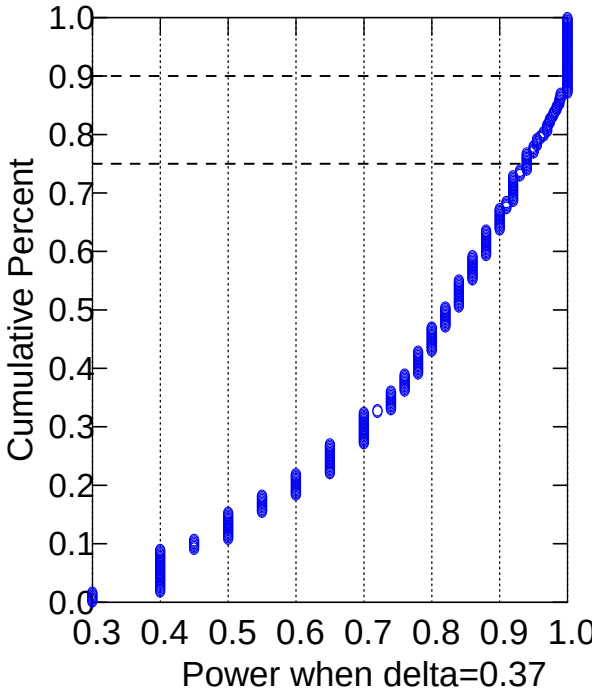
Variability of Ambient Toxicity Tests (409 tests)



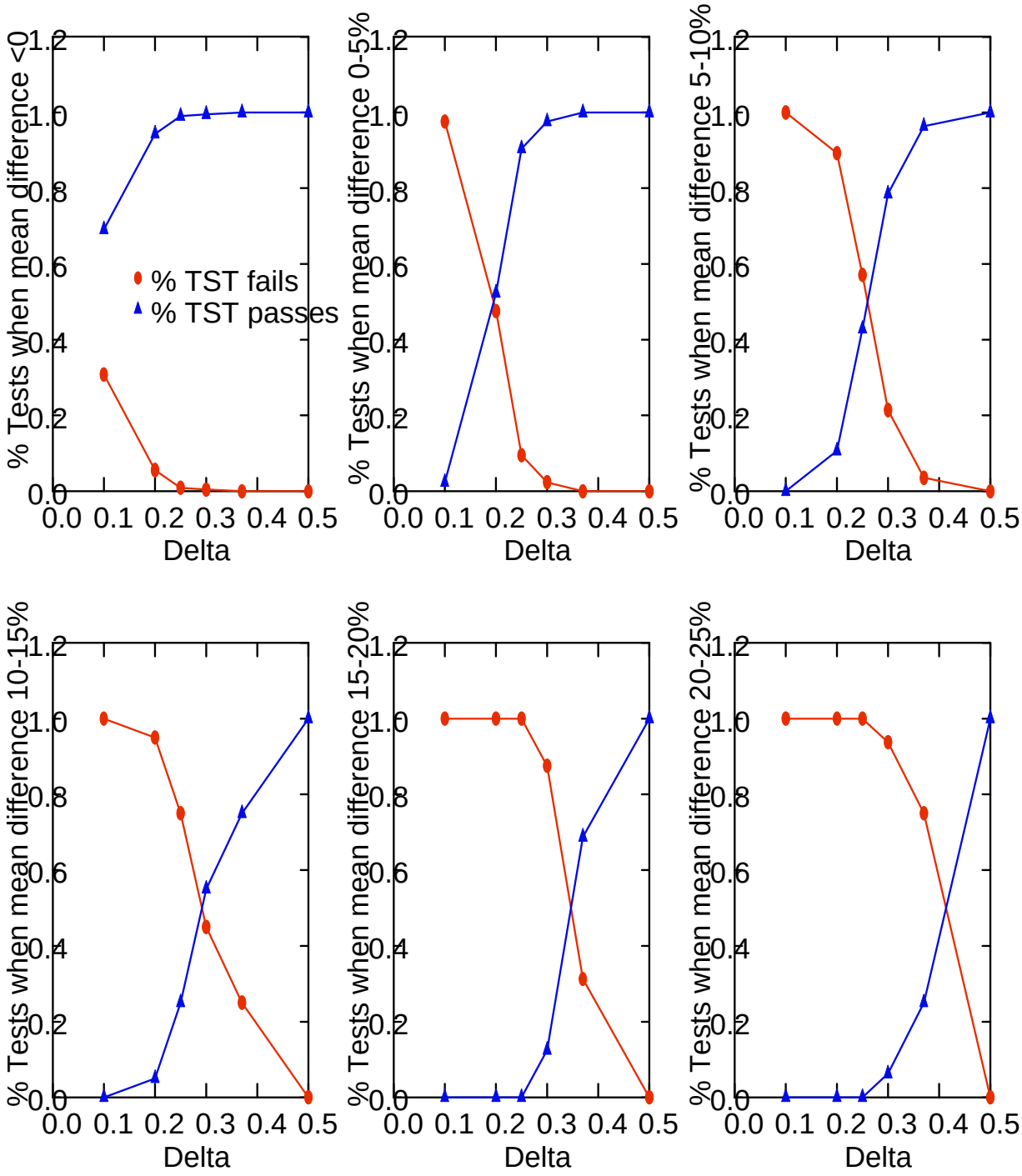


Power as a function of delta = 1-b



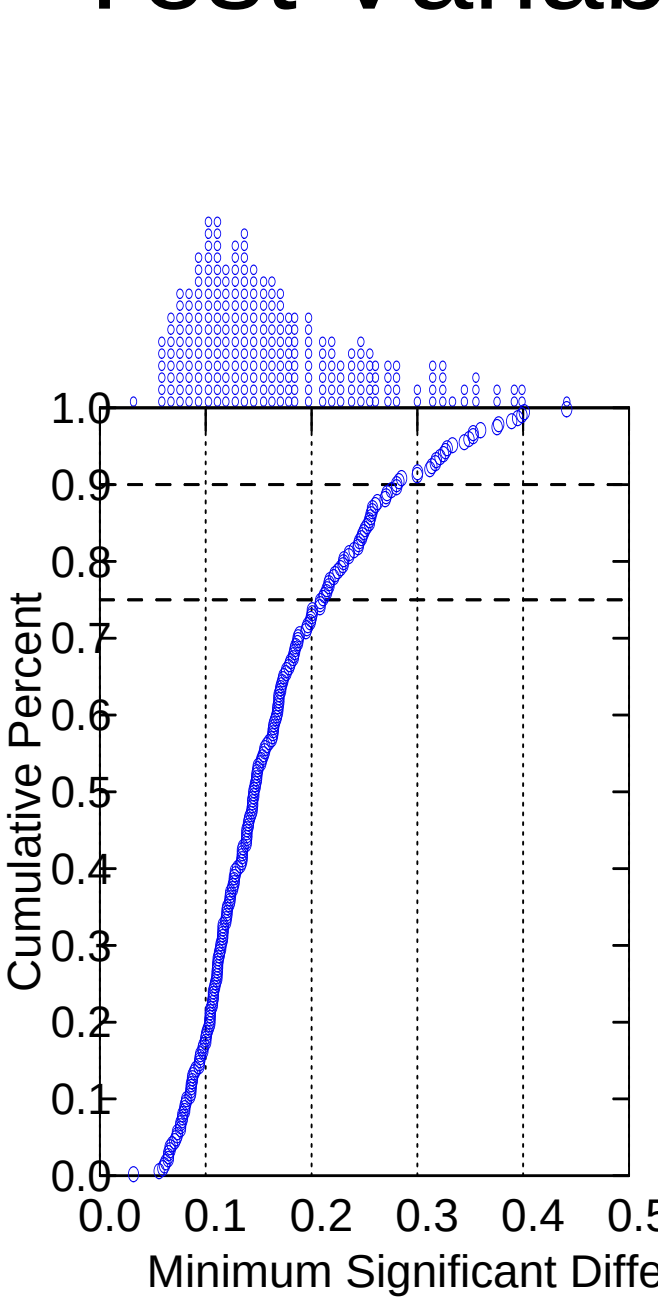


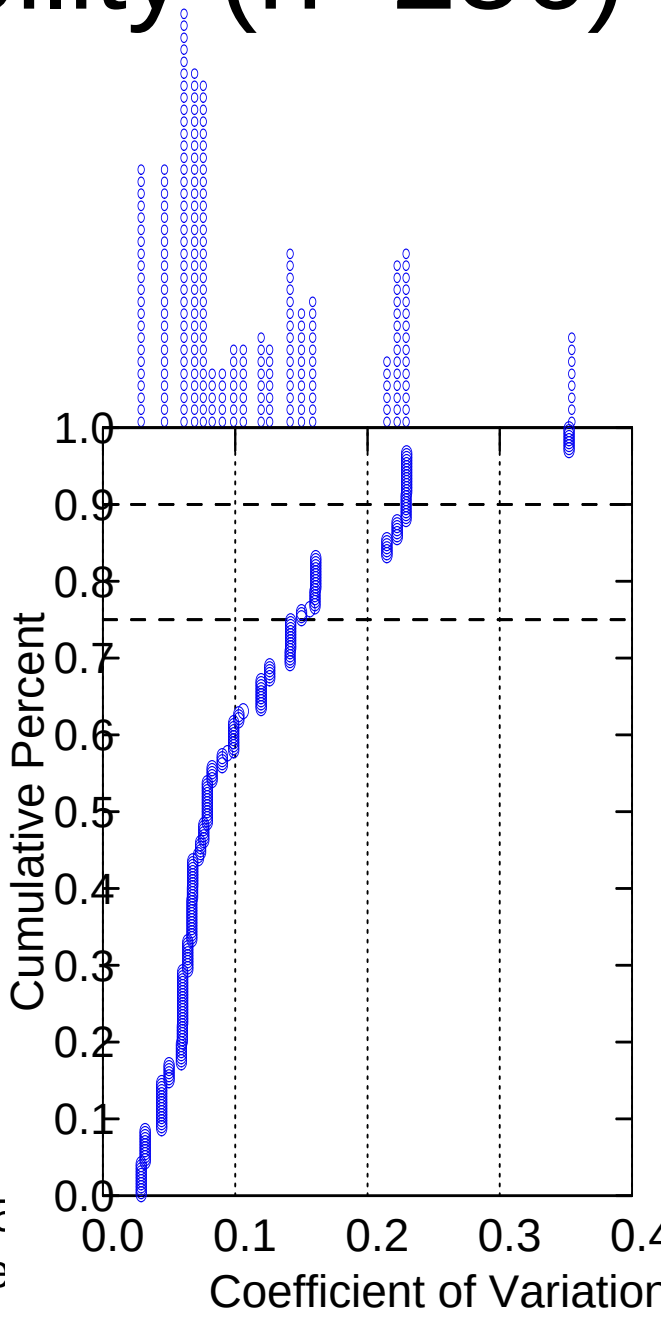
Detailed Analyses
delta = 1-b



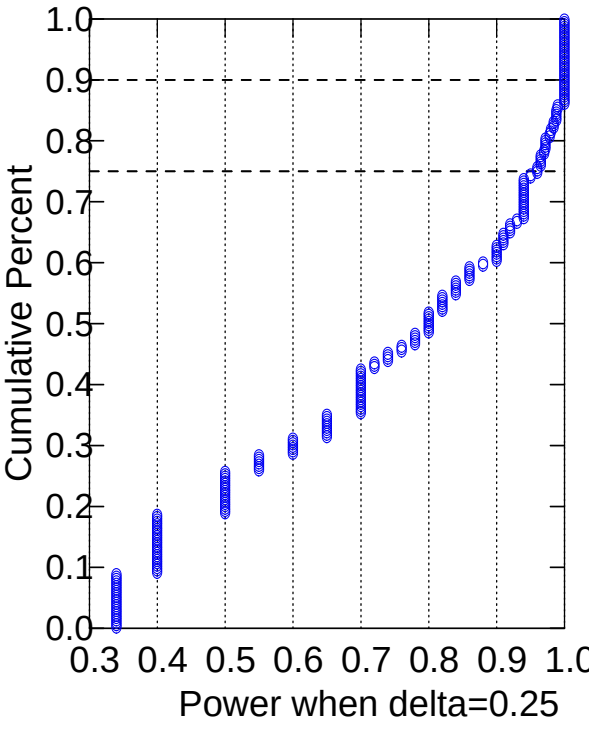
APPENDIX J
P. promelas Ambient Toxicity Data Characterization

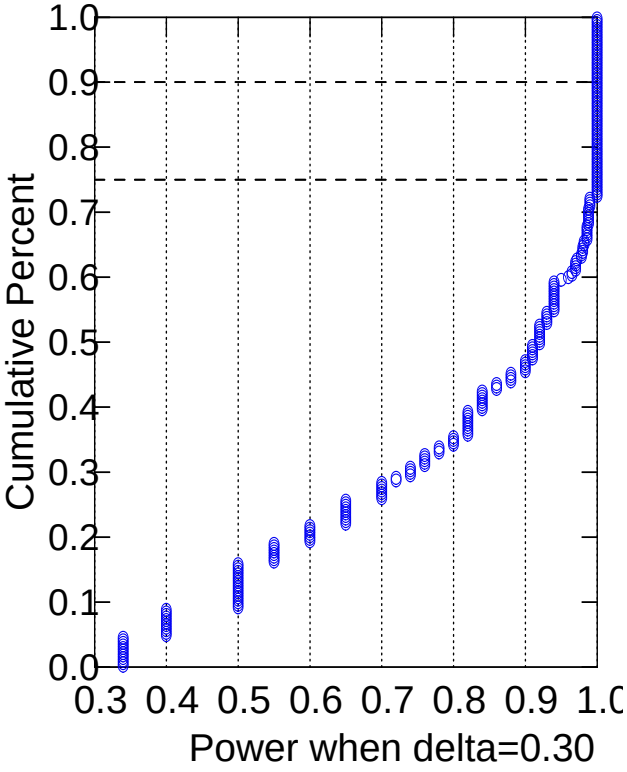
Test Variability (n=256)



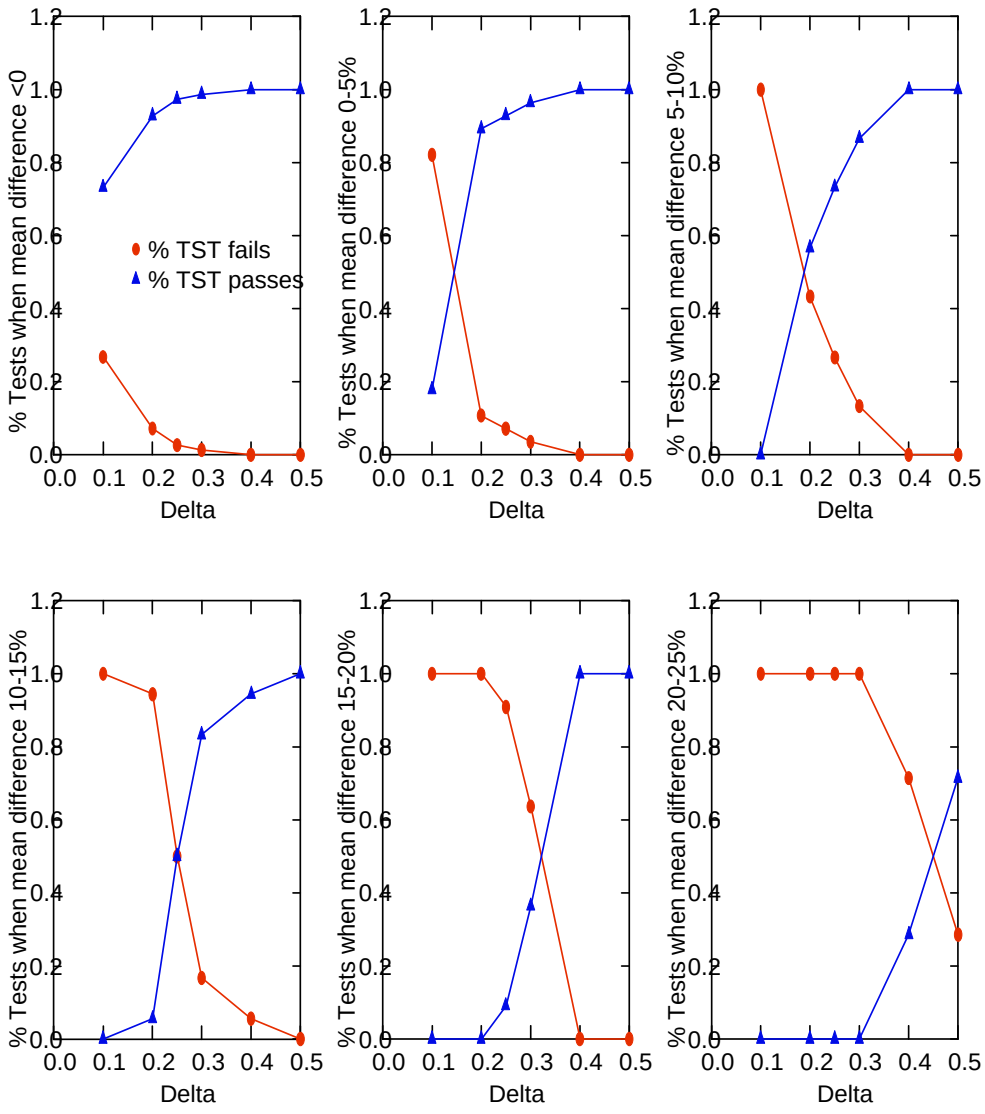


Power in relation to delta = 1-b





Detailed Analysis delta = 1-b



Summary of Peer Review of EPA Document:

Evaluation of the Test of Significant Toxicity (TST) as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity

October 20, 2008

Submitted to:

U.S. Environmental Protection Agency
1200 Pennsylvania Avenue, NW
Washington, DC 20460
CO: Kami Nolte
PO: Laura Phillips

Submitted by:

Avanti Corporation
5520 Cherokee Ave.
Suite 205
Alexandria, VA 22312

Contents

Executive Summary

A peer review of the “Evaluation of the Test of Significant Toxicity as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity” has been performed to obtain a critical, technical appraisal of the Test of Significant Toxicity (TST) Approach. Eight expert external reviewers were identified through Internet, literature, and professional association searches as potential reviewers. Potential reviewers were contacted, provided an overview of the project, and asked to determine their potential conflict of interest. The resumes of the potential reviewers were “sanitized” to remove any information that might identify the reviewers to EPA. The resumes were submitted to EPA for review to determine that their qualifications were sufficient to perform the peer review. From the eight reviewers approved, five were randomly selected and contacted to confirm that they could meet the time constraints of the project.

The evaluations of the TST approach document are provided in this peer review summary document. Section 1 provides an overview of the questions posed to the peer reviewers in the charge document. Section 2 provides each reviewer’s complete response. Section 3 provides a brief overview of the comments received for each question and compiles the comments by topic.

Reviewers were charged with review of the document with reference to:

1) Document's Merit
2) Document's Responsiveness
3) Document's Data Analysis Basis
4) Document Conclusions
5) Document Overall Quality
6) Other Recommendations

A summary of the results of the peer reviewer comments are as follows:

EPA Question 1 - Document Merit

Evaluate the conceptual soundness of the draft TST document's recommendations and the data analysis on which it is based. Is the draft TST approach an improvement over the current accepted hypothesis testing approach used in the NPDES WET program? If so, why, and if not, why not?

All of the commenters concurred that the bioequivalence method used in this study is a sound conceptual approach. Most also agreed that the TST approach is an improvement over the current accepted hypothesis testing approach used in the NPDES WET program. Commenters raised the issue of the method for selecting the value for *b.* Commenters also offered opinions on data analysis: a limitation of real world data is that estimated error rates are based on sample data means and not population means; without an objective standard of comparison, although a reasonable exercise, empirical studies can only provide a comparison of the methods.

EPA Question 2 - Document Responsiveness

Assess whether the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis.

Four of the five commenters agreed the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis. A dissenting commenter believed that hypothesis testing and TST approaches are not all that different in so far

as that both approaches are based on experimental designs reflecting the magnitude of the effect sought.

EPA Question 3 - Document Data Analysis Basis

Assess whether the data supporting the recommendations and conclusions on the draft TST *document are technically correct and defensible. The draft TST document attempts to evaluate existing data comprehensively, but: (1) for the purposes of standardizing comparisons, relies on data developed after 1995; (2) to be comprehensive, evaluates data developed using EPA WET test methods conducted under the current 2002 edition, as well as some earlier editions; and (3) to ensure that conclusions are based on appropriate data, censors some data points. The Agency's reasoning behind each of these aspects of the evaluation is explained in the draft document and related references (i.e., data test acceptance and quality assurance protocol).*

All of the commenters generally agreed that the data supporting the recommendations and conclusions are reasonable and defensible. There was a consensus that it was better to focus on current methods and future data. Commenters had an issue with an apparent assumption of normal distribution, and questioned if the assumption of a normal distribution can be made for all data used. A commenter had a set of specific issues critical of the documentation in the section of the document on the simulation method.

EPA Question 4 - Document Conclusions

Assess whether the draft TST approach as applied is technically defensible especially if challenged by either the NPDES regulated community, permitting authorities or expert consultants hired by permittees or other interested parties. Specifically, bioequivalency "b" values were derived for each test method using several risk management decision criteria which together, were intended to balance desired maximum alpha and beta errors at specific mean effect levels and within-test variability. Comment on the fact that this draft TST approach could be similarly used for additional WET test method(s) in the future. This draft TST approach builds upon EPA's earlier peer reviewed NPDES WET Variability document (USEPA 2000e) to derive and evaluate the "b" values. Evaluate the methodology used in the draft TST document to derive method-specific "b" values and apply the draft TST approach.

Commenters generally agreed that the TST approach is technically defensible. All commenters agreed, however, that the method in determining "b" values is critical to the validity of the TST approach and its acceptance by the regulated community. A commenter was concerned specifically about the way "b" was determined and also about the alpha values chosen. A commenter suggested including different statistical distributions to improve the robustness of the simulation results. A commenter suggested dropping the label "bioequivalency" as it conveys a potentially confusing meaning due to its historical use that is unnecessary to its meaning or use in the TST approach.

EPA Question 5 - Document Quality Overall

Provide any recommendations for how this draft TST document should be presented to the public (or the users of this approach) particularly NPDES regulatory authorities such as NPDES States and EPA Regions (the document will be revised to accommodate readers with a more Plain English version). Suggest, if possible, how it's highly technical content should be translated into a version more readily understood by the NPDES regulatory public (again meeting EPA's Plain English requirements) and yet maintain its clarity given its potential scientific, regulatory, and technical applications. Also critique whether a regulatory authority and their permittees would clearly understand the draft TST document's recommendations and if not how specifically should it be revised to make it easier to implement under EPA's NPDES permit's program.

Commenters were generally sensitive to how the document should be presented to the public. They also were quite critical of the presentation and clarity of the draft document. All of the commenters had substantial issues with the clarity, completeness, and grammatical errors that they found to be common throughout the document. One commenter noted the document needed to be re-written before a “plain English” assessment was attempted.

EPA Question 6 - Recommendations

Provide any recommendations to improve the draft TST document's technical basis and approach for deriving the alternative WET statistical analysis method in the NPDES permitting program.

Commenters stated that many of their recommendations were presented in response to previous questions. Recommendations that were reiterated include the following:

- Eliminate the term “bioequivalence” because it will be met with resistance among NPDES permittees.
- Present the decision criteria for selection of “b” in an explicit tabular form
- Commit to monitor, analyze, and assess WET precision to refine alpha and beta error rates and “b” values
- Selection of the level of “b” should be by consensus
- Present the TST method so that it is mathematically clear, including assumptions, steps, statistics, and criteria for evaluation
- Base simulation results on various non-normal distributions not just normal ones
- Use weighted calculations to estimate “b” and power, not simulations
- Cutpoints for “toxic” and “nontoxic” types of assessments are natural in the context of receiver operating characteristic (ROC) curves, and this type of analysis should have been included as part of this assessment.

1. INTRODUCTION

Five peer reviewers were charged with reviewing the document: “Draft Test of Significant Toxicity Approach as an Alternative to Current Recommended Statistical Analysis Approaches for Acute and Chronic Whole Effluent Toxicity.” Reviewers were asked to evaluate the document with respect to six review questions.

1) Document's Merit: Evaluate the conceptual soundness of the draft TST document's recommendations and the data analysis on which it is based. Is the draft TST approach an improvement over the current accepted hypothesis testing approach used in the NPDES WET program? If so, why, and if not, why not?

2) Document's Responsiveness: Assess whether the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis.

3) Document's Data Analysis Basis: Assess whether the data supporting the recommendations and conclusions on the draft TST document are technically correct and defensible. The draft TST document attempts to evaluate existing data comprehensively, but: (1) for the purposes of standardizing comparisons, relies on data developed after 1995; (2) to be comprehensive, evaluates data developed using EPA WET test methods conducted under the current 2002 edition, as well as some earlier editions; and (3) to ensure that conclusions are based on appropriate data, censors some data points. The Agency's reasoning behind each of these aspects of the evaluation is explained in the draft document and related references (i.e., data test acceptance and quality assurance protocol).

4) Document Conclusions: Assess whether the draft TST approach as applied is technically defensible especially if challenged by either the NPDES regulated community, permitting authorities or expert consultants hired by permittees or other interested parties. Specifically, bioequivalency “*b*” values were derived for each test method using several risk management decision criteria which together, were intended to balance desired maximum alpha and beta errors at specific mean effect levels and within-test variability. Comment on the fact that this draft TST approach could be similarly used for additional WET test method(s) in the future. This draft TST approach builds upon EPA’s earlier peer reviewed NPDES WET Variability document (USEPA 2000e) to derive and evaluate the “*b*” values. Evaluate the methodology used in the draft TST document to derive method-specific “*b*” values and apply the draft TST approach.

5) Document Quality Overall: Provide any recommendations for how this draft TST document should be presented to the public (or the users of this approach) particularly NPDES regulatory authorities such as NPDES States and EPA Regions (the document will be revised to accommodate readers with a more *Plain English* version). Suggest, if possible, how it’s highly technical content should be translated into a version more readily understood by the NPDES regulatory public (again meeting EPA’s *Plain English* requirements) and yet maintain its clarity given its potential scientific, regulatory, and technical applications. Also critique whether a regulatory authority and their permittees would clearly understand the draft TST document's recommendations and if not how specifically should it be revised to make it easier to implement under EPA’s NPDES permit’s program.

6) Recommendations: Provide any recommendations to improve the draft TST document's technical basis and approach for deriving the alternative WET statistical analysis method in the NPDES permitting program.

The responses from these five reviewers have been compiled into this peer review comment document and are presented below. Section 2 presents peer review comments on the above six

review questions sorted by peer reviewer. Section 3 presents a summary of peer reviewer comments for each of the review questions as well as each peer reviewer's comments sorted by question.

2. COMMENTS BY INDIVIDUAL COMMENTER

Commenter 1

1) Document Merit

The draft TST document's recommendations and the data analyses upon which it is based are conceptually sound. The specific recommendations from the document are: 1) to include calculation of TST for NPDES WET testing as well as for testing for ambient water quality monitoring; 2) provide incentives for permitees to provide high quality WET data to permitting authorities with basis in reasonable potential decisions as well as WET limits; 3) provide protection for receiving systems if WET test data have relatively large within-test variability or other inconsistencies while decreasing the probability of "false positives"; and 4) incorporate error rates into the decision process, thereby increasing confidence in test results. This is the essence of adaptive water resource management (the future for water resources in the U.S.). The specific data and analyses in the draft TST document included actual and representative data from over 2000 WET tests as well as appropriate simulated data to further examine the results or consequences of the analysis. The draft TST approach can represent an improvement over the current hypothesis testing approach if the WET testing program for NPDES purposes as well as the ambient toxicity testing continues as they have in the past. As noted above, the TST approach will encourage design of aqueous toxicity testing producing more precision or less variability in the data. Further, the proposed TST approach will permit identification of toxic samples in situations confounded by variable data and will minimize "false positives" (identification of nontoxic samples as toxic). The scientific community is 'evolving' to this approach for much of the experimental data that we collect relative to toxicity. Thus, the TST approach should serve to aid convergence of statistical significance with regulatory toxicological (ecological and biological) significance. Important for acceptance of this approach by the regulated community will be its implementation. I have included more specific recommendations in subsequent sections of this review.

2) Document Responsiveness

The draft TST document is responsive and meaningful in addressing some limitations of the current hypothesis testing statistical WET analysis. It is important to note that we can arrive at the right or "correct" answer through either approach. Both approaches need high quality data and the outcome of analysis of those data is dependent at least in part on the experimental design selected by the permittee or the permit writer. Both approaches have to recognize that high quality data are not necessarily more precise or less variable data. The TST approach offers some advantages in this regard. The TST incorporates the advantages of hypothesis testing and makes the risk management decision or level clear for the permittee. Further, the advantage of more precise data is clearly evident in the TST approach. The decrease in the incidence of identification of non-problems (false positives) should appeal to the regulated community. And the TST overcomes some of the current concerns with traditional hypothesis testing (e.g., *No Observed Effect Concentration* or NOEC) or the point estimate (e.g., *Inhibition Concentration* such as the IC_{25}) approaches (Crane and Newman 2000). Additional method guidance is provided in (USEPA 1995, 2002a, 2000b, 2000c) to analyze WET data and to determine compliance with permit conditions or water quality standards and this guidance provides permittees and permit writers some options for more efficient and effective experimental design.

3) Document Data Analysis Basis

The data supporting the recommendations and conclusions of the draft TST document are technically correct and defensible. For example, data from more than 2000 WET tests were used to analyze this approach. Use of the post-1995 data is justifiable since those data have been developed utilizing more uniform protocols and experienced testing laboratories. The test methods are sufficiently uniform to include them in the data set analyzed. The document emphasizes the importance of experimental design (e.g. replication) in influencing the outcome of a test or analysis. The data selection and processing SOP applied rigorous, logical and defensible criteria for inclusion of data in the sets that were analyzed. Censoring of data that did not meet inclusion criteria as presented in the SOP is appropriate. Importantly, the data were from numerous dischargers and testing laboratories to ensure that they were representative. Key to the initial success of the TST approach is the use of sufficient representative data in its derivation.

4) Document Conclusions

The draft TST approach is technically defensible and will likely be challenged by the NPDES regulated community or other interested parties. The challenges will likely come from those impacted negatively as well as the entities allied with those impacted negatively by the draft TST approach. However the TST approach is technically feasible and defensible. I think the term "bioequivalency" should be abandoned since it will convey an unnecessarily confusing message to the permittees (it has historical precedence that is of no value in this arena). The "*b*" values could be simply explained as an aggregate value accounting for the risk management decision (the level of protection desired for the receiving aquatic system), a balance of both false positive and false negative error rates at a specified effect level, and within test variability. For this TST document, derivation of the method specific "*b*" values was relatively clear and scientifically defensible. A decision support system could be readily developed that essentially codifies the process by which "*b*" values should be calculated (as outlined on pp. ii and iii and Table E-2). The TST approach has clear application to other toxicity testing in the future as data are accumulated and methods are refined.

5) Overall Document Quality

A thorough consequences analysis should be conducted prior to implementation of this draft TST approach. What are the environmental, social and economic costs of implementation? A data base consisting of those permittees that would be negatively impacted could be developed containing data from those permits formerly identified as "nontoxic" that would be identified as "toxic" as a consequence of implementation of the draft TST approach. Are there any data or is there any evidence that the affected parties or sites actually have demonstrable problems?

As with any regulatory program or change in a program, the devil is in the details. How would this new approach be implemented? Would there be a "phase in" period or an option period analogous to "monitor and report"? This is a great opportunity to institute a compliance period (e.g. one year to comply). Both the current approach and the proposed TST approach could be used initially in a 'monitor-and-report' mode and then the new approach could be included in issuance of a new permit. Both permitees and permit writers will need workshops and training materials to implement the TST approach. What does a single "significant" test mean? And failure to reject the null hypothesis of no difference or bioequivalence does not mean that there is no effect (i.e. does not mean that you accept the null).

A user's manual with specific examples (beyond those on pp. 24 and 25) and perhaps software to facilitate implementation will be needed (this is equivalent to the TSD). The present document needs some editorial (English) assistance for clarity and grammar. Some of the definitions are weak or incorrect (e.g. "Reasonable Potential", "Precision", Type II Error [beta]) and some

definitions are missing (e.g. “sensitivity”, “specificity”). Other grammar or “english” errors include subject-verb agreements and incongruous phrases such as “both simultaneously highest” (see p. 33). Failing to reject a null hypothesis is not the same as accepting the alternative hypothesis. This appears to me to be an error throughout the document (e.g. p. 6, 1st line; “accepting the alternative hypothesis – that the effluent is non-toxic” may be the practical outcome of the statistical test, but I do not think that it is the strictly scientifically defensible interpretation).

6) Recommendations

Some suggestions and recommendations are presented in previous sections of this review. The technical basis for critical aspects of the draft TST analysis such as selection of the value for “*b*” in each case should be explicitly stated in tabular form as well as scientifically or technically defensible decision criteria for their selection. Although the decision criteria for selection of “*b*” values are in the report, they should be assembled in tabular form or essentially codified. Permittees would justifiably object if such values are left to arbitrary selection at the level of implementation (e.g. region or state). The consequences of selection of risk management values should be clearly explained and illustrated. This can be accomplished in a user’s manual or through implementation guidance that includes appropriate t-test values (i.e. normal t-test, Welch’s t-test) and examples of toxicity tests with multiple ‘treatments’. The use of the “bioequivalence” terms or terminology will meet with resistance in the NPDES permit arena and should be avoided. It will spawn opposition that can be avoided if more acceptable terms such as “test for significant toxicity” are used. To further avoid confusion, you will likely need another term for “*b*” values since beta (β) is already co-opted by statisticians.

A “confirmation” data set should be assembled if the draft TST approach is intended to be implemented. This data set would contain those data from permittees or site that would be impacted by implementation. For example, data should be scrutinized from a permittee that was formerly designated “nontoxic” under the hypothesis testing analysis and would be designated as “toxic” under the TST analysis. An explicit example or two would go a long way toward convincing the regulated community that this is worthwhile. What is broken that would now be fixed and what would be identified correctly as not broken and not in need of fixing? What will happen to the notion of accelerating testing with indication of a problem (i.e. a failed test)? There are considerable advantages in having additional data to support a decision in this case.

Some maintenance (monitoring) will be required for successful implementation of the TST approach. This will involve periodic evaluation of the precision of WET and other testing to establish and implement alpha and beta error “rates” and the “*b*” values. Commitment to this maintenance will need to be clearly stated (along with the frequency) in the implementation phase. The apparent interpretation that decreased variance in toxicity testing results indicates higher quality data may not be supportable. An alternative situation is that variance in the data indicates a varying process and variance in the sample. So variance may have to be partitioned (within test variance vs. among test variance vs. sample source or generator- specific [industry-specific] variance). Reference toxicant data should be inherently less variable than source sample data, although the opposite is indicated on p. 29 for *Pimephales promelas* chronic growth tests.

In summary, I think the TST approach represents an improvement relative to existing statistical techniques. However, the opportunity to realize the potential enhancement in the aqueous toxicity testing programs depends strongly on the implementation strategy employed.

References:

References were provided by Commenter 1 but removed so as not to potentially jeopardize commenter confidentiality.

Commenter 2

1) Document Merit

The concept of an equivalence test is admirable. Equivalence tests provide a proof of safety while the NOEC provides proof of ignorance. The proposal to use an equivalence test is a big improvement over the current hypothesis test.

The weak part of the document is the rationale for choosing *b*, the equivalence factor. In a traditional equivalence test, *b* is determined by subject matter considerations. For example, FDA guidance for generic drug evaluation considers generic and branded drugs to be equivalent when important pharmacological properties of the generic are significantly between 80% and 125% of those for the branded drug. That equivalence region (80%, 125%) is a biological conclusion. Applied to WET evaluation, the analogous procedure would be to determine a level of effect that is considered important, i.e. not safe. This determination will not be easy because it relies on a lot of ecology. It may be partly a guess. It requires discussion among stakeholders. But, with stakeholder participation, *b* summarizes what is considered not safe. Once *b* is determined, the approaches in this document can be used to evaluate the properties of the test (alpha and beta levels given certain amounts of variability).

The approach in the EPA document is backwards. That is, *b* is determined by considering test characteristics (the risk management criteria). The choice of *b* is driven by the variability (c.v.) of the tests. In other words, safety (or not safe) is determined by the precision of our measurements. If tests were more precise, *b* would be closer to 1.0. This creates a logical conundrum: what do you do if WET tests get more precise (as data briefly discussed in the document suggests has happened since the early 1990's). If a WET test becomes more precise in 2010, do you compute a new b? The logic used in this document says you would. Common sense says this is nonsense. If *b* increases from 0.7 to 0.8, then an effect of 25% reduction from control is safe now (when $b = 0.7$) but not safe in 2010 (when $b = 0.8$). It's the same effect, with the same ecology.

In summary, the document is an improvement over the current hypothesis testing approach. An equivalence test is much better than the NOEC. It would be a huge improvement if subject matter considerations were used to determine *b*.

2) Document Responsiveness

Yes it is. No further comments

3) Document's Data Analysis Basis

The data sets have a minor role in the evaluation. Primarily, the data provides estimates of the c.v. among samples exposed to control and effluent. These estimates are used to provide realistic distributions for the simulation studies. The rationale for subsetting and censoring are very reasonable.

4) Document Conclusions

I repeat my concern with the method of choosing *b*, raised in point 1.

I also repeat my general view that an equivalence test is an appropriate way to demonstrate safety in many toxicity tests. I am concerned with how alpha levels were computed. One detail in the tables in section 4 doesn't seem right. This is either a major failure in the computation or a major failure in communication. Details of my concern are given below in the detail for "Table 4.2 et seq".

5) Overall Document Quality

The document, as currently written, is very hard to understand. I struggled in many places, in spite of my background in these issues. Even after a month of struggling, I'm not sure I completely understand all that you did. If some of my comments don't make sense, it is probably because of my inability to understand what you did.

Will this document be the primary source documenting the TST? If so, it needs to include a worked example of a test, explaining and illustrating how the test is actually done. You do this on p 24, but that is somewhat hidden. The example needs to be prominent.

Parts of the document are repetitive (details below). It needs a thorough proofreading (some specifics included with my detailed comments). Some tables and figures need a major reorganization. Will this document be printed in color? If not, Figures E-1 and 6-1 become very hard to read. Using lines to indicate the rejection regions for the two tests avoids the need for color.

6) Recommendations

Overall comments reiterate what was said in point 1. An equivalence test, like the proposed Test of Significant Toxicity, should be the required statistical method for the analysis of WET data.

My only concern is the proposed method to choose the safety bound, *b*. It should be set by a consensus opinion of what biologically represents not safe. If a stakeholder based choice of *b* is impossible, the proposed method and choices of *b* lead to statistical tests with better properties than the current NOEC. This alone is a substantial improvement.

Other Comments:

The document can be improved by a careful revision. My comments are marked by page and line, with negative numbers indicating lines from bottom:

p 7, l 3. How is statistical power incorporated into the TST? The statement is confusing because the TST also has a power. Which power is being referred to here? The TST and NOEC test have very different null hypotheses. The only connection (and hence the only way the power of the "usual" test of no difference is incorporated into the TST) is because all tests and all powers depend on the coefficient of variation. This claim is repeated many times in the document.

p 10, figure 2.1 Labeling of 'average' on the figures. If the dots are the 90th percentile, I can't see how the lines are the "average CV", as claimed. Perhaps these are the average 90th percentile, but that's not the average coefficient of variation.

p 16, table 2-3. This is a repetition of table 1-2b.

p 16, l -4. "Monte Carlo" is an adjective not a noun, unless you're referring to the city state.

p 17, l 4. "ground truth" Is the truth known in the actual WET data? I doubt it. Without knowing the truth, how does the analysis of actual data sets "ground truth" the simulation?

p 18, l 6. The preceding paragraphs describe the choice of effect size and variability. Properties of a statistical test also depend on the degrees of freedom for the error (i.e. whether variances are pooled, the number of treatments, and the number of replicates). These values affect the power and the relationship between the TST and the usual test. Table 2-1 gives the **minimum** numbers. This makes me suspect that there was variability among tests. How did you choose the rest of the details for each test? If your calculations used a single degree of freedom error, say so, and report that degree of freedom error.

p 16, lines 12 – end of page. This is really important stuff because it describes how you translated risk management criteria into an evaluation of both tests. It is a key feature of your analysis. It must not be buried inside a section described as 'Freshwater and East Coast …'. I found this section very hard to understand. Since it lies at the heart of your evaluation, it needs to be prominent and clearly written.

p 19, l 4. This sounds like a fourth criterion. (or are you implying that a test should satisfy all criteria). If so, say that.

p 20, l 8. Where is exhibit A?

p 20, l 15. The arc-sine square root transformation is a **variance stabilizing** transformation. It does not correct for non-normality.

p 20, l -10. I don't see why four replicates matter here.

p 20, l -8. You simulated data for four replicates, even though reality is only 2 replicates. That means you are reporting properties for a non-existent statistical test (using 4 reps).

p 20, l -3. This seems like you're trying to force non-normal data into a t-test framework. This introduces all sorts of complications, as you discuss. However, I'm not sure that alternatives (e.g. binomial exact tests or beta-binomial tests) are any easier. They certainly are not part of the standard statistical toolbox.

p 21, l 8. I'm not sure what you're doing here. In fact, I'm completely confused by what you say you're trying to do. The problem with establishing alpha and beta from actual data is that both require that you know the true difference (is it zero or not). You don't know that.

One ways to bypass this problem is to use only the control data and add a specific effect into the effluent mean. I couldn't tell whether you used observed test data 'as is' (without knowing truth) or whether you added known values to control data.

p 21, l -6. This is a very unusual definition of alpha level. It isn't clear whether 'exceeding the toxicity threshold' applies to population quantities or sample means. If this is based on sample means, the computation is completely wrong, since sample means often exceed (and even as likely as 50% to exceed) the population quantities.

Also, aren't the definitions backwards, since % effect is calculated as (control – effluent)/control, so a large number is a 'bad', i.e. toxic result.

p 21, l -2. The number of data sets has nothing to do with their appropriateness for evaluating the simulations.

Figure 2.2. This is one of many examples of poor graphics. Specific problems include:

a) the legend describes %, with a range of 0 – 100%, but the y axis is scaled from 0 – 1.

b) the y-axis label is the title of the plot, not a description of what is plotted on the y axis.

c) the two lines are redundant. One is 1- the other.

p 23-24. 3 issues:

a) Isn't this out of place and said earlier?

b) Distributions of t: these aren't very useful, since the t statistics are very different for the two tests.

c) l -2 on p 24. Where did 25% effect come from? b is 0.68, which translates into a 32% difference from the control. Same issue on next page.

p 26, all. This material needs to be combined with that in section 2.4, which also discusses QA.

p 26. l 3. What is “Table 3”, since tables are numbered as section-table?

p 26, l -8. What is computed in Excel?

p 29, table 4-2 (and all tables through 4-8 that are similar). These baffled me for far too long. The really important piece of information is the coefficient of variation for each test, which is hidden in footnotes. Nothing in this table made sense until I realized you were changing coefficient of variation for different rows.

I suggest a complete reorganization of this and comparable tables for other tests.

Effect level %	c.v.	Risk Criterion	B value				
			0.63	0.68	0.70	0.75	NOEC
15	≤ 75th%	Toxic < 0.2	0.0	0.001	0.003	**0.35**	**0.99**
20	25th-50th%	Toxic < ??	0.00	0.02	0.21	1.00	1.00
25	> 50th%	Non-toxic < 0.05	**0.99**	0.00	0.00	0.00	**0.78**
25	< 25th%	Toxic = 0	**0.90**	**0.19**	**0.62**	**1.00**	**1.00**
30	Any	Non-toxic = 0	**0.99**	0.00	0.00	0.00	**0.22**

This suggested revision a) includes the coefficient of variation as a specific element of the table, b) eliminates the meaningless columns (the –‘s), and c) indicates where the risk management guidelines are exceeded (by the bold entries). Some of my entries for the 20% effect line are hypothetical, because I couldn’t figure out the important information for that line in your table 4.2. If you feel that combining toxic and non-toxic endpoint s is too confusing, then separate the above table into a part for the toxic endpoints and another part for the non-toxic endpoints.

p 30. Figure 4.1 Why didn’t you use connected lines, as in figures 4.2 and 4.3?

Table 4.2 et seq. Is the value of 0.00 for the alpha level for 30% effect and b=0.70 correct? I think there is either a major failure in the computations or a major failure in the communication of how these results were computed. Here’s why:

The usual interpretation of 30% effect is that control mean - effluent mean = 0.3 * control mean. When b = 0.70, then the effluent mean is exactly on the boundary of the equivalence region (because the boundary of the equivalence region is effluent mean = 0.7 control mean, i.e. control – effluent = 1 – 0.7 = 0.3 = 30% effect). When the population mean for the effluent lies exactly on the boundary of the equivalence region, the TST should have alpha = 5% no matter what the coefficient of variation is. This is follows from the construction of the test and the definition of the alpha level. I am very confused why the reported values are 0.00 (e.g. for 30% effect, b = 0.70, non-toxic (alpha) reported as 0.00) .

p 32, line -6. You refer to table 4-1. Shouldn’t this be table 4-4?

p 44. The pie charts are terrible graphics. The false 3D only hinders the visual interpretation. A table presents the information much more concisely. You essentially include the table information when you give the actual %’s for each category. A graphical alternative is a mosaic plot.

p 44. This approach, a pairwise comparison of TST and NOEC test results for each actual data set, would be a very good way to summarize results for all the ‘actual data’ analyses in the previous sections.

p 47, lines 3 et seq. This seems to be based on the simulations. This claim would be much stronger if was based on the actual data. That is, you could use the actual WET data, compute both the NOEC and the TST and compare the results using the approach on p 44. More protective is then shown by t-test fails = 0 while TST fails = something larger than 0.

Figure 6.1 (and also E-1). Most of this graph is blank space because the largest % effect is 40%, but the Y axis maximum in 100%. Re-draw the graph with Y max = 40% to focus on the interesting stuff.

p 54, l -5. Again, I don't see how power is incorporated into the decision process because power of the t-test has little to do with the decision from the TST.

p 55. l -5. Need year for the Grothe et al citation.

Commenter 3

1) Document Merit

The proposed approach is an improvement over the current, accepted hypothesis testing approach used in the WET program. The TST approach or what I would call the bioequivalence approach is a well-studied approach that is commonly used in biomedical studies. The problem with the current WET approach is that the type II errors are not controlled by the procedure. Because of this, I would expect that there is a tendency to not reject an effluent as toxic. The TST approach basically switches the null and alternative hypothesis, putting more burden on industry to show the effluent is safe. The novel addition to the TST approach is that there is an attempt to account for both types of errors by recommending the value of the bioequivalence factor (b in the report).

I have two major problems with the TST approach relative to the current approach. First, the formulas for the TST approach are never defined in the document. It is not clear if the intention is always to use a modified two-sample t-test with the assumption of unequal variances or if other tests are used in the data analysis. Second, the current WET test approach provides a roadmap for decision making that is intended to help a user decide which test to apply in a given situation (i.e. equal or unequal variance, transformation, etc). Such a roadmap could easily be done in this document.

2) Document Responsiveness

There has been a need to revise the WET approach for quite a while as the design component of the approach has never been a strong component. This has resulted in a limitation to the testing approach. The two main problems have been the need to control variability and the need to control error rates. The TST approach is one way to do this by essentially switching the null an alternative hypothesis. The null essentially becomes that the effluent is toxic and the role of the laboratory study is to prove otherwise.

3) Document's Data Analysis Basis

The simulation that is used to determine the value of the bioequivalence parameter relies on the assumption that the data that are collected are representative of WET tests that are carried out in the future. Therefore they try to obtain a more or less balanced sample of the current data rather than a census of current data. The census approach would give more weight to laboratories that carry out a large number of tests. I think it is quite reasonable to focus on current methods and data rather than past data.

The section on the simulation method leaves much to be desired. For the simulation to be valid there is an assumption that the screened data represents the population of WET testing. It is not at all clear what is being simulated. It is not clear why the control group is truncated. What is the

justification for this? It is not clear from reading the document what the data are. For example, for Ceriodaphnia do you have counts or is it simply the mean and coefficient of variation? I do not understand why the coefficients of variation are randomly selected. For a t-test one would just have to sample means and standard deviations. Is “n” fixed? It is not mentioned on pages 17 or 18 yet is critical for power. I think the authors also assume that the data are normal. This is not justified.

After reading this several times I think this is what was done:

Use the screened data from controlled samples to obtain an empirical distribution for the mean and coefficient of variation

Select 1000 samples of means and coefficients of variation from the trimmed empirical distribution

Compute the test statistic for the TST test for different means and coefficients of variation. For each test statistics decide to accept or reject the hypothesis

Compute the number of rejections in the 1000 tests.

Some things that are missing:

What is the sample size? Is the sample size for control treated the same as for the effluent

Was the control sample used with all levels of effluent mean and coefficient of variation (as in a block design) or was a new sample selected for each effluent level.

Was the choice of 1000 justified? For power calculations, I am used to having 10,000 simulations rather than 1000 as this sample size will lead to smaller simulation error.

What are the formulas for the tests?

What are the assumptions?

It appears that the mean and coefficient of variation were sampled independently. I think this is not justified. The mean and standard deviation are attributes of the individual WET test. I think they should be sampled together rather than separately.

I also do not understand why one would sample mean and coefficient of variation rather than mean and standard deviation. For the test statistic, the standard deviation would have to be calculated. Are the mean and coefficient of variation independent?

I question whether a Monte Carlo simulation is necessary. I think a more accurate approach is a direct calculation. Given the means, standard deviations, and a value of b compute the power of the test assuming normality. Then weight by the probability of the mean and standard deviation occurring in the population. Repeat this for a set of means and standard deviations (over a grid, say) and then compute the weighted average power. This approach should be more accurate.

It is also common to use cross validation in these cases to evaluate error rates. Since the value of b is calculated using all the data, shouldn’t a test using the value of b be evaluated on a separate data set?

Appendix A seems to have been pulled from the grant proposal as it is written in future tense rather than present or past tense. In reading this section, there is an implication that the EPA flowcharts will be used to select a test. This may result for example, in a two-sample t-test assuming equal variance (which should have greater power that the test assuming unequal variance). It is not clear that the same approach is applied to the TST procedure. What is the justification for this approach?

One additional criticism is that if the TST is an alternative to the WET approach, the simulation needs to compare the “approaches” not the “tests”. The simulation as described compares the

tests. The WET approach allows for a variety of tests depending on the assumptions. The conclusions therefore apply to a test not the approach.

4) Document Conclusions

The bioequivalence approach is a valid statistical approach. It does not depend on this document for validation. The choice of b is critical in the method. The value in the document is based on the two error rates. If these error rates are reasonable then the approach has justification.

5) Document Quality Overall

I think the writing is rather weak and the document needs to be rewritten before the Plain English requirement is attempted. There are many minor problems with the descriptions, examples and general layout of the document. Many specific examples of problems are given below. To improve the clarity of the document, I think there is a need to carefully consider the layout of the document, the position of the figures and tables and the order of these figures and tables. There is a general lack of formulas and inconsistencies given when formulas are used. Although formulas might not be a good idea for some readers, they give a much clearer description of what was done than the written description. For example, to simply state that a two-sample test was done is not precise because variances could be equal or not. To be precise, I think the formulas should be general. The general description of the TST approach should include null and alternative hypotheses, assumptions, the test statistic and the criteria (distribution to be used along with degrees of freedom). One cannot infer these from the examples because the examples are not general. They are based on equal sample sizes and include nothing about how to calculate degrees of freedom.

6) Recommendations

Write out a specific description of the TST method that is mathematically clear. Include steps, assumptions, statistic and criteria for evaluation.

Write out the assumptions of the simulation and clearly state the steps used in the simulation.

Change the order of the presentation to ensure that figures, formulas, tables, etc are in the correct location for understanding the document.

Use weighted calculations rather than simulation to estimate b and power.

Other Comments:

Line numbers would help with the review process

Figure E-1 What is the purpose of the colors? If these mean something why is it not described in the figure legend?

Page iv top low error rates than t-test should be lower error rates than the t-test.

Figure E-2 What is the standard t-test. The two sample t with equal variance? What is meant by TST fails? It is difficult to see that the TST test is better unless a higher fail rate is better. In fact, one might argue that the degree of concordance is quite high for the two approaches.

The results of the project listed on the bottom of page iv really don't have anything to do with the project; these statements basically can be made just from reading papers on bioequivalence.

Table E-4. What is the difference between specificity and relative specificity? Here is the medical definition of relative specificity: The specificity of a medical screening test as determined by comparison with an established test of the same type. I am not even sure that what you have calculated is correctly termed specificity since specificity implies knowing the true state.

Page xi Glossary: why not use standard definitions of terms rather than creating misleading ones. For example: hypothesis test, power and significant difference (why connect this to a confidence interval) are not consistent with statistical definitions of these terms. Type I and II error are expressed in terms of hypotheses but power is not.

Table 1-1:

NOEC disadvantage 2: does not explicitly estimate statistical power – I don't think any method explicitly estimates power – I think you mean controls statistical power

NOEC disadvantage 5 is also a problem with estimation of an IC50. It is also a potential problem with the equivalence approach.

Point estimate disadvantage 4: confidence intervals are also affected by assumptions

Page 1. Line 7: the NOEC is not a hypothesis test rather it is level (concentration or dilution) that is derived from a test of hypothesis.

Line 16: the LC50 is an estimated value and it seems to me answers the question at what concentration is the 50% effect predicted. If the WET test uses a criterion for evaluation that is based on a permitted IWC less than the IC25 it seems to be ignoring the uncertainty in the estimate of the IC25. Is this not a problem?

Page 4 line 9. Power depends on the variability of both groups (see figure 1-1). I think what you really want to describe here is the potential for a small biological effect being significant. If any decrease in survival relative to the control is an indication of toxic then is this not relevant?

Figure 1.1 Very small intra-test variability – This would be better if there is not a pattern in the second group of data i.e. two lines of data. I would change the figure to show no linear pattern. The use of "very small" to characterize variability seems odd – why not just use "low".

Table 1-2B: I think "b" should be defined in the table since you take the opportunity here to define Type I and II errors. Should b<1 also be added? I think the order of table and figures needs to be looked at as it seems that table 1-2B should come before fibure 1.2

Page 6: It seems that the main value of the bioequivalence test is that it puts the burden on industry to use a sufficient sample size that has good power for the test. If sample sizes are not sufficient the effluent will not be declared "not toxic". Why would the approach be better than to require a post-hoc power analysis i.e. require the industry to have power of 0.8 for a change of say 30%?

Page 7: I am not sure what is meant by "maximum" desired alpha and beta rates. It seems one would want to minimize these or to be as close as possible to specified rates.

Page 8: the project objectives are not well written. The first paragraph is an awkward read. First, the primary objective (purpose) is stated (which seems to actually be two objectives). Then the second sentence lists another set of primary objectives.

Figure 2-1. Is it appropriate to talk about the coefficient of variation as test variability?

In reading the document, I was very confused by the connection between the "test" used to statistically evaluate an hypothesis, the data and the simulated data. I did not find in the document a formula for the bioequivalence test statistic. Is the test being used just a variation of the two sample t-test? If so, why not write it out. I think it also has to be connected to Table 2-1 which describes typical data for the study design. For an uninitiated reader it would be useful to know what the minimum number of effluent concentrations corresponds to. For the bioequivalence test, is the reference compared to all of these or just the full concentration? Table 2-1 needs to connect to section 2.5 and the test statistics. In addition, if different methods are

applied to the data from these studies in different situations, should you not also consider different tests for bioequivalence?

The study is based on the assumption that the data represents a sample from laboratories or facilities that is in some sense probabilistic. This is needed to treat the screened sample (20 most recently conducted tests) as a sample from a population. How can one tell if this is an unbiased collection of observations? It might be useful to know what the universe of samples is and how the observed facilities and laboratories represents this universe. What are your assumptions?

Page 15: There is a need to be explicit as to the formula for the test statistics, the assumptions, degrees of freedom, etc., for the bioequivalence test.

Although beta=0.8 is probably the first size of power that comes to mind for most statisticians, there is also an argument for an alpha:beta compromise.

Table 2-3: I like that the hypotheses are defined in terms of parameters. However the mean of the treatment is defined incorrectly. It is not the response of the effluent concentration rather it is the mean response to or the population mean response to… Why then switch to effluent<b*control which is not very descriptive?

Section 2.5.2 Is “several different” good grammar?

The notation used on page 23 bottom is confusing. Sc is referred to as the variance of the control yet in the formula Sc2 is used. Why use SDEVc rather than simply Sc. I would use the standard notation Sc2 for a sample variance and Sc for a standard deviation. Include degrees of freedom and critical value. I think it should be □=.05 not □<0.5. A clear way to write it is t.05(1),18=1.73. It also seems that Step 1 should be to state the hypotheses (this way one knows it is a one-sided test).

I think the example using the unequal variance case (page 25) should provide a general formula not a formula for the special case when Nc=Ne. Otherwise this could be misleading for someone using the example as a template for their data.

Page 24: the standard statistical statement is “do not reject” the null hypothesis rather than “accept” the null hypothesis.

Page 27 why mention Ceriodaphnia is a freshwater invertebrate – water flea when this is mentioned in the header for 4.1. It is stated that 65% has power >= 0.8 and 35% is <=0.8. Should one of these be a strict inequality?

Appendix B. Are results based on 1000 simulations?

The axis legend is often cutoff i.e. Minimum Significant Differen. The truncation of the labels occurs throughout the graphs in the appendix. Also the numbers on the y-axes on many of the graphs are difficult to see as the overlap the axis line.

Commenter 4

1) Document's Merit

I agree with the central idea around the TST approach: that changing the NOEC null hypothesis (the effluent is not toxic) to the TST null hypothesis (the effluent is toxic) will likely improve the detectability of toxic effects. I also agree with changing the hypotheses by including the b bioequivalence factor to avoid the possibility of categorizing highly precise results as toxic when the effect is small. Whether the TST approach is an improvement over the current approach will ultimately depend on the choice of b. As shown in the data and report, a too-low value of b would reduce the ability to detect toxic effluents below that of the current approach. The document justifies the chosen b factors for the evaluated methods reasonably effectively. As discussed in

the later items, however, additional analyses may be necessary to assess the robustness of the estimated b values to departures from various data assumptions.

I think that including data analyses based on both simulated and real-world data is generally a good one. However, the real-world data have the limitation that the categorization of “toxic” vs. “non-toxic” can only be made based on the mean of sample data, and therefore the estimated error rates would reflect those means rather than the population mean being estimated. Therefore, the real-world data analyses do not add much support to the document recommendations and in some cases could undercut them. Data analyses using the Monte Carlo data were generally sound; however, additional conditions should be included in the simulations, as discussed in Section 3.

2) Document's Responsiveness

Table 1-1 presents many of the limitations of the current NOEC approach fairly clearly. However, the TST approach does not address all of these limitations because both approaches use the same data. Therefore, only the disadvantages 2 (Does not explicitly evaluate statistical power) and 3 (No incentive for permittee to increase test precision) are by the TST approach. The data analyses described throughout demonstrate that the TST approach addresses these two issues better than the current approach.

Using the Monte Carlo analyses, it is demonstrated that toxic effluents are detected more effectively under the TST approach, and that the likelihood of the TST approach incorrectly categorizing non-toxic effluents as toxic decreases as variability decreases. High false positive (concluding that a test concentration is toxic when it has a low mean effect) and false negative (concluding that a test concentration is non-toxic when it has a high mean effect) rates for the NOEC approach can be found in the appendices and the results descriptions. It may be beneficial, however, to summarize these more succinctly using tables/graphs (see item 5 for further discussion).

Unlike the simulated data, the real-world data presented and analyzed in the draft document do not clearly show the stated limitations of the NOEC approach. Tabular results are presented in the Appendices for simulated results only, and the figures presenting real-world data are more limited than those presenting simulated data. Fore example, Figure 5-1 shows the frequency of the approaches coming to different conclusions (approximately 8% of the time for Ceriodaphnia dubia and 12% of the time for *Pimephales promelas*), but does not show these rates vary by observed mean effect. Additionally, Table 6-2 gives the sensitivity and specificity for the TST approach only.

The discussion of the real-world results also does not strongly support the stated limitations of the NOEC approach. In the conclusions section, it is stated:

“7% of the 256 tests (18 tests) were declared non-toxic using t-test, despite mean effect levels as high as 33%. In addition, 5% (13 tests) were declared toxic using the t-test approach at effect levels as low as 17%. By comparison, TST rarely declared samples as non-toxic at mean effect levels > 20% and rarely declared samples as toxic at mean effect levels < 20%.”

This does not strongly support the stated limitations of the NOEC/t-test approach, as the audience may not see a large difference between 7% and 5% rates and “rarely,” especially because it is unclear which of the two approaches is giving the correct answer for those data. While rates for specific mean effects were not presented for the data presented in Section 5, other methods yielded rates higher than those quoted above when the toxicity decision was based on the TST approach for the chosen b value. For example, 25% of tests were declared as toxic when the mean effect was ≤20% using b=0.8 for *Macrocystis pyrifera*, as shown in Figure 4-6.

3) Document Data Analysis Basis

In terms of the range of WET test methods and time period of data development, the choice of data appears to be fine (though I do not have the full knowledge of different WET tests to comment on this fully). However, there are a few factors that either were not included in the data, or were not assessed in the analyses:

A t-test was used for all analyses to assess both the NOEC and TST hypotheses. This suggests that all of the results followed a normal distribution (after any data transformations). In Section 2, it is stated that Welch's t-test is used when the assumption of equal variances is not met, but there is no discussion of whether the issue of non-normality occurred with the data. The Monte Carlo results may have been simulated based on the assumption of a normal distribution, but it seems unlikely that all of the real-world data would have followed a normal distribution, especially when the CV was high. Even if this was the case, it would be worth simulating data following some non-normal distribution. The Wilcoxon Rank Sum test is often used in NOEC analyses when the normality assumption is not met, and could easily be used for TST analyses as well, by adjusting either the control or effluent results by b prior to ranking. The statistical power of this test would likely be less when a nonparametric test is necessary, which would decrease the probability of concluding a truly toxic effluent is toxic under the NOEC approach, and concluding that an effluent that has an effect below that of b*µc is not toxic under the TST approach. Therefore, the difference in performance between the two approaches, and the appropriateness of the chosen factor b, could differ when the normality assumption is not met.

In Appendix A, it is stated that "Any test that used more replicates than the minimum number of replicates as indicated in Table 1 will be flagged as such and subjected to separate statistical analyses initially because number of replicates is expected to influence WET test performance characteristics." This is true, and a very important point. The report itself does not mention that this ever occurred. If none of the data ever exceeded the minimum amount of replication required by the method, this is another factor that should be included in the simulations. A greater number of replicate analyses would increase the statistical power of the NOEC and TST tests, and would influence the comparison of the two approaches, and could influence the appropriateness of the chosen b factor.

One difference between the simulated and real-world data is that one "knows" the correct answer when simulating results because one starts with the population mean effect, whereas with the real world data one starts with the observed mean effect and the population mean is unknown. Unfortunately, this is an issue always faced when using existing data. As a result, the percentage of false toxic and false non-toxic decisions would not necessarily be accurate when determined from the real-world data. The variability of the observed mean effects will be a function of the number and variability of replicate results. This will have a greater effect on the choice of the most appropriate b value than on the TST vs. NOEC comparison because both approaches use the same data. However, it should be stated in the report that the error rates, sensitivity, and specificity for the TST approach may be inaccurate because it does not take into account the variability of the observed effect means.

4) Document Conclusions

Given the change of the null hypothesis from "sample is not toxic" to "sample is toxic" under the TST process, challenges from the regulated community are likely. That particular change is mainly a policy issue rather than a technical one. Emphasizing the decreased rate of falsely categorizing samples as toxic under the TST approach when variability is small will help with making this approach more palatable to this community. However, the ability of permittees to control the level of variability in test data may lower than anticipated, as they can only indirectly control the amount of analytical variability produced by the laboratory.

Because of the change in the null hypothesis under the TST process, the choice of "b" would be very important to the NPDES regulated community. As stated above, the assessments of "b" values using real-world data could be inaccurate because it is not known whether a sample is truly toxic or not. This would not be an issue with the Monte Carlo results, because the simulated results start with an assumed population mean effect. However, the simulated results would be more robust if additional factors, such as different statistical distributions, were included. If nonparametric tests are used instead of t-tests in the TST assessments, the optimal "b" value may change.

5) Overall Document Quality

There will likely be some implementation difficulties with the change from the current approach to the TST approach, because it will be perceived as going from "innocent until proven guilty" to "guilty until proven innocent" by NPDES permittees. Therefore, it is necessary that the improvements of the TST performance, and the added protection of the bioequivalency factor, be clear to the readers. Additionally, while some might see the benefit of increasing the amount of replication, taking account analytical precision when choosing laboratories or undertaking other approaches to reduce variability, others may not. Unfortunately, the data and analyses are not always presented clearly. This is especially important for tables and graphs, as they may be presented without the context given in the accompanying text.

The main difference between the TST and NOEC approaches, i.e., switching the null hypothesis from "assume non-toxic" to "assume toxic" complicates the discussion of the results. False positive decisions, false positive decisions, and statistical power each have very different meanings for the two different approaches. Therefore, it is recommended that these phrases be avoided when describing the results, even when the discussion is focusing on the TST approach only. For example, the definition of power given on page xii is correct for the NOEC approach, but not the TST approach (and is inconsistent with Table 1-2B). It is also recommended that the mean effect level cutoff for what is assumed to be "toxic" be emphasized when discussing the results. It is often unclear in the tables and results summary text whether a stated result is "correct" or not, as the mean effect cutoff varies from WET method to WET method.

6) Recommendations

As stated above, most of the weight of evidence supporting the TST approach is from the Monte Carlo analyses, as the real world data were only observational. Simulating results based on various non-normal distributions will yield additional information on the difference in the TST and NOEC approaches, and on the most appropriate value of "b" for a given method. Simulating different numbers of replicates also will be helpful, especially because it would be in the permittees' best interests to have the lab run more replicate analyses under the TST approach to reduce variability.

Other comments on tables/graphs are listed below:

Figure E-1, page iv: The labels for the symbols may be confusing to the audience. The phrase "NOEC passes," for example, implies that the NOEC test gave the correct answer, but really seems to be stating that the sample was categorized as non-toxic based on the NOEC test.

Figure E-2, page vi: The previous comment applies to this figure as well. Additionally, unlike the prior graph, it is unclear how the summarized results relate to the target percent effect. Perhaps this graph could be presented different effect level ranges, or effect level ranges could be included in footnotes under the graph.

Table E-2: Because the meaning of α and β differ between the two test approaches, it would be helpful to define them in a footnote for this table.

Figure 1-1, page 5: This is a useful graph for portraying the difference between the approaches.

Figure 2-2, page 22: This figure (and all subsequent figures in this style) is a bit confusing. Depicting the TST failure rate and passing rate as two lines is redundant, because one rate is always 1 minus the other. Perhaps this information could be depicted as a set of stacked bar charts. Additionally, the legend states that the Y-axis is in percent units, while the axis label is in proportion units.

Table 4-2, page 29: This table is unclear. It appears that the values under α and β correspond to the proportion of simulated analyses that were categorized as toxic (β) and non-toxic (α). When a number is presented, it implies that that result would be the "wrong" one given the simulated effect level, while a '-' implies that that conclusion was the "right" one, and therefore would not be an error. However, the 25% effect level includes proportions for both alpha and beta. From the footnotes, this appears to be due to the data being produced by different assumed CVs; however, this still implies that both results are "wrong." Much of this is likely due to the use of α and β, as they imply that the presented proportions represent "errors." It would be more meaningful to the audience to label the columns "test concludes toxic" and "test concludes non-toxic" without α and β, and explain that these two values will not add up to 1 in all cases.

Figure 4-1, page 30: This figure is useful, but it may be helpful to add dashed lines at 0.95 and 0.2 to emphasize the target error rates. Also note that this figure does not include interpolation lines between points, while all subsequent graphs in this format do.

Figure 5-2, page 45: This figure isn't as useful without knowing the observed mean effects for the two sets of data. While it can show NPDES permittees the benefits of reducing variability, for other segments of the audience what really matters is whether the approach gave the "correct" answer or not.

Table 6-1, page 50: These tables are the most useful for the public, as it gives the frequency of "wrong" answers under both approaches, though it would be more accurate to show the rates when the variability is larger as well. There may be some value in re-arranging the columns so the two fractions of samples incorrectly categorized as toxic are paired together, rather than the two rates for a given approach. However, this is not a vital change.

Table 6-2, page 51: It may be useful to include the sensitivity and specificity based on the NOEC approach for comparison purposes.

Table B-1: Rather stating that the percents were calculated as the percent toxic using a separate result column, the table would be more readily understandable to state this in the column headings. The heading for this table also incorrectly states that the next page shows the percent of samples categorized as non-toxic.

Table E-1: Why does this table show the percentage of tests categorized as non-toxic, while all previous ones show the percentage of tests categorized as toxic? This will confuse the audience.

Miscellaneous additional items:

Table E-4, page ix: Footnotes 3 and 4 quote mean effect cutoffs of 20% and 25%, respectively. These should probably be the same.

Glossary, page xi: The definition of confidence interval is rather vague, though it may suffice for this document (as confidence intervals generally don't play a role in either the NOEC or TST approaches).

Section 2.5.1, page 16, 3rd sentence: "p" is not defined

Section 5.0, page 42, last sentence of 1st paragraph: The statement “the t-test, a type of hypothesis test, is not usually designed to minimize the rate of false negatives in the WET program” is a bit inaccurate. It should really say that studies for which t-tests are applied are not usually designed to minimize the rate of false negatives.

Appendix B, “Detailed Analysis” figures: The label says “TET” rather than “TST”

Appendix Graphs: Many of the graphs appear to have been cutoff when inserted. For example, those on page 99 of the PDF.

Commenter 5

1) Document Merit

Conceptual soundness: Bioequivalence testing is an established method with justification beyond this study. This document requires much additional work before it would be ready for a broader distribution. In fact, review by an appropriate subcommittee of EPA’s Scientific Advisory Board might provide even a broader perspective on the impact and accessibility of this report.

Data analysis: An empirical study is reasonable although without knowing the truth toxic/non-toxic state, you are simply comparing two methods to each other and not to a “gold” standard. The simulation strategy is reasonable although poorly described.

2) Document Responsiveness

The HT strategy requires good experimental design that reflects the size of the effect that is important in light of test variability with a balancing of decision errors. The so-called TST approach is really not different in this regard.

3) Document Data Analysis Basis

As noted above, the empirical comparison of HT and TST approaches with an extensive data base is reasonable; however, this can not provide a definitive basis for selecting one method over another. We don’t know which approach yields a “true” classification of toxicity; we only know which is detecting a specific decrement relative to a control response.

4) Document Conclusions

I believe that the “b” values should first and foremost be set by scientists who can comment on what impact represents a meaningful change to some exposed population of organisms. I don’t find this empirical exercise looking at a large number of experiments to be compelling. It is useful and interesting but I believe that “b” should be more linked to the biologically meaningful changes.

5) Overall Document Quality

I fear that this report will not be easily understood by the regulatory and regulated communities. There are a host of poorly explained, incorrectly defined concepts that are central to understanding the proposals in this report. In the SPECIFIC COMMENTS section below, I list a number of places throughout this report where the presentation is inadequate and confusing.

6) Recommendations

Assuming that samples that are truly toxic or non-toxic are identified, the choice of the “b” factor is essentially a balance between false positive and false negative error rates. This “b” factor could be conceptualized as determining the cutpoint for declaring a test result as “toxic” or “non-toxic.” These types of assessments are natural in the context of receiver operating characteric (ROC) curves, and this type of analysis should have been included as part of this assessment. One concern is that it is not clear that the toxicity/non-toxicity of samples are “known” in this empirical

exploration. Finally, other, specific comments below provide suggestions for improving the technical quality and clarity of this report.

Other Comments:

[page] comment

[title] The title of this report is somewhat misleading. The “test of significant toxicity” (TST) is compared to the “hypothesis testing’ (HT) approach but not really compared to the “point estimate” approach. In addition, this newer approach is more of a test of equivalent toxicity (TET) versus a test of significant toxicity (TST) if conceptualized from a bioequivalence perspective. The term “significant” can refer to either statistically detectable differences or to biologically meaningful changes.

[ii] TST is simply known as “bioequivalence” in the literature. It is confusing and misleading to introduce new terms when this is already well described in the literature. Thus, the TST references should be changed throughout the report to bioequivalence.

[ii] HT is not equivalent to NOEC. Hypothesis testing is a general strategy for evaluating competing hypotheses and TST clearly employs HT as well.

[ii] The “advantage” of hypothesis testing may be a reflection of a common misinterpretation of NOECs and no-effect levels. The concentration associated with a response that is not statistically different from controls is not necessarily a safe concentration. The testing reflects the variability in the system, size of the effect that would be declared different, the power of the test, and the false positive/Type I error rate.

[ii] The so-called point estimate method (again an unfortunate label since confidence intervals are often constructed) yields what is more commonly considered a potency endpoint in other toxicology applications. Potency is estimated at a particular risk management (RM) level (e.g. IC25 or IC50) and there is an incentive to generate higher quality data since the CI for this endpoint would be narrower and the standard error for the endpoint would be decreased as well. The focus on the two-concentration test data may suggest the reason why this analysis alternative was ignored.

[ii] The objective of finding the “b” for the TST to compare this to the HT approach implies that the “point estimation” approaches are not even in the mix any longer as is explicitly stated in the charge for this review.

[ii] The Table E-2 summary is confusing. A figure or table might help with this description. For example, would something like the following help?

$0.75\times\mu_0$ [25% mean effect]	$0.8\times\mu_0$ [20% mean effect]	$0.9\times\mu_0$ [10% mean effect]	μ_0
$\mu_E<0.75\times\mu_0$	$\mu_E>0.8\times\mu_0$	$\mu_E>0.9\times\mu_0$	Mean response in reference / control group
Correctly Declare TOXIC 100%	Incorrectly Declare TOXIC < 5%	Incorrectly Declare Non-TOXIC < 20%	

There is a potential confusion between a true (yet unknowable) difference in population mean responses versus an observed difference in sample mean responses. This distinction may be lost on the reader. This is an important point since this exercise is based on observed differences in sample mean responses relative to observed variability.

[ii] "β error rate" – this does not make sense. The errors that can be made in a decision framework are to reject the null hypothesis when the null is true (a Type I error or a False Positive error) or to accept null hypothesis when the alternative is true (a Type II error or False Negative error). The standard notation for the probability of these errors is α = Pr(Type I error) and β = Pr(Type II error). While most readers would be able to figure out what is meant, this type of statement is misleading and demonstrates insufficient care in presenting technical information.

[ii] It may be easier and make more sense to talk about this in terms of increased power vs. decreased Type II error rates.

[iii] Reference to the simulation method here is a surprise. Early on in the summary, the focus was on empirical comparisons while here we see a simulation component mentioned.

[iii] statements such as "TST never declared a 10% mean effect …" should be restated as "TST never declared an OBSERVED 10% difference in sample means between effluent and control conditions …"

[iii] Care must be taken when considering all of the percentage quantities described here. There are 1) observed sensitivity / power (%); 2) observed false positive/type I error rates (%); 3) the observed decrement in sample mean responses (% mean effect); 4) the value of "b" which is expressed as a % of the mean response; and related, the percentile of the CV distribution. This should be presented in a table or text box to make sure this doesn't lead to additional confusion.

[iii] Sensitivity and specificity are described without formally defining them. Unless the readers are familiar with health screening studies, these terms may be unfamiliar.

[iii] The declaration of when a test was non-toxic appeared to be based on a subjective decrement from control/reference group responses.

[iv] If you are going to compare two methods for detecting toxicity in terms of error rates in the decision, then **receiver-operating-characteristic (ROC) curves** are the most common and natural way to display comparisons. In fact, the area under an ROC curve is a measure of the quality of screening procedure.

[iv] on the figure … "Percent effect in effluent" = $\mu_E / \mu_0 \times 100\%$ or = $(\mu_0 - \mu_E)/\mu_0 \times 100\%$ (assuming a continuous response characterized by a population mean, E=effluent 0=control and a decrease in response is adverse).

[iv] on the figure … "CV Percentile" – refer to CV in control condition or the effluent condition? Are you assuming that these are the same?

[iv] now describing α and β error rates vs. Type I and Type II error rates.

[v] "rank a sample as toxic" – No, you are not ranking anything here. You are making a decision to declare a sample as toxic or not.

[vi] These graphical displays are inappropriate and poor displays of the comparison of the two methods. Three-dimensional pie graphs are often criticized in the statistical graphics community (chart junk – more picture than data presentation – displaying a non-existent third dimension – etc.). More importantly, a table would be a much cleaner display here. For example, the first chart could be replaced by

		t-test (HT procedure)	
		Pass	Fail
TST	Pass	73.6%	3.2%
	Fail	4.9%	18.3%

This table provides a much better sense of concordance between the test results. In fact, we can easily see that the tests have a concordance of almost 92% here (concordance is a formal characteristic that is commonly defined in categorical data).

[vii] Table E-2. Mixing Type I and II error descriptors in the same column is confusing at best. These are observed rejection rates based upon various choices of "b." Extensive clarification is needed here.

[viii] Where is the Monte Carlo simulation analysis described? (Answer: Section 2.5.2) This was alluded to in the summary but not presented in this table.

[ix] Table E-4. What is relative specificity? Relative sensitivity? It is defined as "The specificity of a medical screening test as determined by comparison with an established test of the same type" by the American Heritage Medical Dictionary. Is that what you mean?

[x] Add "HT" to list of acronyms? Add "aka bioequivalence" to TST?

[xi] Glossary. A number of the definitions included in the glossary are imprecise and somewhat misleading and occasionally incorrect.

Confidence interval = interval estimate of a population parameter (not "around a point estimate of a population"). CIs can be one-sided or two-sided but this is not a critical point here.

EC = parameter that corresponds to the concentration of a toxicant associated with a specified level of impact. If a statistical model is fit, then the EC is derived from an inversion of the statistical model, i.e. a function of the regression coefficients. An estimate of this EC can be obtained after fitting the regression model.

HT = refers to using statistical hypothesis testing to identify, which if any, concentration condition differs from control conditions.

Alternative definition? HT "hypothesis testing method" – Using NOEC derived from t-tests (2 groups) or anova with multiple comparisons (if >2 groups) to evaluate mean differences under discharge and control conditions.

MSD = magnitude of difference OF WHAT? In the responses in an effluent group relative to responses in a control group?

NSEC/LSEC = I am not convinced that it is helpful to add more acronyms to the collection already in use in this arena.

RP = ? ecologically determined?

Significant difference = means of two distributions of sampling results? This is unclear. It appears to be defining the CI of the difference between two population means to be the Sig. Diff. Shouldn't significance be a function of an ecologically relevant change?

Type I Error (alpha)

Type II Error (beta) = alpha/beta are the PROBABILITIES of these errors, not the errors themselves.

All of the Glossary presentation appears to emphasize measured responses (continuous variates) versus proportions.

[1] NOEC endpoint IS DEFINED BY A STATISTICAL hypothesis test that … - The endpoint is not the HT approach.

[2] I don't agree with much of what is contained in this summary. Since the "point estimate" approach is not within the scope of the comparison, it is somewhat odd to see this included in the

table. The point estimate approach alluded to here appears to be the ICp method and many of the criticisms relate to this method. The choice of effect level is no different than the choice of "b" in the TST as it is associated with some risk management level. The endpoint can be concentration dependent is listed as a disadvantage. I don't understand this criticism. Do you mean the spacing of concentrations may influence this? There have been 12+ years of scientific contributions to methods for aquatic toxicity testing that have appeared after this cited Pellston workshop, and these appear to be completely ignored in this report. One huge disadvantage of the HT approach is that people often misinterpret a NOEC as a threshold of no concern instead of an artifact of detectable effect sizes.

[3] MSD relates to Fisher's LSD or Tukey's HSD from multiple comparisons methods. It is the mean difference required to declare two population means different. This already includes the standard error of the difference in sample means. To further divide this by control mean is an attempt to give this a CV kind of interpretation.

[4] Note that well designed experiments balance the Type I/II error rates. Again, these are NOT alpha errors and beta errors.

[5] The figures are a nice way to communicate that the same mean difference may not be declared different if the data are more variable. Although in both plots it is important to note that the effluent is DECLARED toxic or non-toxic. You don't know truth. You only know the outcome of this decision.

[6] The null hypothesis is a statement about parameters of the population being equal and NOT an assertion that they are "not statistically significant." This is a fundamental concept and this type of mistake is fatal for this report.

[6] What does "Treatment > Control" here mean? Are you only interested in one-sided alternatives?

[6] Table 1-2A. The footnote in this table is one of the few places where the error rates will carefully and correctly defined.

[6] Table 1-2B. "Effluent ≤ b*Control" is not precise or clear. Do you mean "$\mu_E \leq b^* \mu_0$"?

[7] Sensitivity=statistical power? This first sentence is confusing.

[7] I'm not sure how this picture clarifies the story about HT vs. TST. Note that the HT approach is sometimes referred to as a t-test approach and here as the NOEC approach. This type of switching of description will confuse the general readership of such a report.

[8] As I commented earlier, the "b" factor should reflect important biological / ecological shifts in the population response.

[8] What does "degree of protectiveness" in objective 2 mean?

[8] If you can't compare TST to the "point estimate" approach, then how was it possible to have permits written using either HT or point estimate approaches?

[10] CV on the y-axis is for controls? Effluent group? Both?

[10] 90%-tile of CV from 1989 and 2000 – CVs from control group?

[11] So what are you doing for the survival endpoints? What are you doing with counts? Are you using transformations (e.g. arc-sine-sqrt for proportions, sqrt for counts)?

[12] Does this table imply that a control condition was run with each of these tests (e.g. effluent, reference toxicity)?

[13] What does MSD mean for responses that are proportions (e.g. survival, germination)?

[15] Defining the mean response of the effluent here as μ_T should be done much earlier in this presentation. This would allow the bioequivalence/TST and HT approaches to be formally stated in terms of parameters.

[15] The level of change described as decision 3 is equivalent to the choice of "p" in ICp.

[16] The description of the Monte Carlo simulation is inadequate and confusing. Monte Carlo was not used to simulate WET data. You used a Monte Carlo simulation to study the TST and HT approaches by first generating WET data with known underlying characteristics and then applying the approaches.

[17] second analysis EMPIRICALLY DERIVED Type I and Type II error rates … with different "b" values defined for each calculation of these error rates.

[17] Shouldn't you also simulate cases when the effluent mean was equal to the control mean?

[19] Type I error rate was equal to 0? [Here, incorrectly stated as α error=0.] This is an observed Type I error rate.

[20] It is bad statistical practice to talk about preceding a test of means with a test of variances. The F-test is notoriously sensitive to violations of the normality assumptions while the test of means are very robust. You can use an unequal variance t-test routinely as an alternative. Finally, other tests of variances such as Levene's test are preferred to the F test for variance homogeneity or Bartlett's >2 group generalization (assuming you want to formally test this which I believe is debatable).

[20] 2 replicates vs. 4 replicates? What is a replicate here? Any reported simulation should be presented in sufficient detail so that someone could repeat your computer experiment. This presentation does not meet such a standard. Not only are the conditions unclearly presented but the implementation of the simulation is sketchy at best. For example, how was the simulation programmed (in Excel? FORTRAN? SAS? R?)?

[21] mean effect levels versus an effect level defined as a change in mean response?

[21] mean percent effect ranges? What are these? Why these ranges?

[21] Not as robust statistically? What does this mean? You don't know "truth" in this empirical exercise. This comparison simply tells you how often the 2 methods lead to similar/dissimilar decisions.

[23] No, the test statistic is NOT formed from the population means μ_c and μ_e (what happened to the μ_T formulation earlier?) but in terms of sample means such as $\overline{Y}_C$

In this figure, doesn't nontoxic means $\mu_e > b\ \mu_c$

[23] The formula for calculating the pooled variance makes sense only if there are the same number of observations in both effluent and control groups.

[24] No, this is not the SE(mean) but the SE(difference in sample means)

[24] Doesn't the TST approach calculates the SE($\overline{Y}_e - b * \overline{Y}_C$)?

[25] No, the t-test statistics do NOT involve population means; they are functions of sample means. This is fundamental and critical notation.

[25] Doesn't b=0.68 imply toxic if $\mu_e \leq 0.68\ \mu_c$? Here, and elsewhere in the report, a decrease in response is considered adverse. Was this ever explicitly stated in the report?

[25] Isn't it better to say "equivalent to control response" instead of "not toxic?"

[27] How was the MSD determined? How did you determine the power > 80%?

[27] A 1993 paper reported a similar result? Isn't this backwards and the 1996 paper reported a similar result to the 1993 paper?

[29] Need to comment/formally define the relationship between sensitivity and Type II error rates? Between specificity and Type I error rates?

[31] The table legend needs to be enhanced here. For example, how is effect level defined here? What do the Risk Management columns mean here? Isn't "b" a RM decision?

[32] How does "mean difference" in this figure relate to "effect size?"

[*] Many of the later pages of this report contain figures and discussion that were already criticized as part of the review of the summary. These observations will not be repeated here.

[45] A standard boxplot is a better display here (e.g. box with lines at Q1, median, Q3 and whiskers extending from min to Q1 and from Q3 to max).

[47] TST was a viable alternative prior to this empirical investigation. Bioequivalence has a long and well studied history with pharmaceutical applications.

3. COMMENTS ORGANIZED BY REVIEW QUESTION

EPA Question 1 - Document Merit

Evaluate the conceptual soundness of the draft TST document's recommendations and the data analysis on which it is based. Is the draft TST approach an improvement over the current accepted hypothesis testing approach used in the NPDES WET program? If so, why, and if not, why not?

Summary

All of the commenters concurred that the bioequivalence method used in this study is a sound conceptual approach. Most also agreed that the TST approach is an improvement over the current accepted hypothesis testing approach used in the NPDES WET program. Commenters raised the issue of the method for selecting the value for *b.* One noted the traditional approach is to derive a value based on subject matter context (e.g., comparison to a known "true" value) whereas the EPA approach is developed in a reverse fashion: the value of b is determined by the precision of test measurements, not some objective standard of safety. Another noted any improvement over the current approach would depend on the choice of *b*, and that the chosen values in the document were reasonably justified. A third made a comment on data analysis that is relevant here: that the data analyses could not be evaluated absent a "gold standard," but was simply a comparison of methods. Commenters offered opinions on data analysis. One commenter thought real world and simulation data were good, but a limitation of real world data is that estimated error rates are based on sample data means and not population means. A second commenter indicated that without an objective standard of comparison, although a reasonable exercise, empirical studies can only provide a comparison of the methods.

Commenter 1

The draft TST document's recommendations and the data analyses upon which it is based are conceptually sound. The specific recommendations from the document are: 1) to include calculation of TST for NPDES WET testing as well as for testing for ambient water quality monitoring; 2) provide incentives for permitees to provide high quality WET data to permitting authorities with basis in reasonable potential decisions as well as WET limits; 3) provide protection for receiving systems if WET test data have relatively large within-test variability or other inconsistencies while decreasing the probability of "false positives"; and 4) incorporate error rates into the decision process, thereby increasing confidence in test results. This is the essence of adaptive water resource management (the future for water resources in the U.S.). The specific data and analyses in the draft TST document included actual and representative data from over 2000 WET tests as well as appropriate simulated data to further examine the results or consequences of the analysis. The draft TST approach can represent an improvement over the current hypothesis testing approach if the WET testing program for NPDES purposes as well as the ambient toxicity testing continues as they have in the past. As noted above, the TST approach will encourage design of aqueous toxicity testing producing more precision or less variability in the data. Further, the proposed TST approach will permit identification of toxic samples in situations confounded by variable data and will minimize "false positives" (identification of nontoxic samples as toxic). The scientific community is 'evolving' to this approach for much of the experimental data that we collect relative to toxicity. Thus, the TST approach should serve to aid convergence of statistical significance with regulatory toxicological (ecological and biological) significance. Important for acceptance of this approach by the regulated community will be its implementation. I have included more specific recommendations in subsequent sections of this review.

Commenter 2

The concept of an equivalence test is admirable. Equivalence tests provide a proof of safety while the NOEC provides proof of ignorance. The proposal to use an equivalence test is a big improvement over the current hypothesis test.

The weak part of the document is the rationale for choosing *b*, the equivalence factor. In a traditional equivalence test, *b* is determined by subject matter considerations. For example, FDA guidance for generic drug evaluation considers generic and branded drugs to be equivalent when important pharmacological properties of the generic are significantly between 80% and 125% of those for the branded drug. That equivalence region (80%, 125%) is a biological conclusion. Applied to WET evaluation, the analogous procedure would be to determine a level of effect that is considered important, i.e. not safe. This determination will not be easy because it relies on a lot of ecology. It may be partly a guess. It requires discussion among stakeholders. But, with stakeholder participation, *b* summarizes what is considered not safe. Once *b* is determined, the approaches in this document can be used to evaluate the properties of the test (alpha and beta levels given certain amounts of variability).

The approach in the EPA document is backwards. That is, *b* is determined by considering test characteristics (the risk management criteria). The choice of *b* is driven by the variability (c.v.) of the tests. In other words, safety (or not safe) is determined by the precision of our measurements. If tests were more precise, *b* would be closer to 1.0. This creates a logical conundrum: what do you do if WET tests get more precise (as data briefly discussed in the document suggests has happened since the early 1990's). If a WET test becomes more precise in 2010, do you compute a new b? The logic used in this document says you would. Common sense says this is nonsense. If *b* increases from 0.7 to 0.8, then an effect of 25% reduction from control is safe now (when $b = 0.7$) but not safe in 2010 (when $b = 0.8$). It's the same effect, with the same ecology.

In summary, the document is an improvement over the current hypothesis testing approach. An equivalence test is much better than the NOEC. It would be a huge improvement if subject matter considerations were used to determine *b*.

Commenter 3

The proposed approach is an improvement over the current, accepted hypothesis testing approach used in the WET program. The TST approach or what I would call the bioequivalence approach is a well-studied approach that is commonly used in biomedical studies. The problem with the current WET approach is that the type II errors are not controlled by the procedure. Because of this, I would expect that there is a tendency to not reject an effluent as toxic. The TST approach basically switches the null and alternative hypothesis, putting more burden on industry to show the effluent is safe. The novel addition to the TST approach is that there is an attempt to account for both types of errors by recommending the value of the bioequivalence factor (*b* in the report).

I have two major problems with the TST approach relative to the current approach. First, the formulas for the TST approach are never defined in the document. It is not clear if the intention is always to use a modified two-sample t-test with the assumption of unequal variances or if other tests are used in the data analysis. Second, the current WET test approach provides a roadmap for decision making that is intended to help a user decide which test to apply in a given situation (i.e. equal or unequal variance, transformation, etc). Such a roadmap could easily be done in this document.

Commenter 4

I agree with the central idea around the TST approach: that changing the NOEC null hypothesis (the effluent is not toxic) to the TST null hypothesis (the effluent is toxic) will likely improve the detectability of toxic effects. I also agree with changing the hypotheses by including the *b* bioequivalence factor to avoid the possibility of categorizing highly precise results as toxic when the effect is small. Whether the TST approach is an improvement over the current approach will ultimately depend on the choice of *b*. As shown in the data and report, a too-low value of *b* would reduce the ability to detect toxic effluents below that of the current approach. The document justifies the chosen *b* factors for the evaluated methods reasonably effectively. As discussed in the later items, however, additional analyses may be necessary to assess the robustness of the estimated *b* values to departures from various data assumptions.

I think that including data analyses based on both simulated and real-world data is generally a good one. However, the real-world data have the limitation that the categorization of "toxic" vs. "non-toxic" can only be made based on the mean of sample data, and therefore the estimated error rates would reflect those means rather than the population mean being estimated. Therefore, the real-world data analyses do not add much support to the document recommendations and in some cases could undercut them. Data analyses using the Monte Carlo data were generally sound; however, additional conditions should be included in the simulations, as discussed in Section 3.

Commenter 5

Conceptual soundness: Bioequivalence testing is an established method with justification beyond this study. This document requires much additional work before it would be ready for a broader distribution. In fact, review by an appropriate subcommittee of EPA's Scientific Advisory Board might provide even a broader perspective on the impact and accessibility of this report.

Data analysis: An empirical study is reasonable although without knowing the truth toxic/non-toxic state, you are simply comparing two methods to each other and not to a "gold" standard. The simulation strategy is reasonable although poorly described.

EPA Question 2 - Document Responsiveness

Assess whether the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis.

Summary

Four of the five commenters agreed the draft TST document is responsive and meaningful in addressing the limitations of current hypothesis testing statistical WET analysis. A dissenting commenter believed that hypothesis testing and TST approaches are not all that different in so far as that both approaches are based on experimental designs reflecting the magnitude of the effect sought. One commenter noted the real world data and analysis did not clearly show the stated limitations of the NOEC approach and the TST approach does not address all of the limitations because both approaches used the same data.

Commenter 1

The draft TST document is responsive and meaningful in addressing some limitations of the current hypothesis testing statistical WET analysis. It is important to note that we can arrive at the right or "correct" answer through either approach. Both approaches need high quality data and the outcome of analysis of those data is dependent at least in part on the experimental design selected by the permitee or the permit writer. Both approaches have to recognize that high quality data are not necessarily more precise or less variable data. The TST approach offers some

advantages in this regard. The TST incorporates the advantages of hypothesis testing and makes the risk management decision or level clear for the permitee. Further, the advantage of more precise data is clearly evident in the TST approach. The decrease in the incidence of identification of non-problems (false positives) should appeal to the regulated community. And the TST overcomes some of the current concerns with traditional hypothesis testing (e.g., *No Observed Effect Concentration* or NOEC) or the point estimate (e.g., *Inhibition Concentration* such as the IC_{25}) approaches (Crane and Newman 2000). Additional method guidance is provided in (USEPA 1995, 2002a, 2000b, 2000c) to analyze WET data and to determine compliance with permit conditions or water quality standards and this guidance provides permitees and permit writers some options for more efficient and effective experimental design.

Commenter 2

Yes it is. No further comments.

Commenter 3

There has been a need to revise the WET approach for quite a while as the design component of the approach has never been a strong component. This has resulted in a limitation to the testing approach. The two main problems have been the need to control variability and the need to control error rates. The TST approach is one way to do this by essentially switching the null an alternative hypothesis. The null essentially becomes that the effluent is toxic and the role of the laboratory study is to prove otherwise.

Commenter 4

Table 1-1 presents many of the limitations of the current NOEC approach fairly clearly. However, the TST approach does not address all of these limitations because both approaches use the same data. Therefore, only the disadvantages 2 (Does not explicitly evaluate statistical power) and 3 (No incentive for permittee to increase test precision) are by the TST approach. The data analyses described throughout demonstrate that the TST approach addresses these two issues better than the current approach.

Using the Monte Carlo analyses, it is demonstrated that toxic effluents are detected more effectively under the TST approach, and that the likelihood of the TST approach incorrectly categorizing non-toxic effluents as toxic decreases as variability decreases. High false positive (concluding that a test concentration is toxic when it has a low mean effect) and false negative (concluding that a test concentration is non-toxic when it has a high mean effect) rates for the NOEC approach can be found in the appendices and the results descriptions. It may be beneficial, however, to summarize these more succinctly using tables/graphs (see item 5 for further discussion).

Unlike the simulated data, the real-world data presented and analyzed in the draft document do not clearly show the stated limitations of the NOEC approach. Tabular results are presented in the Appendices for simulated results only, and the figures presenting real-world data are more limited than those presenting simulated data. Fore example, Figure 5-1 shows the frequency of the approaches coming to different conclusions (approximately 8% of the time for *Ceriodaphnia dubia* and 12% of the time for *Pimephales promelas*), but does not show these rates vary by observed mean effect. Additionally, Table 6-2 gives the sensitivity and specificity for the TST approach only.

The discussion of the real-world results also does not strongly support the stated limitations of the NOEC approach. In the conclusions section, it is stated:

“7% of the 256 tests (18 tests) were declared non-toxic using t-test, despite mean effect levels as high as 33%. In addition, 5% (13 tests) were declared toxic using the t-test approach at effect

levels as low as 17%. By comparison, TST rarely declared samples as non-toxic at mean effect levels > 20% and rarely declared samples as toxic at mean effect levels < 20%."

This does not strongly support the stated limitations of the NOEC/t-test approach, as the audience may not see a large difference between 7% and 5% rates and "rarely," especially because it is unclear which of the two approaches is giving the correct answer for those data. While rates for specific mean effects were not presented for the data presented in Section 5, other methods yielded rates higher than those quoted above when the toxicity decision was based on the TST approach for the chosen b value. For example, 25% of tests were declared as toxic when the mean effect was ≤20% using *b*=0.8 for *Macrocystis pyrifera*, as shown in Figure 4-6.

Commenter 5

The HT strategy requires good experimental design that reflects the size of the effect that is important in light of test variability with a balancing of decision errors. The so-called TST approach is really not different in this regard.

EPA Question 3 - Document Data Analysis Basis

Assess whether the data supporting the recommendations and conclusions on the draft TST document are technically correct and defensible. The draft TST document attempts to evaluate existing data comprehensively, but: (1) for the purposes of standardizing comparisons, relies on data developed after 1995; (2) to be comprehensive, evaluates data developed using EPA WET test methods conducted under the current 2002 edition, as well as some earlier editions; and (3) to ensure that conclusions are based on appropriate data, censors some data points. The Agency's reasoning behind each of these aspects of the evaluation is explained in the draft document and related references (i.e., data test acceptance and quality assurance protocol).

Summary

All of the commenters generally agreed that the data supporting the recommendations and conclusions are reasonable and defensible. There was a consensus that it was better to focus on current methods and future data. However on commenter thought that a census approach would be preferred to a balanced approach. This commenter also had an issue with an apparent assumption of normal distribution, a concern also stated by another commenter. They both question if the assumption of a normal distribution can be made for all data used. Lastly, one commenter had a set of specific issues critical of the documentation in the section of the document on the simulation method.

Commenter 1

The data supporting the recommendations and conclusions of the draft TST document are technically correct and defensible. For example, data from more than 2000 WET tests were used to analyze this approach. Use of the post-1995 data is justifiable since those data have been developed utilizing more uniform protocols and experienced testing laboratories. The test methods are sufficiently uniform to include them in the data set analyzed. The document emphasizes the importance of experimental design (e.g. replication) in influencing the outcome of a test or analysis. The data selection and processing SOP applied rigorous, logical and defensible criteria for inclusion of data in the sets that were analyzed. Censoring of data that did not meet inclusion criteria as presented in the SOP is appropriate. Importantly, the data were from numerous dischargers and testing laboratories to ensure that they were representative. Key to the initial success of the TST approach is the use of sufficient representative data in its derivation.

Commenter 2

The data sets have a minor role in the evaluation. Primarily, the data provides estimates of the c.v. among samples exposed to control and effluent. These estimates are used to provide realistic distributions for the simulation studies. The rationale for subsetting and censoring are very reasonable.

Commenter 3

The simulation that is used to determine the value of the bioequivalence parameter relies on the assumption that the data that are collected are representative of WET tests that are carried out in the future. Therefore they try to obtain a more or less balanced sample of the current data rather than a census of current data. The census approach would give more weight to laboratories that carry out a large number of tests. I think it is quite reasonable to focus on current methods and data rather than past data.

The section on the simulation method leaves much to be desired. For the simulation to be valid there is an assumption that the screened data represents the population of WET testing. It is not at all clear what is being simulated. It is not clear why the control group is truncated. What is the justification for this? It is not clear from reading the document what the data are. For example, for *Ceriodaphnia* do you have counts or is it simply the mean and coefficient of variation? I do not understand why the coefficients of variation are randomly selected. For a t-test one would just have to sample means and standard deviations. Is “n” fixed? It is not mentioned on pages 17 or 18 yet is critical for power. I think the authors also assume that the data are normal. This is not justified.

After reading this several times I think this is what was done

Use the screened data from controlled samples to obtain an empirical distribution for the mean and coefficient of variation

Select 1000 samples of means and coefficients of variation from the trimmed empirical distribution

Compute the test statistic for the TST test for different means and coefficients of variation. For each test statistics decide to accept or reject the hypothesis

Compute the number of rejections in the 1000 tests.

Some things that are missing

What is the sample size? Is the sample size for control treated the same as for the effluent

Was the control sample used with all levels of effluent mean and coefficient of variation (as in a block design) or was a new sample selected for each effluent level.

Was the choice of 1000 justified? For power calculations, I am used to having 10,000 simulations rather than 1000 as this sample size will lead to smaller simulation error.

What are the formulas for the tests?

What are the assumptions?

It appears that the mean and coefficient of variation were sampled independently. I think this is not justified. The mean and standard deviation are attributes of the individual WET test. I think they should be sampled together rather than separately.

I also do not understand why one would sample mean and coefficient of variation rather than mean and standard deviation. For the test statistic, the standard deviation would have to be calculated. Are the mean and coefficient of variation independent?

I question whether a Monte Carlo simulation is necessary. I think a more accurate approach is a direct calculation. Given the means, standard deviations, and a value of *b* compute the power of the test assuming normality. Then weight by the probability of the mean and standard deviation occurring in the population. Repeat this for a set of means and standard deviations (over a grid, say) and then compute the weighted average power. This approach should be more accurate.

It is also common to use crossvalidation in these cases to evaluate error rates. Since the value of *b* is calculated using all the data shouldn't a test using the value of *b* be evaluated on a separate data set?

Appendix A seems to have been pulled from the grant proposal as it is written in future tense rather than present or past tense. In reading this section, there is an implication that the EPA flowcharts will be used to select a test. This may result for example, in a two-sample t-test assuming equal variance (which should have greater power that the test assuming unequal variance). It is not clear that the same approach is applied to the TST procedure. What is the justification for this approach?

One additional criticism is that if the TST is an alternative to the WET approach, the simulation needs to compare the "approaches" not the "tests". The simulation as described compares the tests. The WET approach allows for a variety of tests depending on the assumptions. The conclusions therefore apply to a test not the approach.

Commenter 4

In terms of the range of WET test methods and time period of data development, the choice of data appears to be fine. However, there are a few factors that either were not included in the data, or were not assessed in the analyses:

A t-test was used for all analyses to assess both the NOEC and TST hypotheses. This suggests that all of the results followed a normal distribution (after any data transformations). In Section 2, it is stated that Welch's t-test is used when the assumption of equal variances is not met, but there is no discussion of whether the issue of non-normality occurred with the data. The Monte Carlo results may have been simulated based on the assumption of a normal distribution, but it seems unlikely that all of the real-world data would have followed a normal distribution, especially when the CV was high. Even if this was the case, it would be worth simulating data following some non-normal distribution. The Wilcoxon Rank Sum test is often used in NOEC analyses when the normality assumption is not met, and could easily be used for TST analyses as well, by adjusting either the control or effluent results by *b* prior to ranking. The statistical power of this test would likely be less when a nonparametric test is necessary, which would decrease the probability of concluding a truly toxic effluent is toxic under the NOEC approach, and concluding that an effluent that has an effect below that of $b^*\mu_c$ is not toxic under the TST approach. Therefore, the difference in performance between the two approaches, and the appropriateness of the chosen factor *b*, could differ when the normality assumption is not met.

In Appendix A, it is stated that "Any test that used more replicates than the minimum number of replicates as indicated in Table 1 will be flagged as such and subjected to separate statistical analyses initially because number of replicates is expected to influence WET test performance characteristics." This is true, and a very important point. The report itself does not mention that this ever occurred. If none of the data ever exceeded the minimum amount of replication required by the method, this is another factor that should be included in the simulations. A greater number of replicate analyses would increase the statistical power of the NOEC and TST tests, and would influence the comparison of the two approaches, and could influence the appropriateness of the chosen *b* factor.

One difference between the simulated and real-world data is that one "knows" the correct answer when simulating results because one starts with the population mean effect, whereas with the real world data one starts with the observed mean effect and the population mean is unknown. Unfortunately, this is an issue always faced when using existing data. As a result, the percentage of false toxic and false non-toxic decisions would not necessarily be accurate when determined from the real-world data. The variability of the observed mean effects will be a function of the number and variability of replicate results. This will have a greater effect on the choice of the most appropriate *b* value than on the TST vs. NOEC comparison because both approaches use the same data. However, it should be stated in the report that the error rates, sensitivity, and specificity for the TST approach may be inaccurate because it does not take into account the variability of the observed effect means.

Commenter 5

As noted above, the empirical comparison of HT and TST approaches with an extensive data base is reasonable; however, this can not provide a definitive basis for selecting one method over another. We don't know which approach yields a "true" classification of toxicity; we only know which is detecting a specific decrement relative to a control response.

EPA Question 4 - Document Conclusions

Assess whether the draft TST approach as applied is technically defensible especially if challenged by either the NPDES regulated community, permitting authorities or expert consultants hired by permittees or other interested parties. Specifically, bioequivalency "*b*" values were derived for each test method using several risk management decision criteria which together, were intended to balance desired maximum alpha and beta errors at specific mean effect levels and within-test variability. Comment on the fact that this draft TST approach could be similarly used for additional WET test method(s) in the future. This draft TST approach builds upon EPA's earlier peer reviewed NPDES WET Variability document (USEPA 2000e) to derive and evaluate the "*b*" values. Evaluate the methodology used in the draft TST document to derive method-specific "*b*" values and apply the draft TST approach.

Summary

Commenters generally agreed that the TST approach is technically defensible. All commenters agreed, however, that the method in determining "b" values is critical to the validity of the TST approach and its acceptance by the regulated community. Two commenters noted that the derivation of the method specific "b" values was scientifically defensible or could be reasonable based on the error rates chosen. One commenter was concerned specifically about the way "b" was determined and also about the alpha values chosen. One commenter suggested emphasizing the reduction in false positives to help the acceptance of the approach by the regulated community. One commenter suggested including different statistical distributions to improve the robustness of the simulation results. One commenter believes that "b" should be chosen to reflect biologically meaningful changes. A commenter also suggested dropping the label "bioequivalency" as it conveys a potentially confusing meaning due to its historical use that is unnecessary to its meaning or use in the TST approach.

Commenter 1

The draft TST approach is technically defensible and will likely be challenged by the NPDES regulated community or other interested parties. The challenges will likely come from those impacted negatively as well as the entities allied with those impacted negatively by the draft TST approach. However the TST approach is technically feasible and defensible. I think the term "bioequivalency" should be abandoned since it will convey an unnecessarily confusing message

to the permitees (it has historical precedence that is of no value in this arena). The "*b*" values could be simply explained as an aggregate value accounting for the risk management decision (the level of protection desired for the receiving aquatic system), a balance of both false positive and false negative error rates at a specified effect level, and within test variability. For this TST document, derivation of the method specific "*b*" values was relatively clear and scientifically defensible. A decision support system could be readily developed that essentially codifies the process by which "*b*" values should be calculated (as outlined on pp. ii and iii and Table E-2). The TST approach has clear application to other toxicity testing in the future as data are accumulated and methods are refined.

Commenter 2

I repeat my concern with the method of choosing *b*, raised in point 1.

I also repeat my general view that an equivalence test is an appropriate way to demonstrate safety in many toxicity tests.

I am concerned with how alpha levels were computed. One detail in the tables in section 4 doesn't seem right. This is either a major failure in the computation or a major failure in communication. Details of my concern are given below in the detail for "Table 4.2 et seq".

Commenter 3

The bioequivalence approach is a valid statistical approach. It does not depend on this document for validation. The choice of *b* is critical in the method. The value in the document is based on the two error rates. If these error rates are reasonable then the approach has justification.

Commenter 4

Given the change of the null hypothesis from "sample is not toxic" to "sample is toxic" under the TST process, challenges from the regulated community are likely. That particular change is mainly a policy issue rather than a technical one. Emphasizing the decreased rate of falsely categorizing samples as toxic under the TST approach when variability is small will help with making this approach more palatable to this community. However, the ability of permittees to control the level of variability in test data may lower than anticipated, as they can only indirectly control the amount of analytical variability produced by the laboratory.

Because of the change in the null hypothesis under the TST process, the choice of "*b*" would be very important to the NPDES regulated community. As stated above, the assessments of "*b*" values using real-world data could be inaccurate because it is not known whether a sample is truly toxic or not. This would not be an issue with the Monte Carlo results, because the simulated results start with an assumed population mean effect. However, the simulated results would be more robust if additional factors, such as different statistical distributions, were included. If nonparametric tests are used instead of t-tests in the TST assessments, the optimal "*b*" value may change.

Commenter 5

I believe that the "b" values should first and foremost be set by scientists who can comment on what impact represents a meaningful change to some exposed population of organisms. I don't find this empirical exercise looking at a large number of experiments to be compelling. It is useful and interesting but I believe that "b" should be more linked to the biologically meaningful changes.

EPA Question 5 - Document Quality Overall

Provide any recommendations for how this draft TST document should be presented to the public (or the users of this approach) particularly NPDES regulatory authorities such as NPDES States and EPA Regions (the document will be revised to accommodate readers with a more *Plain English* version). Suggest, if possible, how it's highly technical content should be translated into a version more readily understood by the NPDES regulatory public (again meeting EPA's *Plain English* requirements) and yet maintain its clarity given its potential scientific, regulatory, and technical applications. Also critique whether a regulatory authority and their permittees would clearly understand the draft TST document's recommendations and if not how specifically should it be revised to make it easier to implement under EPA's NPDES permit's program.

Summary

Commenters were generally sensitive to how the document should be presented to the public. They also were quite critical of the presentation and clarity of the draft document. Suggestions as to distribution/implementation of the TST approach included: a consequence (cost) analysis, including which permittees may be adversely affected and whether monitoring data support potential receiving water impacts; and overcoming the perception that effluents will be presumed "guilty" rather than "innocent" with the TST approach. All of the commenters had substantial issues with the clarity, completeness, and grammatical errors that they found to be common throughout the document. One commenter noted the document needed to be re-written before a "plain English" assessment was attempted.

Commenter 1

A thorough consequences analysis should be conducted prior to implementation of this draft TST approach. What are the environmental, social and economic costs of implementation? A data base consisting of those permitees that would be negatively impacted could be developed containing data from those permits formerly identified as "nontoxic" that would be identified as "toxic" as a consequence of implementation of the draft TST approach. Are there any data or is there any evidence that the affected parties or sites actually have demonstrable problems?

As with any regulatory program or change in a program, the devil is in the details. How would this new approach be implemented? Would there be a "phase in" period or an option period analogous to "monitor and report"? This is a great opportunity to institute a compliance period (e.g. one year to comply). Both the current approach and the proposed TST approach could be used initially in a 'monitor-and-report' mode and then the new approach could be included in issuance of a new permit. Both permitees and permit writers will need workshops and training materials to implement the TST approach. What does a single "significant" test mean? And failure to reject the null hypothesis of no difference or bioequivalence does not mean that there is no effect (i.e. does not mean that you accept the null).

A user's manual with specific examples (beyond those on pp. 24 and 25) and perhaps software to facilitate implementation will be needed (this is equivalent to the TSD). The present document needs some editorial (English) assistance for clarity and grammar. Some of the definitions are weak or incorrect (e.g. "Reasonable Potential", "Precision", Type II Error [beta]) and some definitions are missing (e.g. "sensitivity", "specificity"). Other grammar or "english" errors include subject-verb agreements and incongruous phrases such as "both simultaneously highest" (see p. 33). Failing to reject a null hypothesis is not the same as accepting the alternative hypothesis. This appears to me to be an error throughout the document (e.g. p. 6, 1st line; "accepting the alternative hypothesis – that the effluent is non-toxic" may be the practical outcome of the statistical test, but I do not think that it is the strictly scientifically defensible interpretation).

Commenter 2

The document, as currently written, is very hard to understand. I struggled in many places, in spite of my background in these issues. Even after a month of struggling, I'm not sure I completely understand all that you did. If some of my comments don't make sense, it is probably because of my inability to understand what you did.

Will this document be the primary source documenting the TST? If so, it needs to include a worked example of a test, explaining and illustrating how the test is actually done. You do this on p 24, but that is somewhat hidden. The example needs to be prominent.

Parts of the document are repetitive (details below). It needs a thorough proofreading (some specifics included with my detailed comments). Some tables and figures need a major reorganization. Will this document be printed in color? If not, Figures E-1 and 6-1 become very hard to read. Using lines to indicate the rejection regions for the two tests avoids the need for color.

Commenter 3

I think the writing is rather weak and the document needs to be rewritten before the Plain English requirement is attempted. There are many minor problems with the descriptions, examples and general layout of the document. Many specific examples of problems are given below. To improve the clarity of the document, I think there is a need to carefully consider the layout of the document, the position of the figures and tables and the order of these figures and tables. There is a general lack of formulas and inconsistencies given when formulas are used. Although formulas might not be a good idea for some readers, they give a much clearer description of what was done than the written description. For example, to simply state that a two-sample test was done is not precise because variances could be equal or not. To be precise, I think the formulas should be general. The general description of the TST approach should include null and alternative hypotheses, assumptions, the test statistic and the criteria (distribution to be used along with degrees of freedom). One cannot infer these from the examples because the examples are not general. They are based on equal sample sizes and include nothing about how to calculate degrees of freedom.

Commenter 4

There will likely be some implementation difficulties with the change from the current approach to the TST approach, because it will be perceived as going from "innocent until proven guilty" to "guilty until proven innocent" by NPDES permittees. Therefore, it is necessary that the improvements of the TST performance, and the added protection of the bioequivalency factor, be clear to the readers. Additionally, while some might see the benefit of increasing the amount of replication, taking account analytical precision when choosing laboratories or undertaking other approaches to reduce variability, others may not. Unfortunately, the data and analyses are not always presented clearly. This is especially important for tables and graphs, as they may be presented without the context given in the accompanying text.

The main difference between the TST and NOEC approaches, i.e., switching the null hypothesis from "assume non-toxic" to "assume toxic" complicates the discussion of the results. False positive decisions, false positive decisions, and statistical power each have very different meanings for the two different approaches. Therefore, it is recommended that these phrases be avoided when describing the results, even when the discussion is focusing on the TST approach only. For example, the definition of power given on page xii is correct for the NOEC approach, but not the TST approach (and is inconsistent with Table 1-2B). It is also recommended that the mean effect level cutoff for what is assumed to be "toxic" be emphasized when discussing the

results. It is often unclear in the tables and results summary text whether a stated result is "correct" or not, as the mean effect cutoff varies from WET method to WET method.

Commenter 5

I fear that this report will not be easily understood by the regulatory and regulated communities. There are a host of poorly explained, incorrectly defined concepts that are central to understanding the proposals in this report. In the SPECIFIC COMMENTS section below, I list a number of places throughout this report where the presentation is inadequate and confusing.

EPA Question 6 - Recommendations

Provide any recommendations to improve the draft TST document's technical basis and approach for deriving the alternative WET statistical analysis method in the NPDES permitting program.

Summary

Commenters stated that many of their recommendations were presented in response to previous questions. Recommendations that were reiterated include the following:

- Eliminate the term "bioequivalence" because it will be met with resistance among NPDES permittees.
- Present the decision criteria for selection of "b" in an explicit tabular form
- Commit to monitor, analyze, and assess WET precision to refine alpha and beta error rates and "b" values
- Selection of the level of "b" should be by consensus
- Present the TST method so that it is mathematically clear, including assumptions, steps, statistics, and criteria for evaluation
- Base simulation results on various non-normal distributions not just normal ones
- Use weighted calculations to estimate "b" and power, not simulations
- Cutpoints for "toxic" and "nontoxic" types of assessments are natural in the context of receiver operating characteristic (ROC) curves, and this type of analysis should have been included as part of this assessment.

Commenter 1

Some suggestions and recommendations are presented in previous sections of this review. The technical basis for critical aspects of the draft TST analysis such as selection of the value for "*b*" in each case should be explicitly stated in tabular form as well as scientifically or technically defensible decision criteria for their selection. Although the decision criteria for selection of "*b*" values are in the report, they should be assembled in tabular form or essentially codified. Permitees would justifiably object if such values are left to arbitrary selection at the level of implementation (e.g. region or state). The consequences of selection of risk management values should be clearly explained and illustrated. This can be accomplished in a user's manual or through implementation guidance that includes appropriate t-test values (i.e. normal t-test, Welch's t-test) and examples of toxicity tests with multiple 'treatments'. The use of the "bioequivalence" terms or terminology will meet with resistance in the NPDES permitee arena and should be avoided. It will spawn opposition that can be avoided if more acceptable terms such as "test for significant toxicity" are used. To further avoid confusion, you will likely need another term for "*b*" values since beta (β) is already co-opted by statisticians.

A “confirmation” data set should be assembled if the draft TST approach is intended to be implemented. This data set would contain those data from permitees or site that would be impacted by implementation. For example, data should be scrutinized from a permitee that was formerly designated “nontoxic” under the hypothesis testing analysis and would be designated as “toxic” under the TST analysis. An explicit example or two would go a long way toward convincing the regulated community that this is worthwhile. What is broken that would now be fixed and what would be identified correctly as not broken and not in need of fixing? What will happen to the notion of accelerating testing with indication of a problem (i.e. a failed test)? There are considerable advantages in having additional data to support a decision in this case.

Some maintenance (monitoring) will be required for successful implementation of the TST approach. This will involve periodic evaluation of the precision of WET and other testing to establish and implement alpha and beta error “rates” and the “*b*” values. Commitment to this maintenance will need to be clearly stated (along with the frequency) in the implementation phase. The apparent interpretation that decreased variance in toxicity testing results indicates higher quality data may not be supportable. An alternative situation is that variance in the data indicates a varying process and variance in the sample. So variance may have to be partitioned (within test variance vs. among test variance vs. sample source or generator- specific [industry-specific] variance). Reference toxicant data should be inherently less variable than source sample data, although the opposite is indicated on p. 29 for *Pimephales promelas* chronic growth tests.

In summary, I think the TST approach represents an improvement relative to existing statistical techniques. However, the opportunity to realize the potential enhancement in the aqueous toxicity testing programs depends strongly on the implementation strategy employed.

Commenter 2

Overall comments reiterate what was said in point 1. An equivalence test, like the proposed Test of Significant Toxicity, should be the required statistical method for the analysis of WET data.

My only concern is the proposed method to choose the safety bound, *b*. It should be set by a consensus opinion of what biologically represents not safe. If a stakeholder based choice of *b* is impossible, the proposed method and choices of *b* lead to statistical tests with better properties than the current NOEC. This alone is a substantial improvement.

Commenter 3

Write out a specific description of the TST method that is mathematically clear. Include steps, assumptions, statistic and criteria for evaluation.

Write out the assumptions of the simulation and clearly state the steps used in the simulation.

Change the order of the presentation to ensure that figures, formulas, tables, etc are in the correct location for understanding the document.

Use weighted calculations rather than simulation to estimate b and power.

Commenter 4

As stated above, most of the weight of evidence supporting the TST approach is from the Monte Carlo analyses, as the real world data were only observational. Simulating results based on various non-normal distributions will yield additional information on the difference in the TST and NOEC approaches, and on the most appropriate value of “*b*” for a given method. Simulating different numbers of replicates also will be helpful, especially because it would be in the permittees’ best interests to have the lab run more replicate analyses under the TST approach to reduce variability.

Miscellaneous additional items:

Table E-4, page ix: Footnotes 3 and 4 quote mean effect cutoffs of 20% and 25%, respectively. These should probably be the same.

Glossary, page xi: The definition of confidence interval is rather vague, though it may suffice for this document (as confidence intervals generally don't play a role in either the NOEC or TST approaches).

Section 2.5.1, page 16, 3rd sentence: "p" is not defined

Section 5.0, page 42, last sentence of 1st paragraph: The statement "the t-test, a type of hypothesis test, is not usually designed to minimize the rate of false negatives in the WET program" is a bit inaccurate. It should really say that studies for which t-tests are applied are not usually designed to minimize the rate of false negatives.

Appendix B, "Detailed Analysis" figures: The label says "TET" rather than "TST"

Appendix Graphs: Many of the graphs appear to have been cutoff when inserted. For example, those on page 99 of the PDF.

Commenter 5

Assuming that samples that are truly toxic or non-toxic are identified, the choice of the "b" factor is essentially a balance between false positive and false negative error rates. This "b" factor could be conceptualized as determining the cutpoint for declaring a test result as "toxic" or "non-toxic." These types of assessments are natural in the context of receiver operating characteric (ROC) curves, and this type of analysis should have been included as part of this assessment. One concern is that it is not clear that the toxicity/non-toxicity of samples are "known" in this empirical exploration. Finally, the SPECIFIC COMMENTS below provide suggestions for improving the technical quality and clarity of this report.

Additional Comments

The following are additional and editorial comments submitted by the commenters.

Commenter 1

None.

Commenter 2

The document can be improved by a careful revision. My comments are marked by page and line, with negative numbers indicating lines from bottom:

p 7, l 3. How is statistical power incorporated into the TST? The statement is confusing because the TST also has a power. Which power is being referred to here? The TST and NOEC test have very different null hypotheses. The only connection (and hence the only way the power of the "usual" test of no difference is incorporated into the TST) is because all tests and all powers depend on the c.v. This claim is repeated many times in the document.

p 10, figure 2.1 Labeling of 'average' on the figures. If the dots are the 90'th percentile. I can't see how the lines are the "average CV", as claimed. Perhaps these are the average 90'th percentile, but that's not the the average c.v.

p 16, table 2-3. This is a repetition of table 1-2b.

p 16, l -4. "Monte Carlo" is an adjective not a noun, unless you're referring to the city state.

p 17, l 4. “ground truth” Is the truth known in the actual WET data? I doubt it. Without knowing the truth, how does the analysis of actual data sets “ground truth” the simulation?

p 18, l 6. The preceding paragraphs describe the choice of effect size and variability. Properties of a statistical test also depend on the d.f. for the error (i.e. whether variances are pooled, the number of treatments, and the number of replicates). These values affect the power and the relationship between the TST and the usual test. Table 2-1 gives the **minimum** numbers. This makes me suspect that there was variability among tests. How did you choose the rest of the details for each test. If your calculations used a single df error, say so, and report that df error.

p 16, lines 12 – end of page. This is really important stuff because it describes how you translated risk management criteria into an evaluation of both tests. It is a key feature of your analysis. It must not be buried inside a section described as ‘Freshwater and East Coast …’. I found this section very hard to understand. Since it lies at the heart of your evaluation, it needs to be prominent and clearly written.

p 19, l 4. This sounds like a fourth criterion. (or are you implying that a test should satisfy all criteria). If so, say that.

p 20, l 8. Where is exhibit A?

p 20, l 15. The arc-sine square root transformation is a **variance stabilizing** transformation. It does not correct for non-normality.

p 20, l -10. I don’t see why four replicates matter here.

p 20, l -8. You simulated data for four replicates, even though reality is only 2 replicates. That means you are reporting properties for a non-existent statistical test (using 4 reps).

p 20, l -3. This seems like you’re trying to force non-normal data into a t-test framework. This introduces all sorts of complications, as you discuss. However, I’m not sure that alternatives (e.g. binomial exact tests or beta-binomial tests) are any easier. They certainly are not part of the standard statistical toolbox.

p 21, l 8. I’m not sure what you’re doing here. In fact, I’m completely confused by what you say you’re trying to do. The problem with establishing alpha and beta from actual data is that both require that you know the true difference (is it zero or not). You don’t know that.

One ways to bypass this problem is to use only the control data and add a specific effect into the effluent mean. I couldn’t tell whether you used observed test data ‘as is’ (without knowing truth) or whether you added known values to control data.

p 21, l -6. This is a very unusual definition of alpha level. It isn’t clear whether ‘exceeding the toxicity threshold’ applies to population quantities or sample means. If this is based on sample means, the computation is completely wrong, since sample means often exceed (and even as likely as 50% to exceed) the population quantities.

Also, aren’t the definitions backwards, since %effect is calculated as (control – effluent)/control, so a large number is a ‘bad’, i.e. toxic result.

p 21, l -2. The number of data sets has nothing to do with their appropriateness for evaluating the simulations.

Figure 2.2 This one of many examples of poor graphics. Specific problems include:

a) the legend describes %, with a range of 0 – 100%, but the y axis is scaled from 0 – 1.

b) the y-axis label is the title of the plot, not a description of what is plotted on the y axis.

c) the two lines are redundant. One is 1- the other.

p 23-24. 3 issues:

a) Isn't this out of place and said earlier?

b) distributions of t. These aren't very useful, since the t statistics are very different for the two tests.

c) l -2 on p 24. Where did 25% effect come from? b is 0.68, which translates into a 32% difference from the control. Same issue on next page.

p 26, all. This material needs to be combined with that in section 2.4, which also discusses QA.

p 26. l 3. What is "Table 3", since tables are numbered as section-table?

p 26, l -8, What is computed in Excel?

p 29, table 4-2 (and all tables through 4-8 that are similar). These baffled me for far too long. The really important piece of information is the c.v. for each test, which is hidden in footnotes. Nothing in this table made sense until I realized you were changing c.v. for different rows.

I suggest a complete reorganization of this and comparable tables for other tests.

Effect level %	c.v.	Risk Criterion	B value				
			0.63	0.68	0.70	0.75	NOEC
15	≤ 75'th %	Toxic < 0.2	0.0	0.001	0.003	**0.35**	**0.99**
20	25-50'th %	Toxic < ??	0.00	0.02	0.21	1.00	1.00
25	> 50'th %	Non-toxic < 0.05	**0.99**	0.00	0.00	0.00	**0.78**
25	< 25'th %	Toxic = 0	**0.90**	**0.19**	**0.62**	**1.00**	**1.00**
30	Any	Non-toxic = 0	**0.99**	0.00	0.00	0.00	**0.22**

This suggested revision a) includes the c.v. as a specific element of the table, b) eliminates the meaningless columns (the –'s), and c) indicates where the risk management guidelines are exceeded (by the bold entries). Some of my entries for the 20% effect line are hypothetical, because I couldn't figure out the important information for that line in your table 4.2. If you feel that combining toxic and non-toxic endpoint s is too confusing, then separate the above table into an a part for the toxic endpoints and another part for the non-toxic endpoints.

p 30. Figure 4.1 Why didn't you use connected lines, as in figures 4.2 and 4.3?

Table 4.2 et seq. Is the value of 0.00 for the alpha level for 30% effect and b=0.70 correct? I think there is either a major failure in the computations or a major failure in the communication of how these results were computed. Here's why:

The usual interpretation of 30% effect is that control mean - effluent mean = 0.3 * control mean. When b = 0.70, then the effluent mean is exactly on the boundary of the equivalence region (because the boundary of the equivalence region is effluent mean = 0.7 control mean, i.e. control – effluent = 1 – 0.7 = 0.3 = 30% effect). When the population mean for the effluent lies exactly on the boundary of the equivalence region, the TST should have alpha = 5% no matter what the c.v. is. This is follows from the construction of the test and the definition of the alpha level. I very confused why the reported values are 0.00 (e.g. for 30% effect, b = 0.70, non-toxic (alpha) reported as 0.00) .

p 32, line -6. You refer to table 4-1. Shouldn't this be table 4-4?

p 44. The pie charts are terrible graphics. The false 3D only hinders the visual interpretation. A table presents the information much more concisely. You essentially include the table information when you give the actual %'s for each category. A graphical alternative is a mosaic plot.

p 44. This approach, a pairwise comparison of TST and NOEC test results for each actual data set, would be a very good way to summarize results for all the 'actual data' analyses in the previous sections.

p 47, lines 3 et seq. this seems to be based on the simulations. This claim would be much stronger if was based on the actual data. That is, you could use the actual WET data, compute both the NOEC and the TST and compare the results using the approach on p 44. More protective is then shown by t-test fails = 0 while TST fails = something larger than 0.

Figure 6.1 (and also E-1). Most of this graph is blank space because the largest % effect is 40%, but the Y axis maximum in 100%. Re-draw the graph with Y max = 40% to focus on the interesting stuff.

p 54, l -5. Again, I don't see how power is incorporated into the decision process because power of the t-test has little to do with the decision from the TST.

p 55. l -5. Need year for the Grothe et al citation.

Commenter 3

Line numbers would help with the review process

Figure E-1 What is the purpose of the colors? If these mean something why is it not described in the figure legend?

Page iv top low error rates than t-test should be lower error rates than the t-test.

Figure E-2 What is the standard t-test. The two sample t with equal variance? What is meant by TST fails? It is difficult to see that the TST test is better unless a higher fail rate is better. In fact, one might argue that the degree of concordance is quite high for the two approaches.

The results of the project listed on the bottom of page iv really don't have anything to do with the project; these statements basically can be made just from reading papers on bioequivalence.

Table E-4. What is the difference between specificity and relative specificity? Here is the medical definition of relative specificity: The specificity of a medical screening test as determined by comparison with an established test of the same type. I am not even sure that what you have calculated is correctly termed specificity since specificity implies knowing the true state.

Page xi Glossary: why not use standard definitions of terms rather than creating misleading ones. For example: hypothesis test, power and significant difference (why connect this to a confidence interval) are not consistent with statistical definitions of these terms. Type I and II error are expressed in terms of hypotheses but power is not.

Table 1-1

NOEC disadvantage 2: does not explicitly estimate statistical power – I don't think any method explicitly estimates power – I think you mean *controls* statistical power

NOEC disadvantage 5 is also a problem with estimation of an IC50. It is also a potential problem with the equivalence approach.

Point estimate disadvantage 4: confidence intervals are also affected by assumptions

Page 1. Line 7: the NOEC is not a hypothesis test rather it is level (concentration or dilution) that is derived from a test of hypothesis.

Line 16: the LC50 is an estimated value and it seems to me answers the question at what concentration is the 50% effect predicted. If the WET test uses a criterion for evaluation that is

based on a permitted IWC less than the IC25 it seems to be ignoring the uncertainty in the estimate of the IC25. Is this not a problem?

Page 4 line 9. Power depends on the variability of both groups (see figure 1-1). I think what you really want to describe here is the potential for a small biological effect being significant. If any decrease in survival relative to the control is an indication of toxic then is this not relevant?

Figure 1.1 Very small intra-test variability – This would be better if there is not a pattern in the second group of data i.e. two lines of data. I would change the figure to show no linear pattern. The use of "very small" to characterize variability seems odd – why not just use "low".

Table 1-2B: I think "b" should be defined in the table since you take the opportunity here to define Type I and II errors. Should b<1 also be added? I think the order of table and figures needs to be looked at as it seems that table 1-2B should come before fibure 1.2

Page 6: It seems that the main value of the bioequivalence test is that it puts the burden on industry to use a sufficient sample size that has good power for the test. If sample sizes are not sufficient the effluent will not be declared "not toxic". Why would the approach be better than to require a post-hoc power analysis i.e. require the industry to have power of 0.8 for a change of say 30%?

Page 7: I am not sure what is meant by "maximum" desired alpha and beta rates. It seems one would want to minimize these or to be as close as possible to specified rates.

Page 8: the project objectives are not well written. The first paragraph is an awkward read. First, the primary objective (purpose) is stated (which seems to actually be two objectives). Then the second sentence lists another set of primary objectives.

Figure 2-1. Is it appropriate to talk about the coefficient of variation as test variability?

In reading the document, I was very confused by the connection between the "test" used to statistically evaluate an hypothesis, the data and the simulated data. I did not find in the document a formula for the bioequivalence test statistic. Is the test being used just a variation of the two sample t-test? If so, why not write it out. I think it also has to be connected to Table 2-1 which describes typical data for the study design. For an uninitiated reader it would be useful to know what the minimum number of effluent concentrations corresponds to. For the bioequivalence test, is the reference compared to all of these or just the full concentration? Table 2-1 needs to connect to section 2.5 and the test statistics. In addition, if different methods are applied to the data from these studies in different situations, should you not also consider different tests for bioequivalence?

The study is based on the assumption that the data represents a sample from laboratories or facilities that is in some sense probabilistic. This is needed to treat the screened sample (20 most recently conducted tests) as a sample from a population. How can one tell if this is an unbiased collection of observations? It might be useful to know what the universe of samples is and how the observed facilities and laboratories represents this universe. What are your assumptions?

Page 15: There is a need to be explicit as to the formula for the test statistics, the assumptions, degrees of freedom, etc., for the bioequivalence test.

Although beta=0.8 is probably the first size of power that comes to mind for most statisticians, there is also an argument for an alpha:beta compromise.

Table 2-3: I like that the hypotheses are defined in terms of parameters. However the mean of the treatment is defined incorrectly. It is not the *response of* the effluent concentration rather it is the *mean response to* or the population mean response to… Why then switch to effluent<b*control which is not very descriptive?

Section 2.5.2 Is “several different” good grammar?

The notation used on page 23 bottom is confusing. S_c is referred to as the variance of the control yet in the formula S_c^2 is used. Why use $SDEV_c$ rather than simply S_c. I would use the standard notation S_c^2 for a sample variance and S_c for a standard deviation. Include degrees of freedom and critical value. I think it should be α=.05 not α<0.5. A clear way to write it is $t_{.05(1),18}$=1.73. It also seems that Step 1 should be to state the hypotheses (this way one knows it is a one-sided test).

I think the example using the unequal variance case (page 25) should provide a general formula not a formula for the special case when N_c=N_e. Otherwise this could be misleading for someone using the example as a template for their data.

Page 24: the standard statistical statement is “do not reject” the null hypothesis rather than “accept” the null hypothesis.

Page 27 why mention *Ceriodaphnia* is a freshwater invertebrate – water flea when this is mentioned in the header for 4.1. It is stated that 65% has power >= 0.8 and 35% is <=0.8. Should one of these be a strict inequality?

Appendix B. Are results based on 1000 simulations?

The axis legend is often cutoff i.e. **Minimum Significant Differen**. The truncation of the labels occurs throughout the graphs in the appendix. Also the numbers on the y-axes on many of the graphs are difficult to see as the overlap the axis line.

Commenter 4

Specific comments on tables/graphs are listed below:

Figure E-1, page iv: The labels for the symbols may be confusing to the audience. The phrase “NOEC passes,” for example, implies that the NOEC test gave the correct answer, but really seems to be stating that the sample was categorized as non-toxic based on the NOEC test.

Figure E-2, page vi: The previous comment applies to this figure as well. Additionally, unlike the prior graph, it is unclear how the summarized results relate to the target percent effect. Perhaps this graph could be presented different effect level ranges, or effect level ranges could be included in footnotes under the graph.

Table E-2: Because the meaning of α and β differ between the two test approaches, it would be helpful to define them in a footnote for this table.

Figure 1-1, page 5: This is a useful graph for portraying the difference between the approaches.

Figure 2-2, page 22: This figure (and all subsequent figures in this style) is a bit confusing. Depicting the TST failure rate and passing rate as two lines is redundant, because one rate is always 1 minus the other. Perhaps this information could be depicted as a set of stacked bar charts. Additionally, the legend states that the Y-axis is in percent units, while the axis label is in proportion units.

Table 4-2, page 29: This table is unclear. It appears that the values under α and β correspond to the proportion of simulated analyses that were categorized as toxic (β) and non-toxic (α). When a number is presented, it implies that that result would be the “wrong” one given the simulated effect level, while a ‘-‘ implies that that conclusion was the “right” one, and therefore would not be an error. However, the 25% effect level includes proportions for both alpha and beta. From the footnotes, this appears to be due to the data being produced by different assumed CVs; however, this still implies that both results are “wrong.” Much of this is likely due to the use of α and β, as they imply that the presented proportions represent “errors.” It would be more meaningful to the

audience to label the columns “test concludes toxic” and “test concludes non-toxic” without α and β, and explain that these two values will not add up to 1 in all cases.

Figure 4-1, page 30: This figure is useful, but it may be helpful to add dashed lines at 0.95 and 0.2 to emphasize the target error rates. Also note that this figure does not include interpolation lines between points, while all subsequent graphs in this format do.

Figure 5-2, page 45: This figure isn’t as useful without knowing the observed mean effects for the two sets of data. While it can show NPDES permittees the benefits of reducing variability, for other segments of the audience what really matters is whether the approach gave the “correct” answer or not.

Table 6-1, page 50: These tables are the most useful for the public, as it gives the frequency of “wrong” answers under both approaches, though it would be more accurate to show the rates when the variability is larger as well. There may be some value in re-arranging the columns so the two fractions of samples incorrectly categorized as toxic are paired together, rather than the two rates for a given approach. However, this is not a vital change.

Table 6-2, page 51: It may be useful to include the sensitivity and specificity based on the NOEC approach for comparison purposes.

Table B-1: Rather stating that the percents were calculated as the percent toxic using a separate result column, the table would be more readily understandable to state this in the column headings. The heading for this table also incorrectly states that the next page shows the percent of samples categorized as non-toxic.

Table E-1: Why does this table show the percentage of tests categorized as non-toxic, while all previous ones show the percentage of tests categorized as toxic? This will confuse the audience.

Commenter 5

SPECIFIC COMMENTS: [page]

[title] The title of this report is somewhat misleading. The “test of significant toxicity” (TST) is compared to the “hypothesis testing’ (HT) approach but not really compared to the “point estimate” approach. In addition, this newer approach is more of a test of equivalent toxicity (TET) versus a test of significant toxicity (TST) if conceptualized from a bioequivalence perspective. The term “significant” can refer to either statistically detectable differences or to biologically meaningful changes.

[ii] TST is simply known as “bioequivalence” in the literature. It is confusing and misleading to introduce new terms when this is already well described in the literature. Thus, the TST references should be changed throughout the report to bioequivalence.

[ii] HT is not equivalent to NOEC. Hypothesis testing is a general strategy for evaluating competing hypotheses and TST clearly employs HT as well.

[ii] The “advantage” of hypothesis testing may be a reflection of a common misinterpretation of NOECs and no-effect levels. The concentration associated with a response that is not statistically different from controls is not necessarily a safe concentration. The testing reflects the variability in the system, size of the effect that would be declared different, the power of the test, and the false positive/Type I error rate.

[ii] The so-called point estimate method (again an unfortunate label since confidence intervals are often constructed) yields what is more commonly considered a potency endpoint in other toxicology applications. Potency is estimated at a particular risk management (RM) level (e.g. IC25 or IC50) and there is an incentive to generate higher quality data since the CI for this

endpoint would be narrower and the standard error for the endpoint would be decreased as well. The focus on the two-concentration test data may suggest the reason why this analysis alternative was ignored.

[ii] The objective of finding the "b" for the TST to compare this to the HT approach implies that the "point estimation" approaches are not even in the mix any longer as is explicitly stated in the charge for this review.

[ii] The Table E-2 summary is confusing. A figure or table might help with this description. For example, would something like the following help?

$0.75\times\mu_0$ [25% mean effect]	$0.8\times\mu_0$ [20% mean effect]	$0.9\times\mu_0$ [10% mean effect]	μ_0
$\mu_E<0.75\times\mu_0$	$\mu_E>0.8\times\mu_0$	$\mu_E>0.9\times\mu_0$	Mean response in reference / control group
Correctly Declare TOXIC 100%	Incorrectly Declare TOXIC < 5%	Incorrectly Declare Non-TOXIC < 20%	

There is a potential confusion between a true (yet unknowable) difference in population mean responses versus an observed difference in sample mean responses. This distinction may be lost on the reader. This is an important point since this exercise is based on observed differences in sample mean responses relative to observed variability.

[ii] "β error rate" – this does not make sense. The errors that can be made in a decision framework are to reject the null hypothesis when the null is true (a Type I error or a False Positive error) or to accept null hypothesis when the alternative is true (a Type II error or False Negative error). The standard notation for the probability of these errors is α = Pr(Type I error) and β = Pr(Type II error). While most readers would be able to figure out what is meant, this type of statement is misleading and demonstrates insufficient care in presenting technical information.

[ii] It may be easier and make more sense to talk about this in terms of increased power vs. decreased Type II error rates.

[iii] Reference to the simulation method here is a surprise. Early on in the summary, the focus was on empirical comparisons while here we see a simulation component mentioned.

[iii] statements such as "TST never declared a 10% mean effect …" should be restated as "TST never declared an OBSERVED 10% difference in sample means between effluent and control conditions …"

[iii] Care must be taken when considering all of the percentage quantities described here. There are 1) observed sensitivity / power (%); 2) observed false positive/type I error rates (%); 3) the observed decrement in sample mean responses (% mean effect); 4) the value of "b" which is expressed as a % of the mean response; and related, the percentile of the CV distribution. This should be presented in a table or text box to make sure this doesn't lead to additional confusion.

[iii] Sensitivity and specificity are described without formally defining them. Unless the readers are familiar with health screening studies, these terms may be unfamiliar.

[iii] The declaration of when a test was non-toxic appeared to be based on a subjective decrement from control/reference group responses.

[iv] If you are going to compare two methods for detecting toxicity in terms of error rates in the decision, then **receiver-operating-characteristic (ROC) curves** are the most common and

natural way to display comparisons. In fact, the area under an ROC curve is a measure of the quality of screening procedure.

[iv] on the figure … "Percent effect in effluent" = $\mu_E/\ \mu_0 \times 100\%$ or = $(\ \mu_0 - \ \mu_E)/\ \mu_0 \times 100\%$ (assuming a continuous response characterized by a population mean, E=effluent 0=control and a decrease in response is adverse).

[iv] on the figure … "CV Percentile" – refer to CV in control condition or the effluent condition? Are you assuming that these are the same?

[iv] now describing α and β error rates vs. Type I and Type II error rates.

[v] "rank a sample as toxic" – No, you are not ranking anything here. You are making a decision to declare a sample as toxic or not.

[vi] These graphical displays are inappropriate and poor displays of the comparison of the two methods. Three-dimensional pie graphs are often criticized in the statistical graphics community (chart junk – more picture than data presentation – displaying a non-existent third dimension – etc.). More importantly, a table would be a much cleaner display here. For example, the first chart could be replaced by

		t-test (HT procedure)	
		Pass	Fail
TST	Pass	73.6%	3.2%
	Fail	4.9%	18.3%

This table provides a much better sense of concordance between the test results. In fact, we can easily see that the tests have a concordance of almost 92% here (concordance is a formal characteristic that is commonly defined in categorical data).

[vii] Table E-2. Mixing Type I and II error descriptors in the same column is confusing at best. These are observed rejection rates based upon various choices of "b." Extensive clarification is needed here.

[viii] Where is the Monte Carlo simulation analysis described? (Answer: Section 2.5.2) This was alluded to in the summary but not presented in this table.

[ix] Table E-4. What is relative specificity? Relative sensitivity? It is defined as "The specificity of a medical screening test as determined by comparison with an established test of the same type" by the American Heritage Medical Dictionary. Is that what you mean?

[x] Add "HT" to list of acronyms? Add "aka bioequivalence" to TST?

[xi] Glossary. A number of the definitions included in the glossary are imprecise and somewhat misleading and occasionally incorrect.

Confidence interval = interval estimate of a population parameter (not "around a point estimate of a population"). CIs can be one-sided or two-sided but this is not a critical point here.

EC = parameter that corresponds to the concentration of a toxicant associated with a specified level of impact. If a statistical model is fit, then the EC is derived from an inversion of the statistical model, i.e. a function of the regression coefficients. An estimate of this EC can be obtained after fitting the regression model.

HT = refers to using statistical hypothesis testing to identify, which if any, concentration condition differs from control conditions.

Alternative definition? HT “hypothesis testing method” – Using NOEC derived from t-tests (2 groups) or anova with multiple comparisons (if >2 groups) to evaluate mean differences under discharge and control conditions.

MSD = magnitude of difference OF WHAT? In the responses in an effluent group relative to responses in a control group?

NSEC/LSEC = I am not convinced that it is helpful to add more acronyms to the collection already in use in this arena.

RP = ? ecologically determined?

Significant difference = means of two distributions of sampling results? This is unclear. It appears to be defining the CI of the difference between 2 population means to be the Sig. Diff. Shouldn’t significance be a function of an ecologically relevant change?

Type I Error (alpha)

Type II Error (beta) = alpha/beta are the PROBABILITIES of these errors, not the errors themselves.

All of the Glossary presentation appears to emphasize measured responses (continuous variates) versus proportions.

[1] NOEC endpoint IS DEFINED BY A STATISTICAL hypothesis test that … - The endpoint is not the HT approach.

[2] I don’t agree with much of what is contained in this summary. Since the “point estimate” approach is not within the scope of the comparison, it is somewhat odd to see this included in the table. The point estimate approach alluded to here appears to be the ICp method and many of the criticisms relate to this method. The choice of effect level is no different than the choice of “b” in the TST as it is associated with some risk management level. The endpoint can be concentration dependent is listed as a disadvantage. I don’t understand this criticism. Do you mean the spacing of concentrations may influence this? There have been 12+ years of scientific contributions to methods for aquatic toxicity testing that have appeared after this cited Pellston workshop, and these appear to be completely ignored in this report. One huge disadvantage of the HT approach is that people often misinterpret a NOEC as a threshold of no concern instead of an artifact of detectable effect sizes.

[3] MSD relates to Fisher’s LSD or Tukey’s HSD from multiple comparisons methods. It is the mean difference required to declare two population means different. This already includes the standard error of the difference in sample means. To further divide this by control mean is an attempt to give this a CV kind of interpretation.

[4] Note that well designed experiments balance the Type I/II error rates. Again, these are NOT alpha errors and beta errors.

[5] The figures are a nice way to communicate that the same mean difference may not be declared different if the data are more variable. Although in both plots it is important to note that the effluent is DECLARED toxic or non-toxic. You don’t know truth. You only know the outcome of this decision.

[6] The null hypothesis is a statement about parameters of the population being equal and NOT an assertion that they are “not statistically significant.” This is a fundamental concept and this type of mistake is fatal for this report.

[6] What does “Treatment > Control” here mean? Are you only interested in one-sided alternatives?

[6] Table 1-2A. The footnote in this table is one of the few places where the error rates will carefully and correctly defined.

[6] Table 1-2B. "Effluent ≤ b*Control" is not precise or clear. Do you mean "$\mu_E \leq b^* \mu_0$"?

[7] Sensitivity=statistical power? This first sentence is confusing.

[7] I'm not sure how this picture clarifies the story about HT vs. TST. Note that the HT approach is sometimes referred to as a t-test approach and here as the NOEC approach. This type of switching of description will confuse the general readership of such a report.

[8] As I commented earlier, the "b" factor should reflect important biological / ecological shifts in the population response.

[8] What does "degree of protectiveness" in objective 2 mean?

[8] If you can't compare TST to the "point estimate" approach, then how was it possible to have permits written using either HT or point estimate approaches?

[10] CV on the y-axis is for controls? Effluent group? Both?

[10] 90%-tile of CV from 1989 and 2000 – CVs from control group?

[11] So what are you doing for the survival endpoints? What are you doing with counts? Are you using transformations (e.g. arc-sine-sqrt for proportions, sqrt for counts)?

[12] Does this table imply that a control condition was run with each of these tests (e.g. effluent, reference toxicity)?

[13] What does MSD mean for responses that are proportions (e.g. survival, germination)?

[15] Defining the mean response of the effluent here as μ_T should be done much earlier in this presentation. This would allow the bioequivalence/TST and HT approaches to be formally stated in terms of parameters.

[15] The level of change described as decision 3 is equivalent to the choice of "p" in ICp.

[16] The description of the Monte Carlo simulation is inadequate and confusing. Monte Carlo was not used to simulate WET data. You used a Monte Carlo simulation to study the TST and HT approaches by first generating WET data with known underlying characteristics and then applying the approaches.

[17] second analysis EMPIRICALLY DERIVED Type I and Type II error rates … with different "b" values defined for each calculation of these error rates.

[17] Shouldn't you also simulate cases when the effluent mean was equal to the control mean?

[19] Type I error rate was equal to 0? [Here, incorrectly stated as α error=0.] This is an observed Type I error rate.

[20] It is bad statistical practice to talk about preceding a test of means with a test of variances. The F-test is notoriously sensitive to violations of the normality assumptions while the test of means are very robust. You can use an unequal variance t-test routinely as an alternative. Finally, other tests of variances such as Levene's test are preferred to the F test for variance homogeneity or Bartlett's >2 group generalization (assuming you want to formally test this which I believe is debatable).

[20] 2 replicates vs. 4 replicates? What is a replicate here? Any reported simulation should be presented in sufficient detail so that someone could repeat your computer experiment. This presentation does not meet such a standard. Not only are the conditions unclearly presented but

the implementation of the simulation is sketchy at best. For example, how was the simulation programmed (in Excel? FORTRAN? SAS? R?)?

[21] mean effect levels versus an effect level defined as a change in mean response?

[21 mean percent effect ranges? What are these? Why these ranges?

[21] not as robust statistically? What does this mean? You don't know "truth" in this empirical exercise. This comparison simply tells you how often the 2 methods lead to similar/dissimilar decisions.

[23] No, the test statistic is NOT formed from the population means μ_c and μ_e (what happened to the μ_T formulation earlier?) but in terms of sample means such as $\bar{Y}_C$

In this figure, doesn't nontoxic means $\mu_e > b\ \mu_c$

[23] The formula for calculating the pooled variance makes sense only if there are the same number of observations in both effluent and control groups.

[24] No, this is not the SE(mean) but the SE(difference in sample means)

[24] Doesn't the TST approach calculates the SE($\bar{Y}_e - b * \bar{Y}_C$)?

[25] No, the t-test statistics do NOT involve population means; they are functions of sample means. This is fundamental and critical notation.

[25] Doesn't b=0.68 imply toxic if $\mu_e \leq 0.68\ \mu_c$? Here, and elsewhere in the report, a decrease in response is considered adverse. Was this ever explicitly stated in the report?

[25] Isn't it better to say "equivalent to control response" instead of "not toxic?"

[27] How was the MSD determined? How did you determine the power > 80%?

[27] A 1993 paper reported a similar result? Isn't this backwards and the 1996 paper reported a similar result to the 1993 paper?

[29] Need to comment/formally define the relationship between sensitivity and Type II error rates? Between specificity and Type I error rates?

[31] The table legend needs to be enhanced here. For example, how is effect level defined here? What do the Risk Management columns mean here? Isn't "b" a RM decision?

[32] How does "mean difference" in this figure relate to "effect size?"

[*] Many of the later pages of this report contain figures and discussion that were already criticized as part of the review of the summary. These observations will not be repeated here.

[45] A standard boxplot is a better display here (e.g. box with lines at Q1, median, Q3 and whiskers extending from min to Q1 and from Q3 to max).

[47] TST was a viable alternative prior to this empirical investigation. Bioequivalence has a long and well studied history with pharmaceutical applications.

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

TABLES

FIGURES

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.

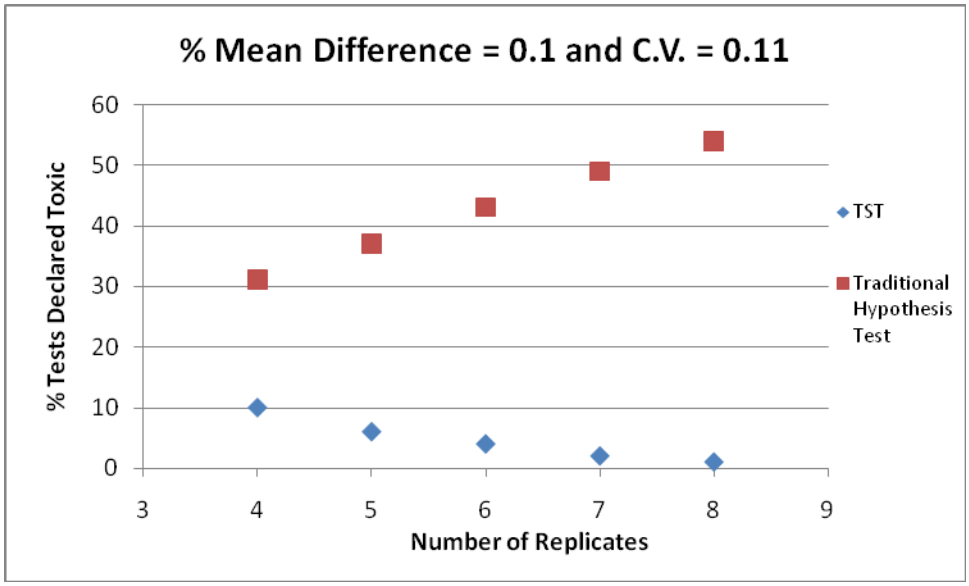


Figure ES-1. Percent of chronic fathead minnow WET tests declared toxic using TST having a mean effluent effect of 10 percent, above average control variability (CV = 0.11 or 11 percent) and an α = 0.25, as a function of the number of within-test replicates. Results using the traditional hypothesis test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

$$\begin{aligned} &\textit{Null Hypothesis}: && \mu_T \geq \mu_C \\ &\textit{Alternative Hypothesis}: && \mu_T < \mu_C \end{aligned}$$

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 - α. 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 - β, 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.

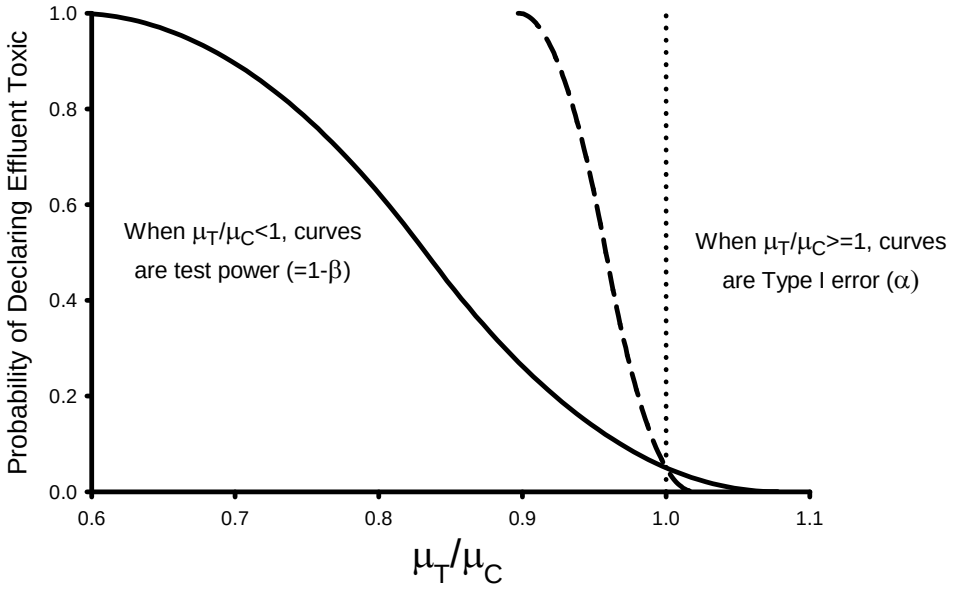


Figure 1-1. Example test performance curves for traditional WET hypothesis tests. The dotted line marks where the true mean biological measure in the effluent equals that in the control. The solid curve is for a high variability test, while the dashed curve is for a low variability test.

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.

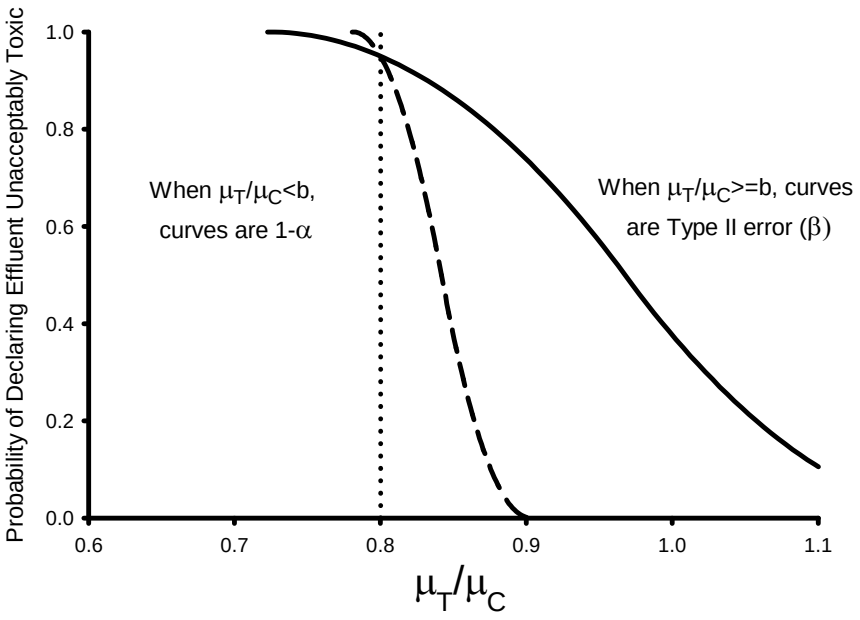


Figure 1-2. Example test performance curves for TST WET hypothesis tests. For this example, *b* is set to 0.8 (denoted by dotted line), with $\alpha = 0.05$. The two curves represent test performance for tests with high (solid line) and low (dashed line) variability.

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%

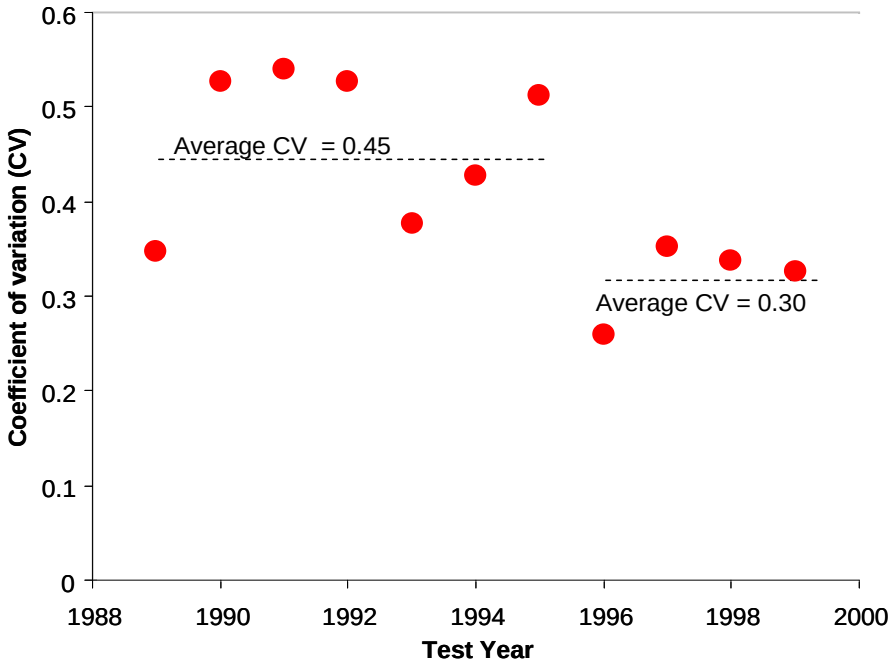


Figure 2-1. Summary of test variability (expressed as the control 90th percentile coefficient of variation or CV) observed between 1989 and 2000 for the chronic *Ceriodaphnia dubia* EPA WET test. This figure illustrates and supports the basis for using test data post 1995, as test precision improved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

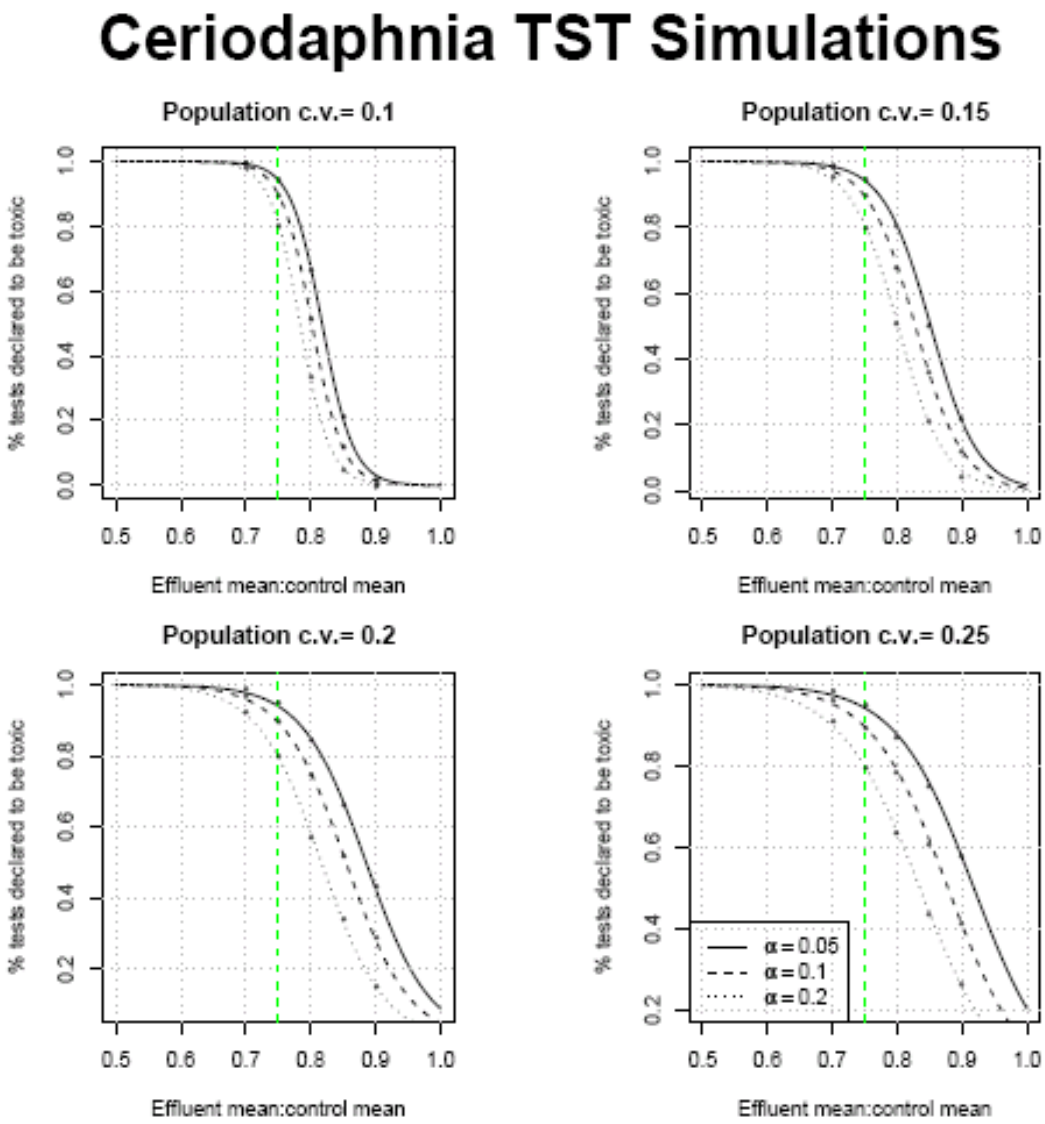


Figure 3-1. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs of 0.1, 0.15, 0.2, and 0.25 correspond to the approximate 25th, 50th, 70th, and 75th 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 α = 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).

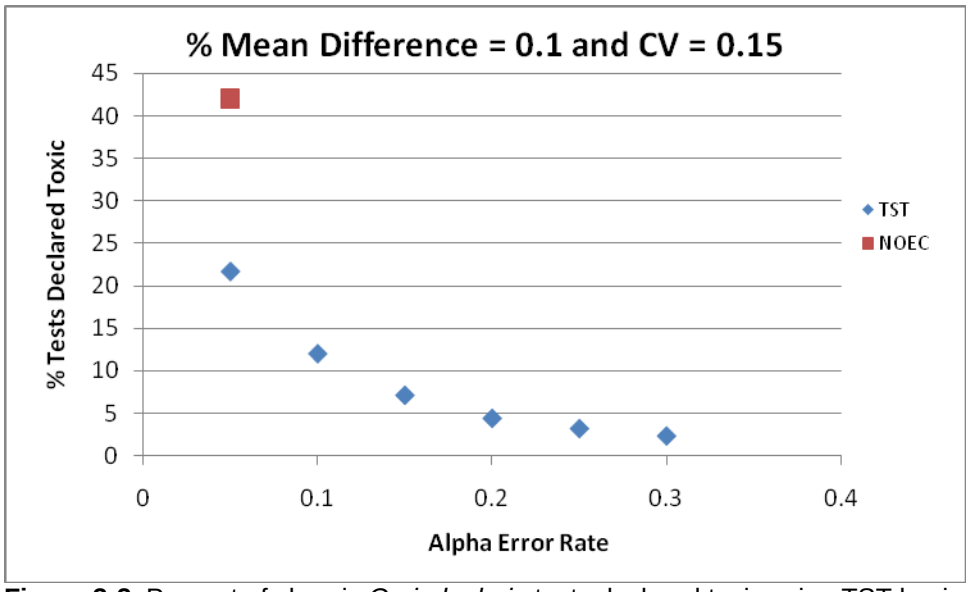


Figure 3-2. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

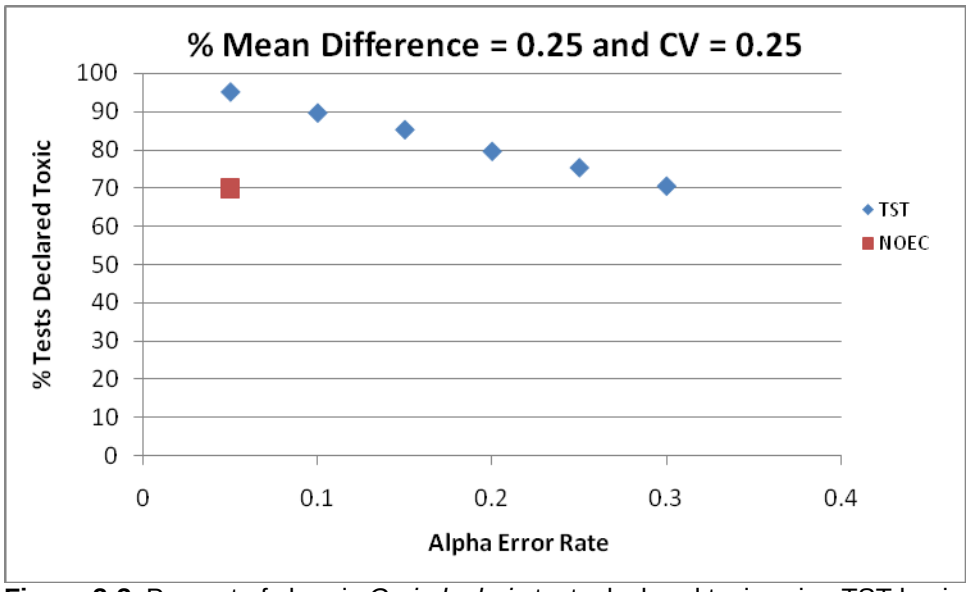


Figure 3-3. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and high control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

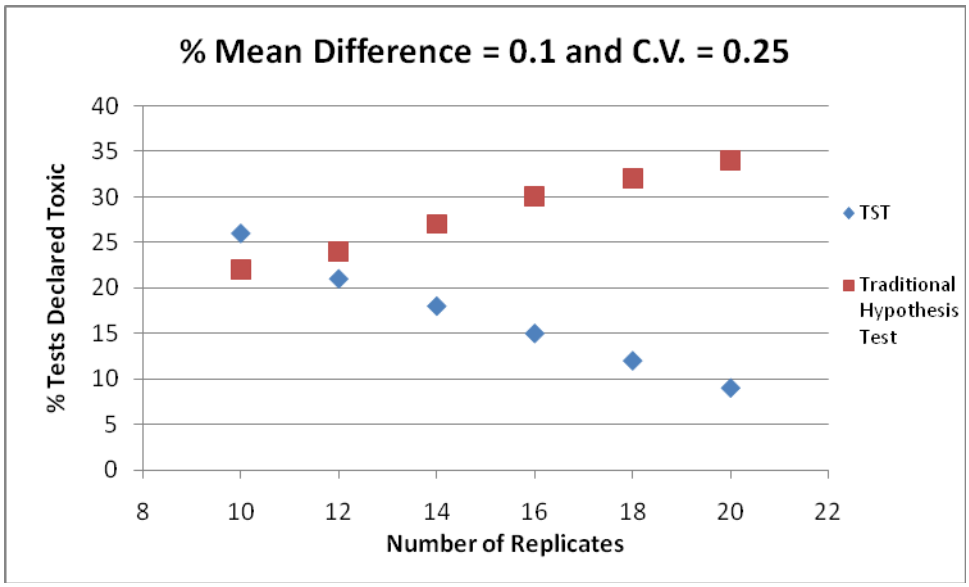


Figure 3-4. Percent of chronic *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 10 percent and above average control variability and α = 0.20, as a function of the number of test replicates. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

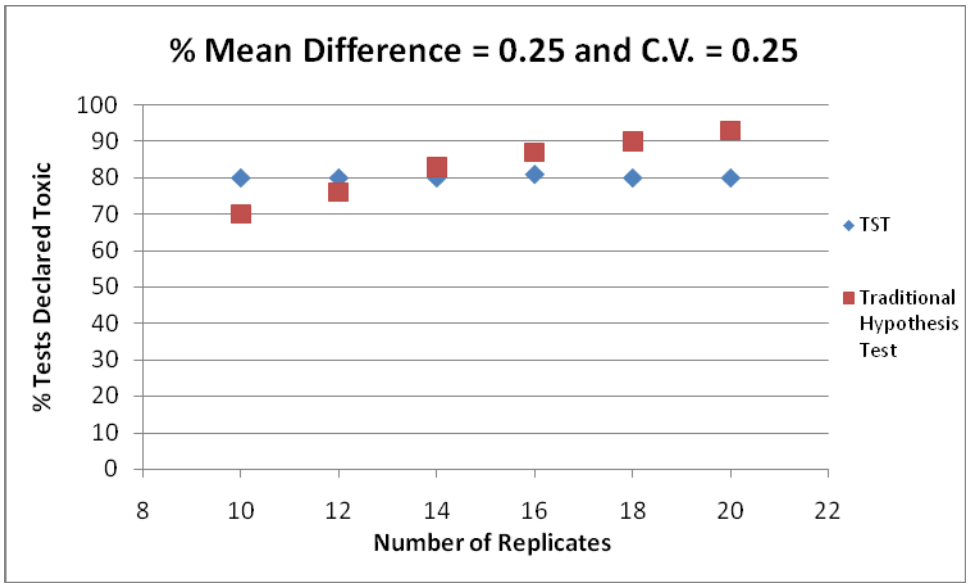


Figure 3-5. Percent of *Ceriodaphnia* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability (α = 0.20) as a function of the number of test replicates. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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 50^{th} and 75^{th} 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 75^{th} 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.

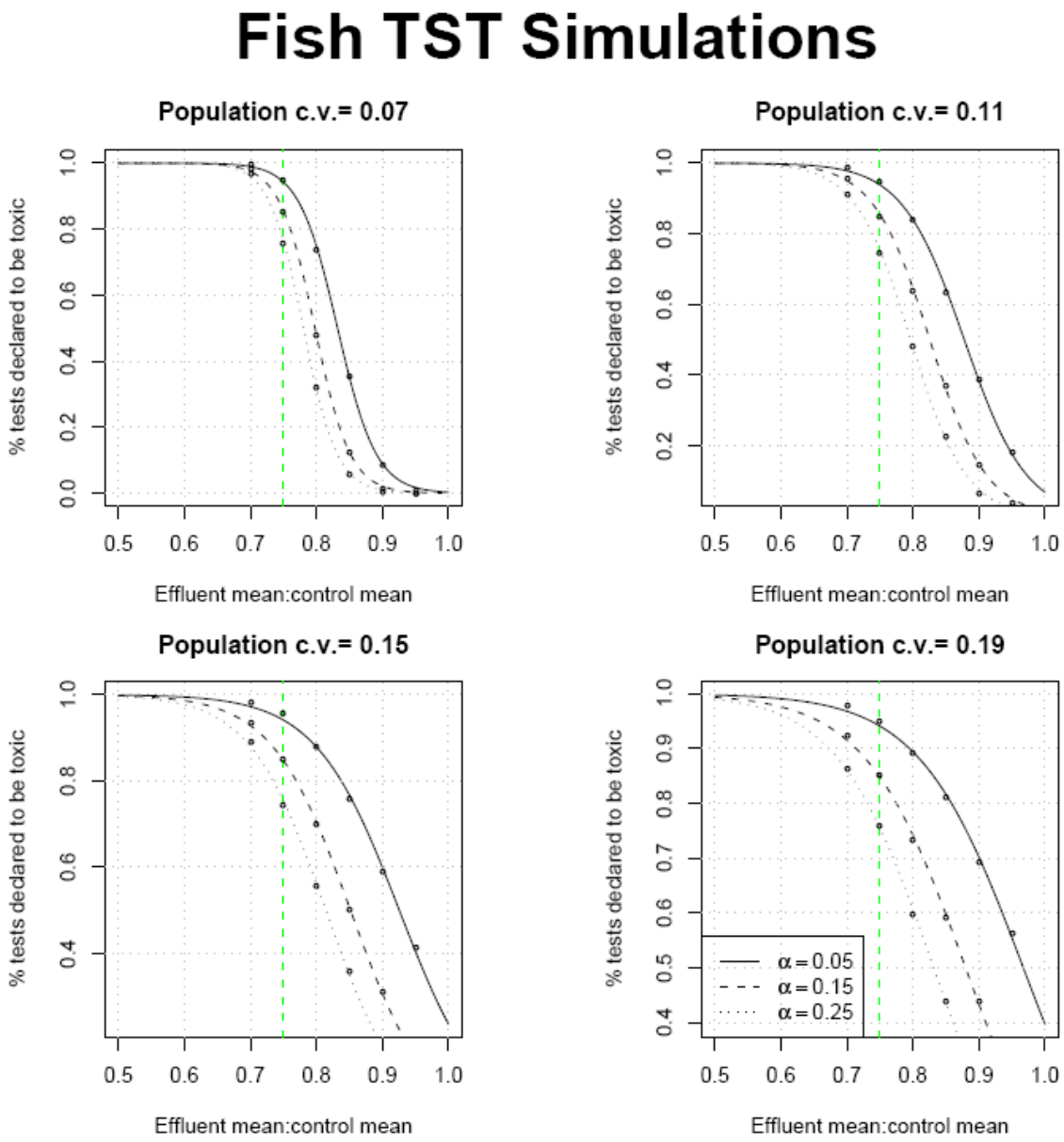


Figure 3-6. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic fathead minnow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

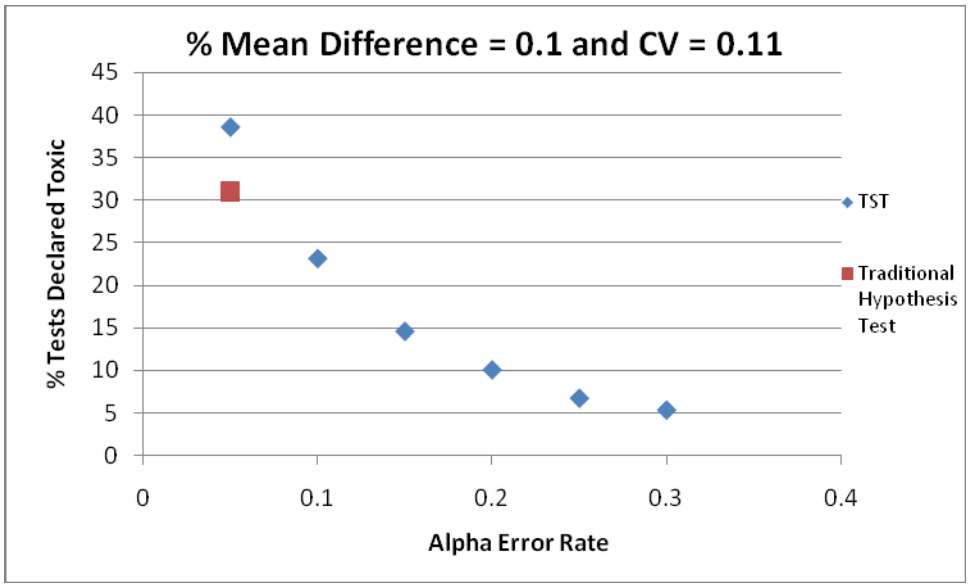


Figure 3-7. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional approach (α = 0.05) is shown as well.

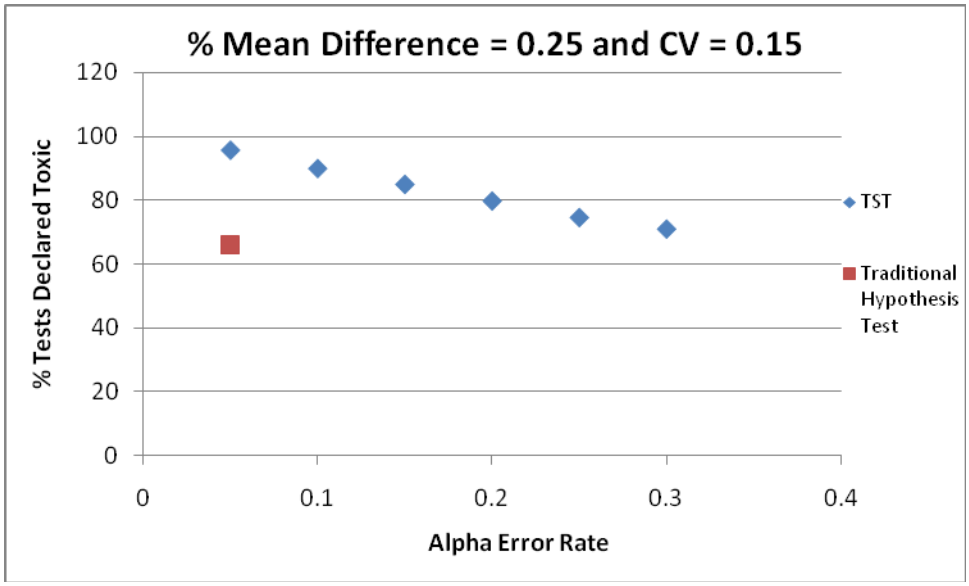


Figure 3-8. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional approach (α = 0.05) is shown as well.

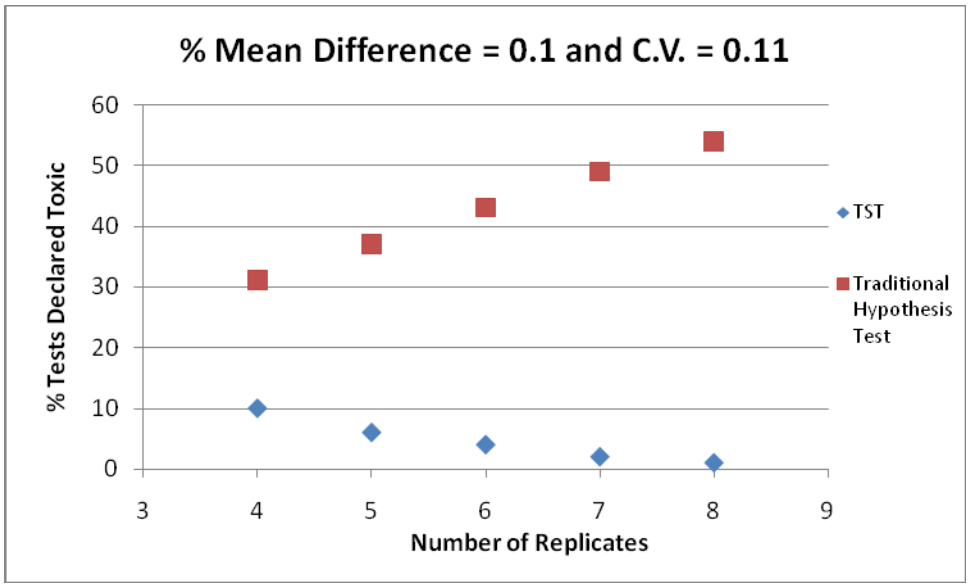


Figure 3-9. Percent of chronic fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability and an α = 0.25, as a function of the number of test replicates. Result using the traditional approach (α = 0.05) is shown as well.

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.

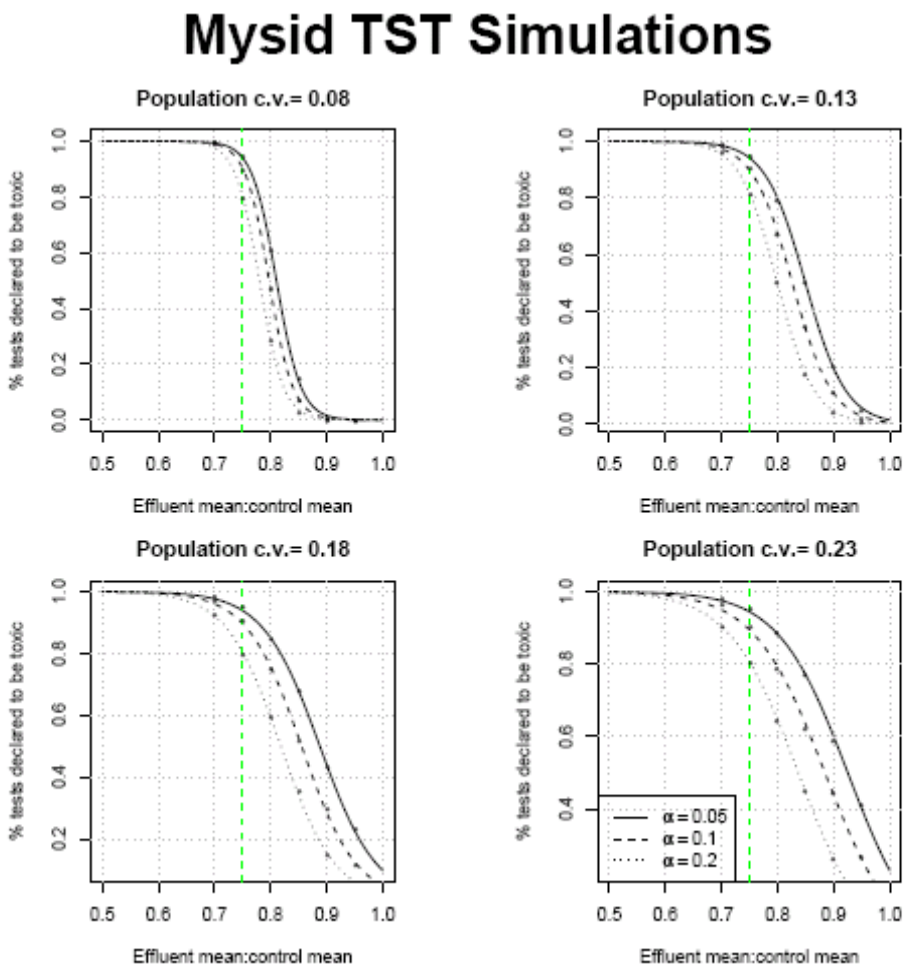


Figure 3-10. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 70th, and 90th percentiles for the chronic mysid WET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

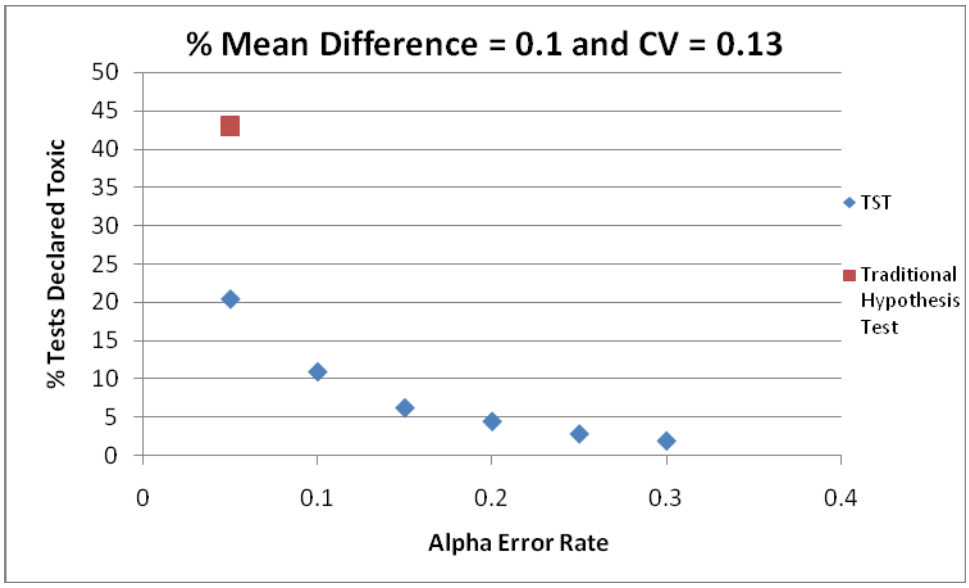


Figure 3-11. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

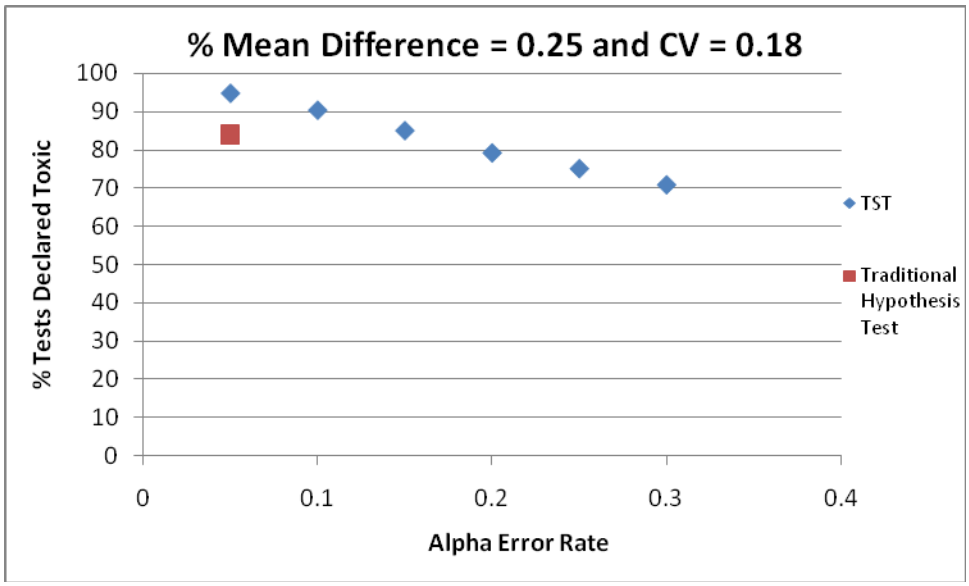


Figure 3-12. Percent of chronic mysid tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

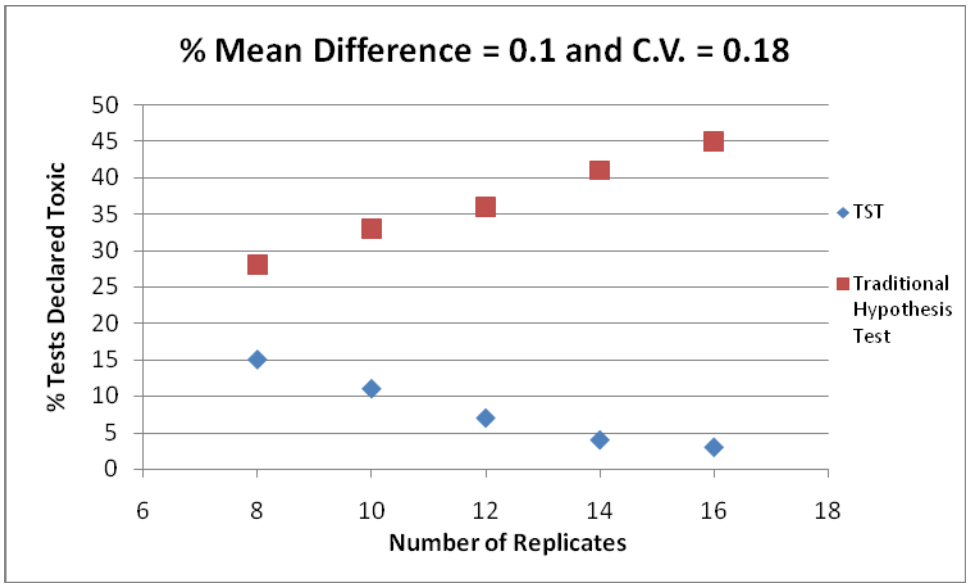


Figure 3-13. Percent of chronic mysid tests having a mean effluent effect of 10 percent and above average control variability declared toxic using TST and an α = 0.15, as a function of the number of test replicates. Results using the traditional hypothesis approach (α = 0.05) are shown as well.

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 ($\leq$ 10 percent mean effect) is declared toxic $\leq$ 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.

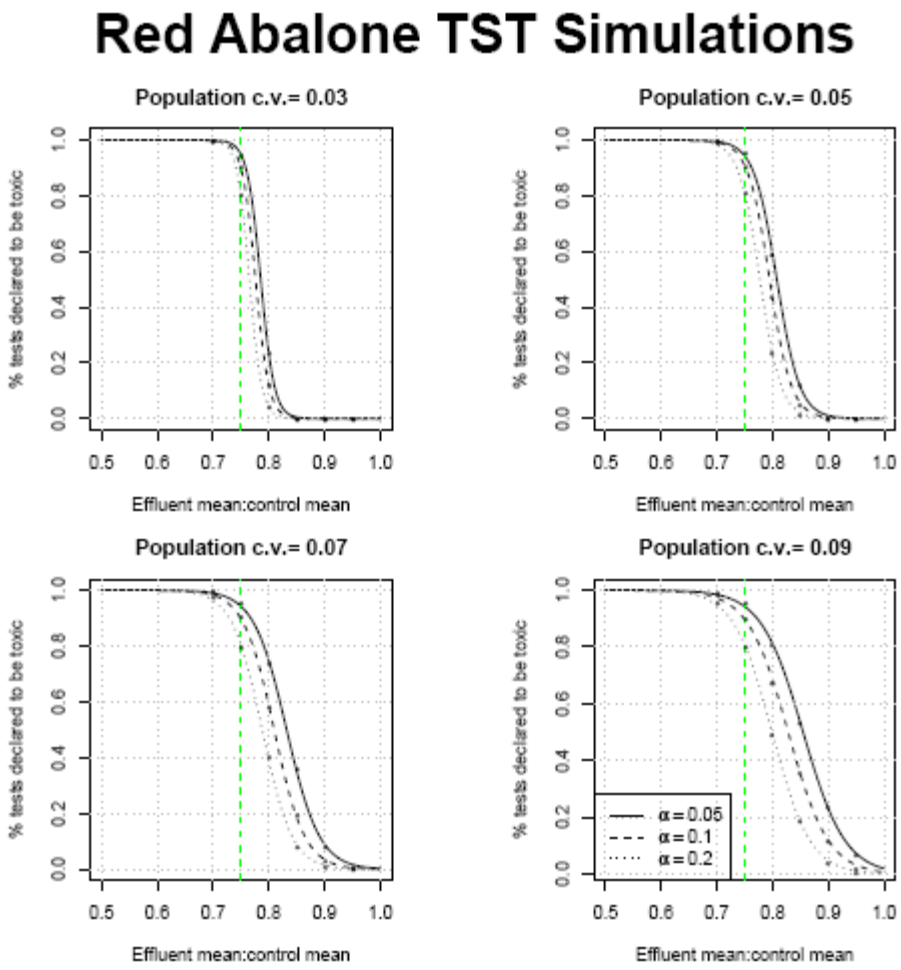


Figure 3-14. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 98th percentiles for the chronic red abalone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

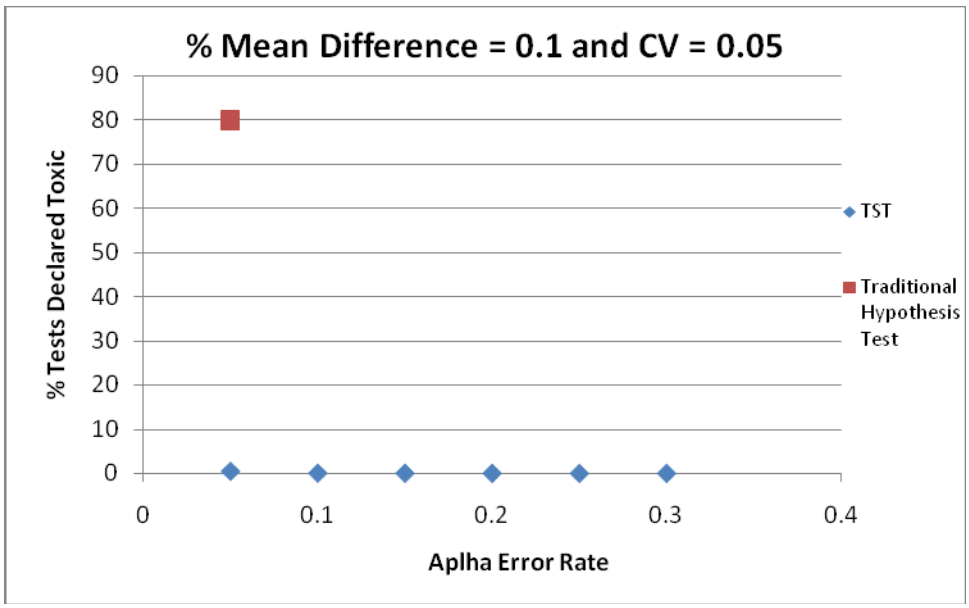


Figure 3-15. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

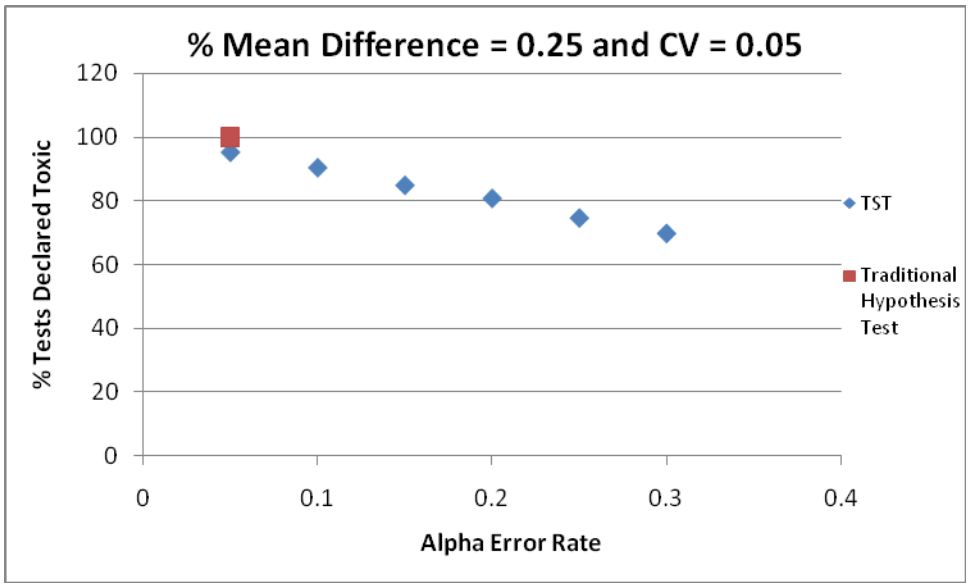


Figure 3-16. Percent of chronic red abalone tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

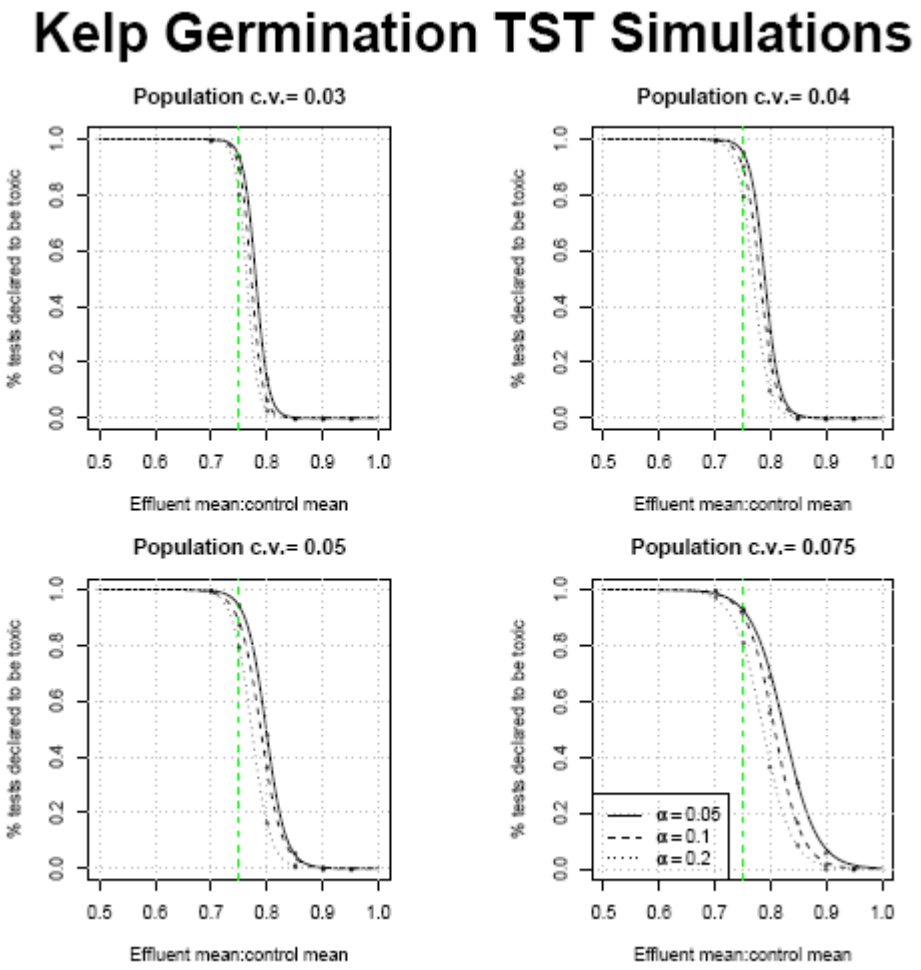


Figure 3-17. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 95th percentiles for the chronic giant kelp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

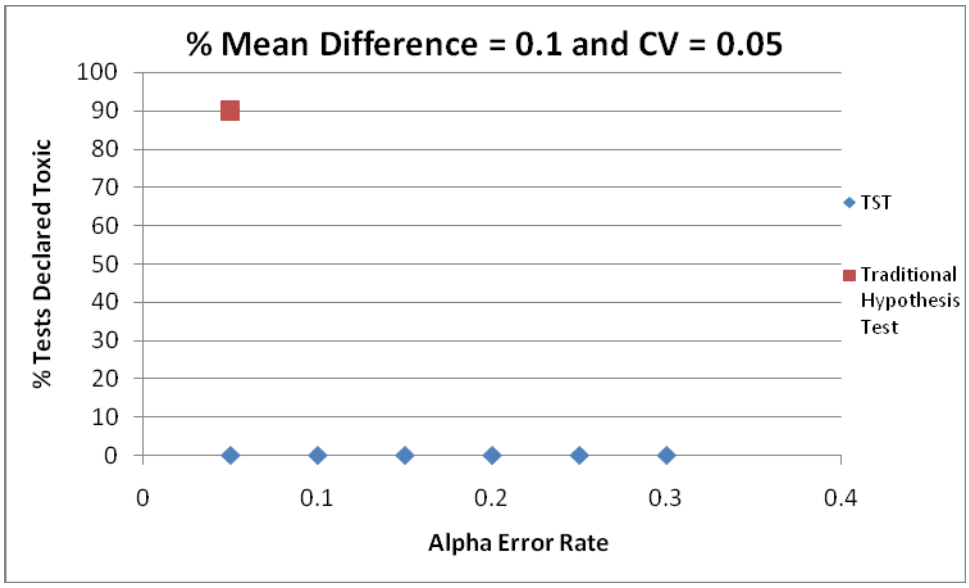


Figure 3-18. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

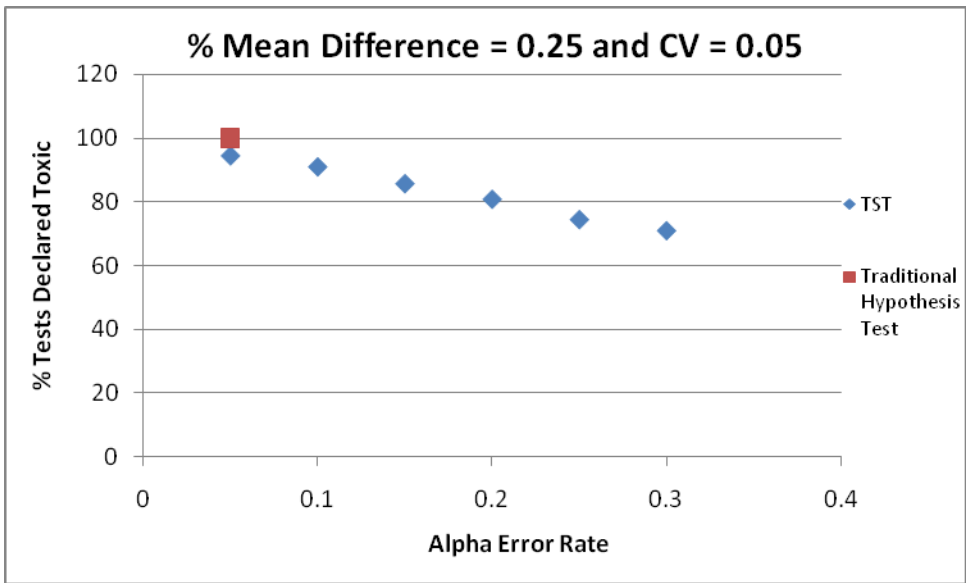


Figure 3-19. Percent of chronic giant kelp germination tests declared toxic using TST having a mean effluent effect of 25 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

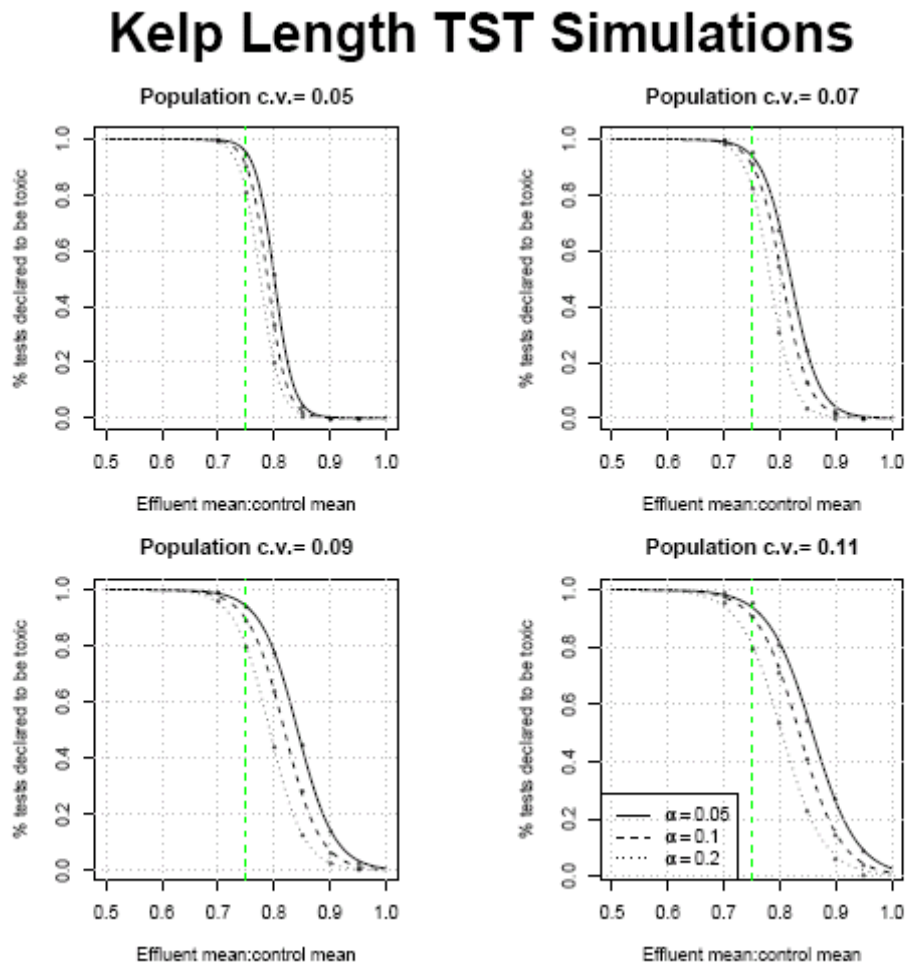


Figure 3-20. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic giant kelp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

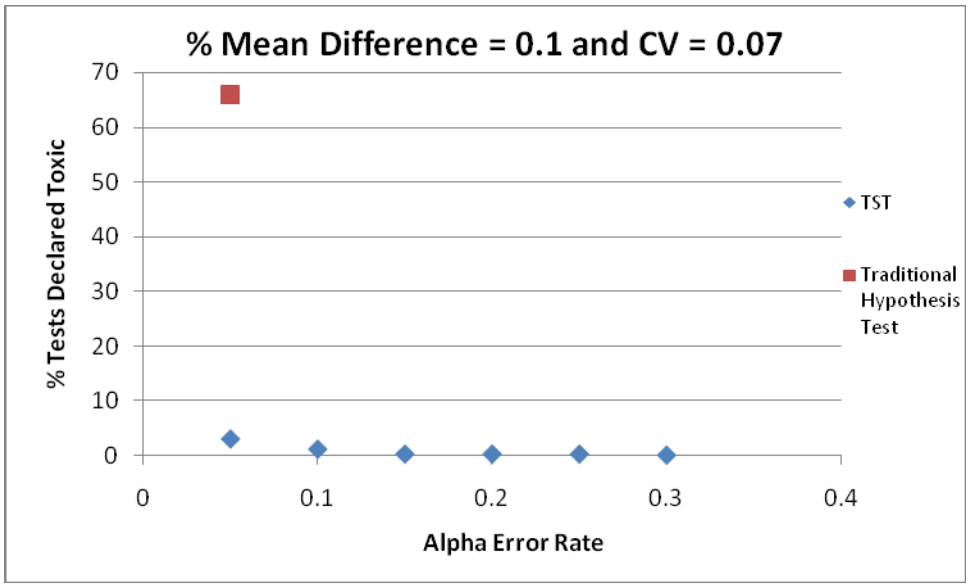


Figure 3-21. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

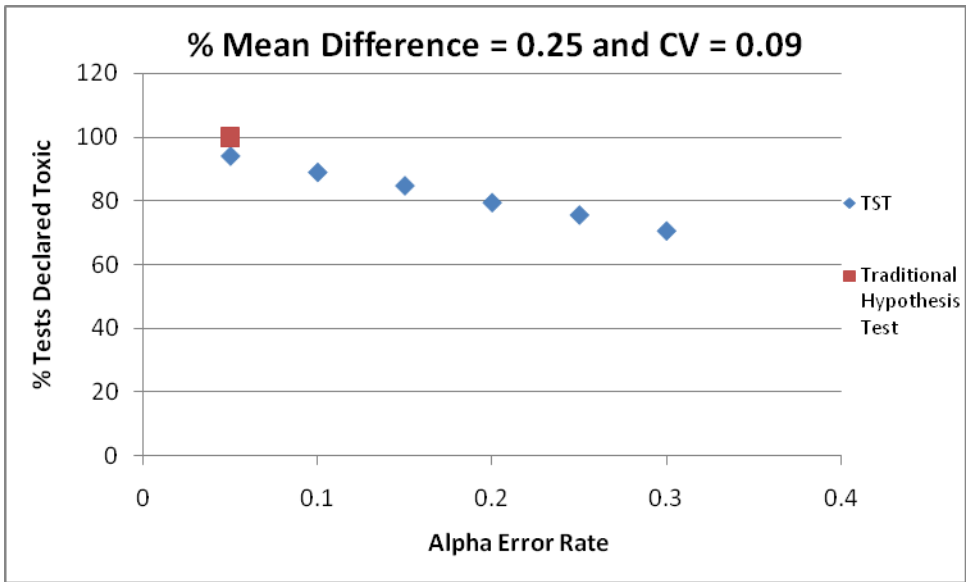


Figure 3-22. Percent of chronic giant kelp germ-tube length tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

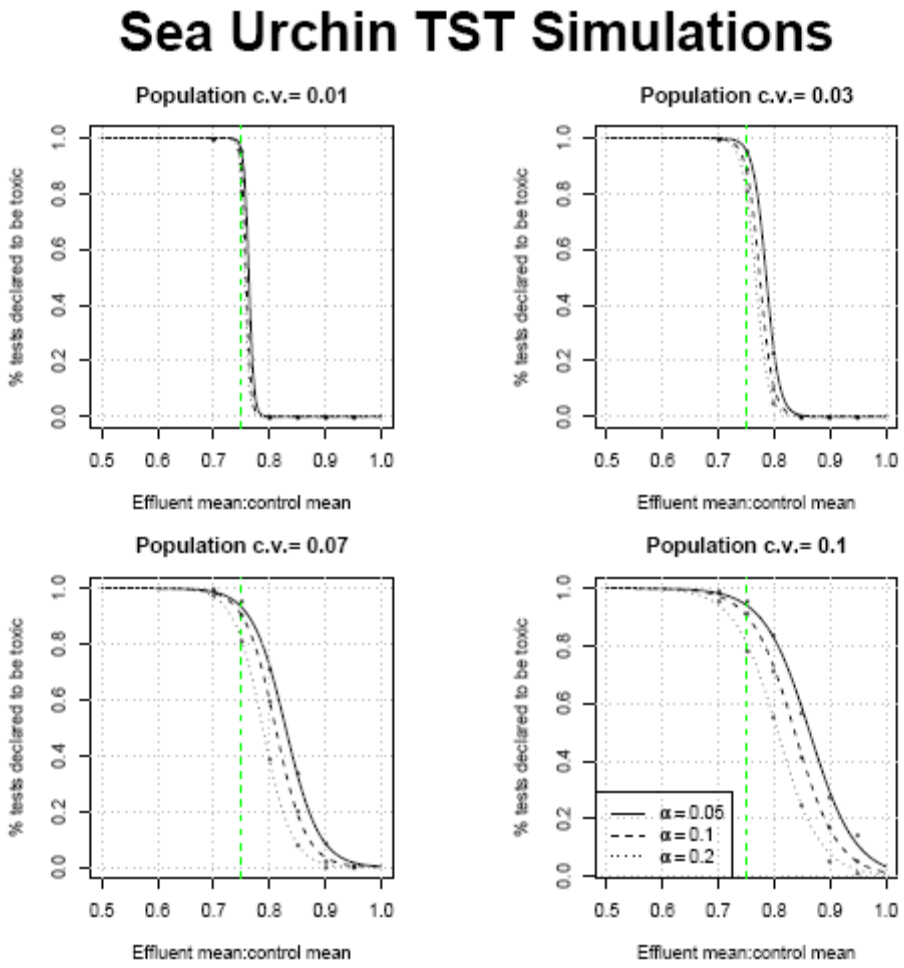


Figure 3-23. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 25th, 50th, 75th, and 90th percentiles for the chronic echinoderm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

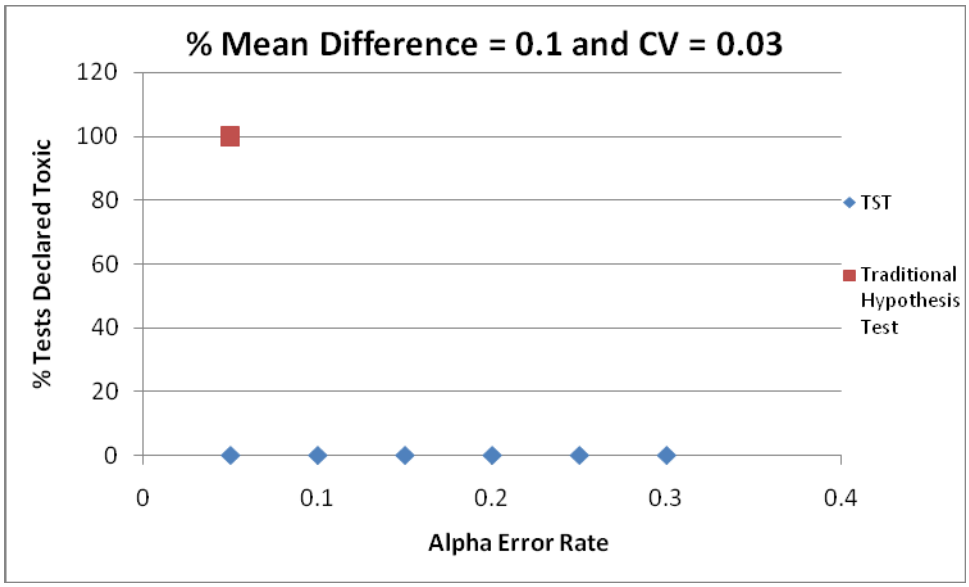


Figure 3-24. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of the α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

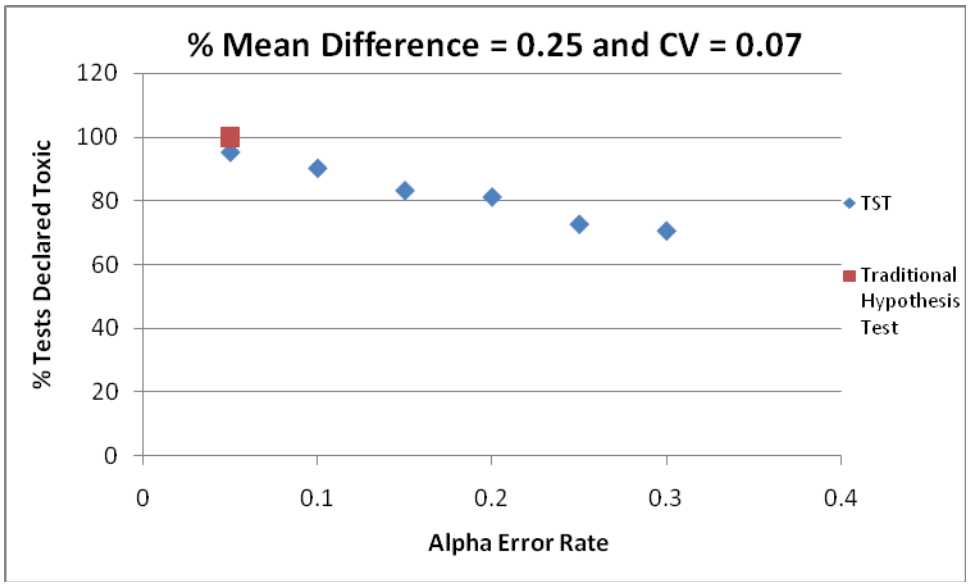


Figure 3-25. Percent of chronic echinoderm tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

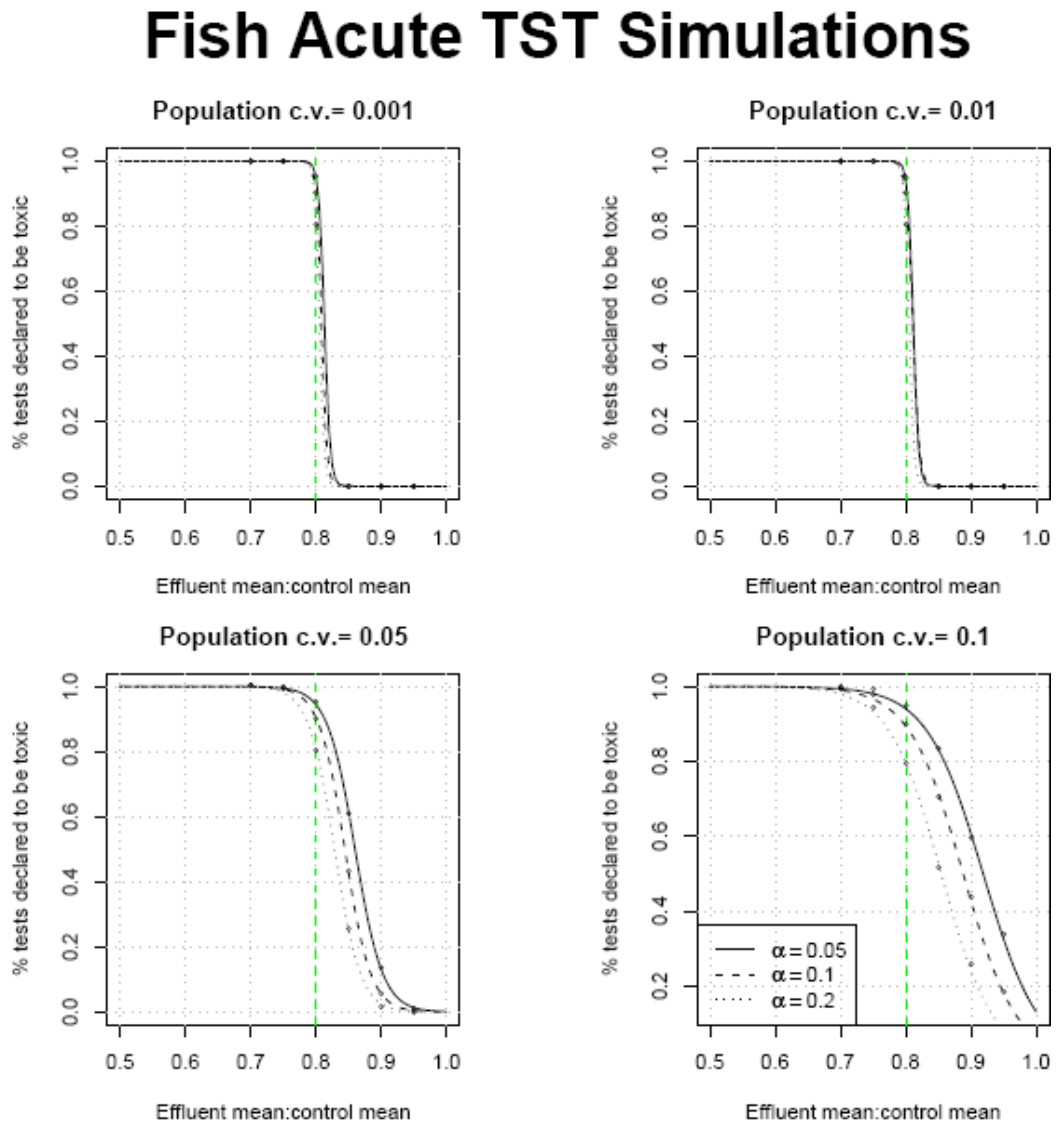


Figure 3-26. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 75th, 80th, 85th, and 88th percentiles for the acute fathead minnow 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.

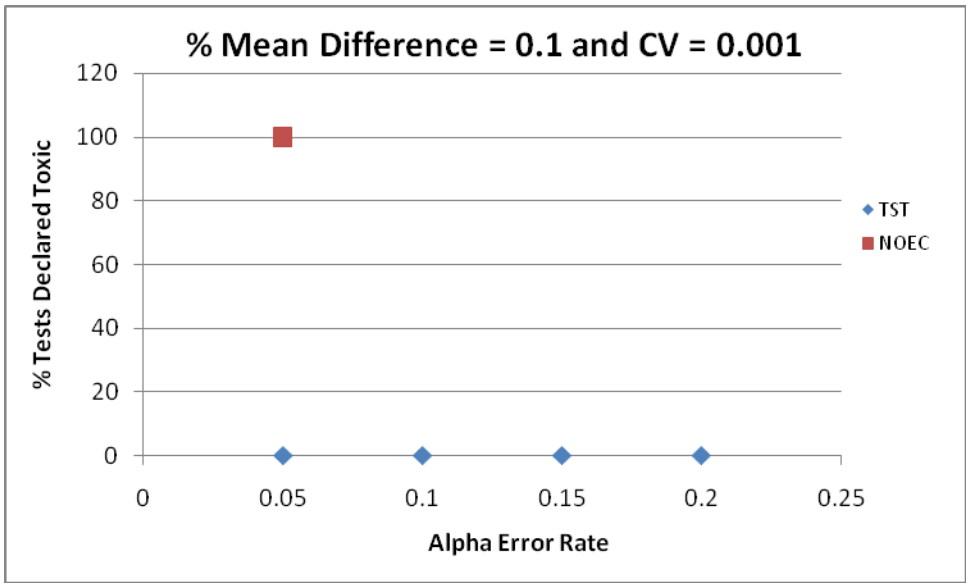


Figure 3-27. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

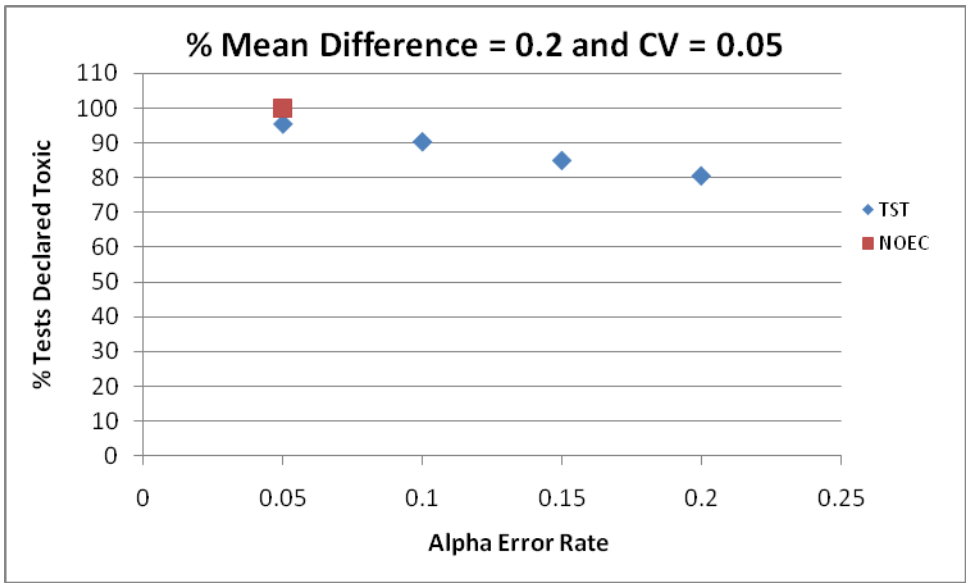


Figure 3-28. Percent of acute fathead minnow tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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 (~70^{th} 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 ~70^{th} 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.

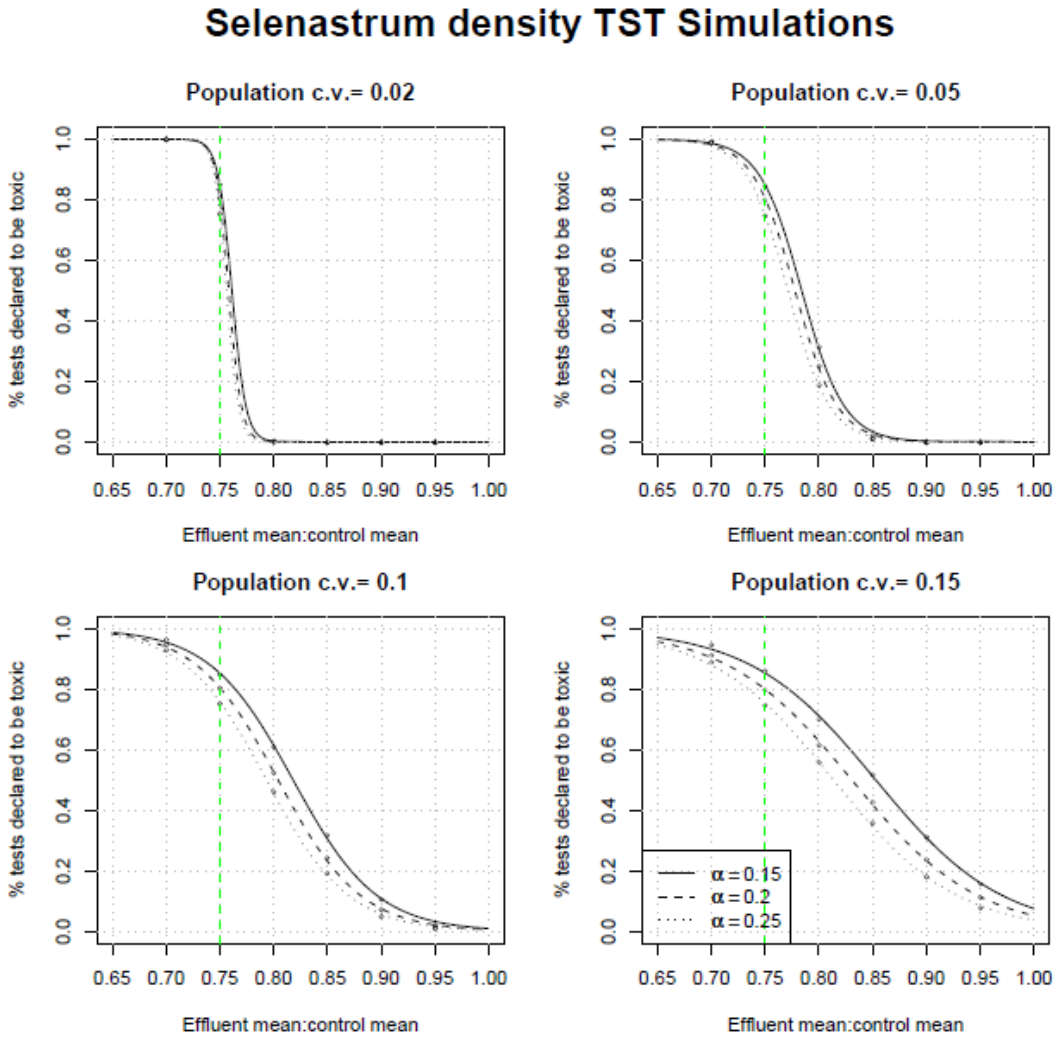


Figure 3-29. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. CVs correspond to the 10th, 40th, 70th, and 85th percentiles for the chronic *Selenastrum*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

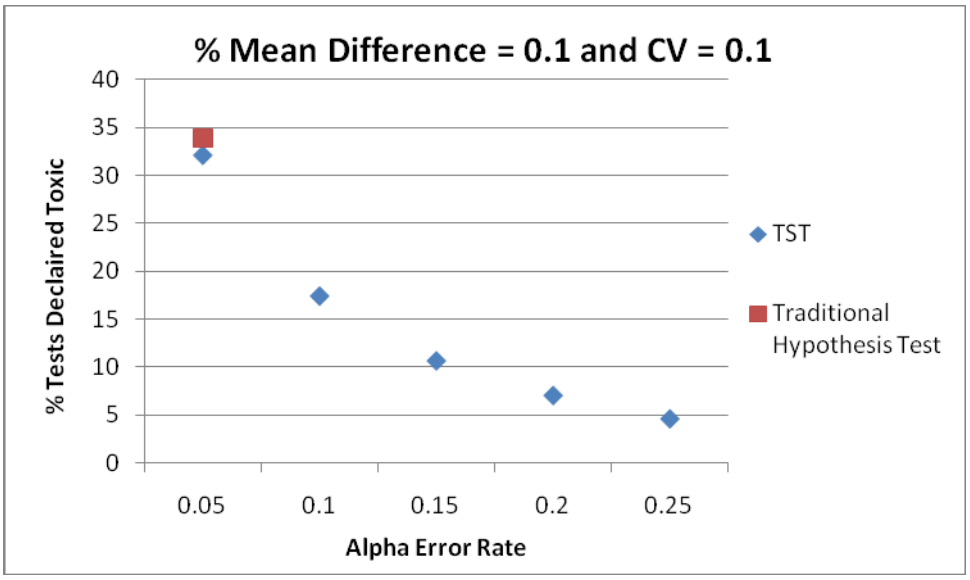


Figure 3-30. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

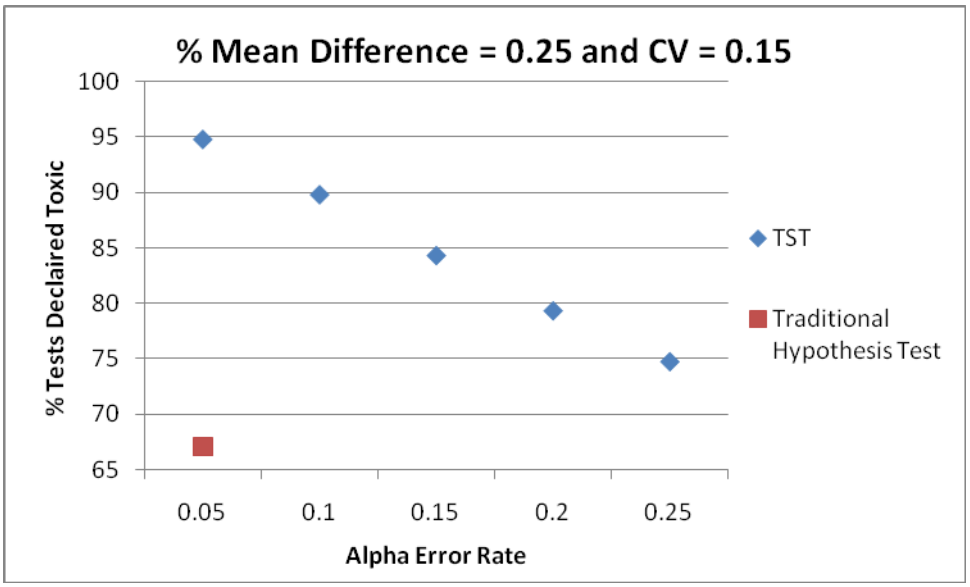


Figure 3-31. Percent of *Selenastrum* tests declared toxic using TST having a mean effluent effect of 25 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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.

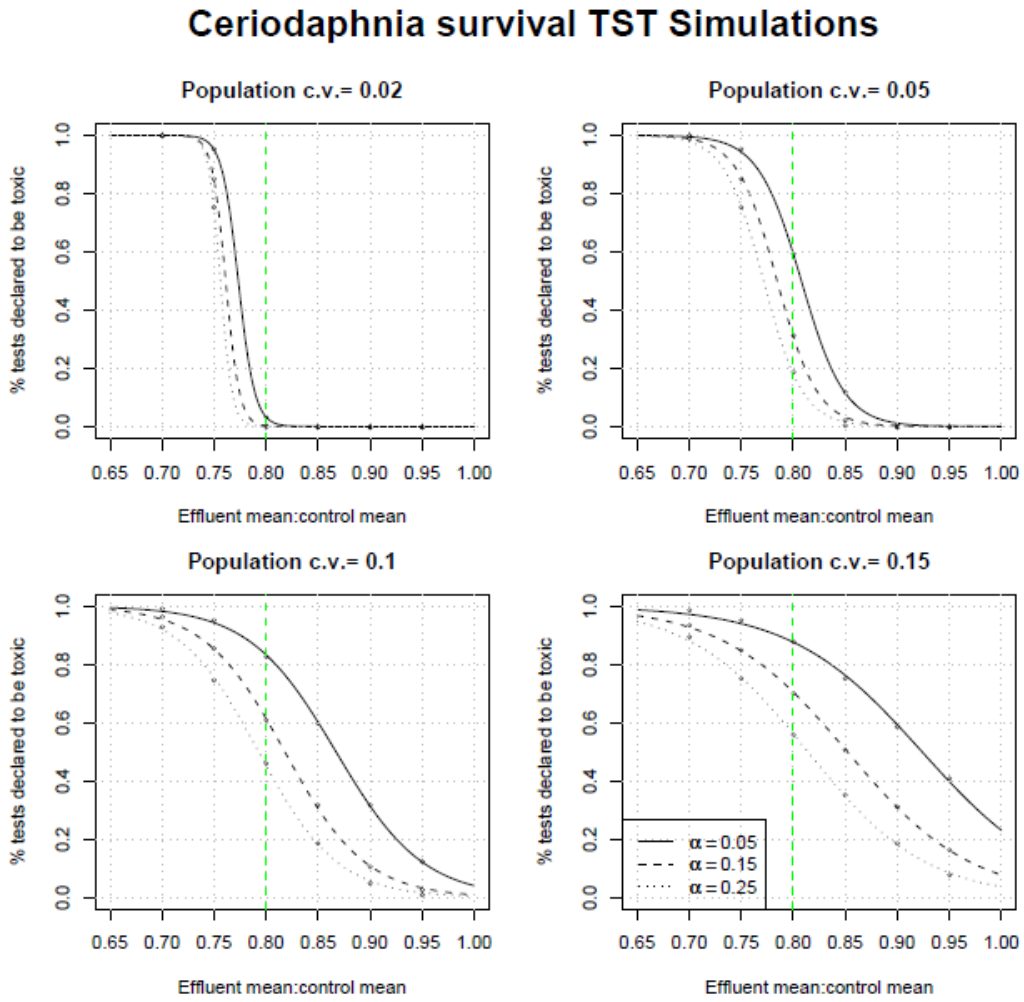


Figure 3-32. Power curves showing the percentage of tests declared toxic as a function of the ratio of effluent mean to control mean response and α level categorized by the level of control within-test variability. The first two CVs correspond to the 85th percentile, and the following two correspond 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.

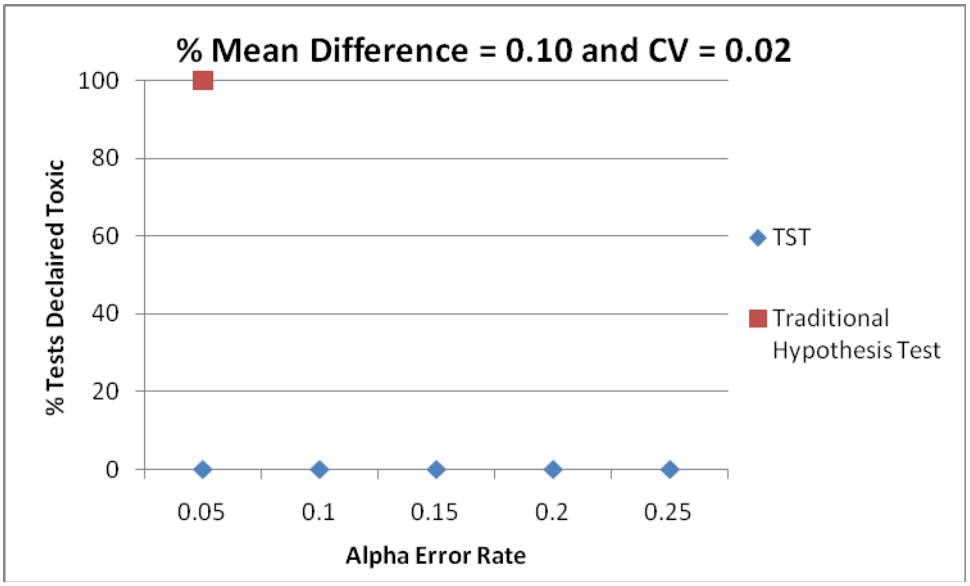


Figure 3-33. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 10 percent and average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

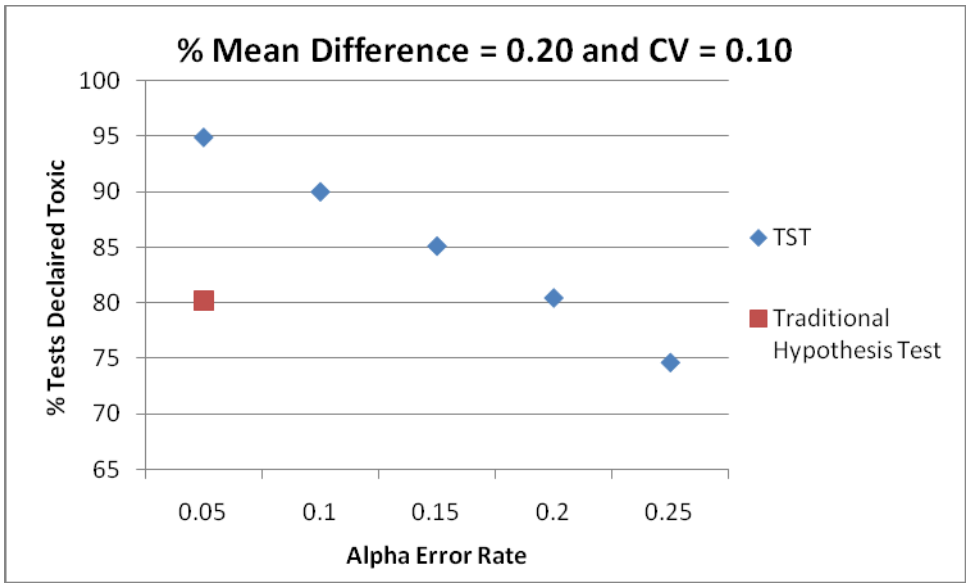


Figure 3-34. Percent of acute *C. dubia* tests declared toxic using TST having a mean effluent effect of 20 percent and above average control variability as a function of α error rate. Result using the traditional hypothesis approach (α = 0.05) is shown as well.

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 ($< 50^{th}$ 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 ($\geq 80^{th}$ 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 80^{th} 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 $\underline{\geq}\ b \times$ 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 $\leq 75^{th}$ 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.

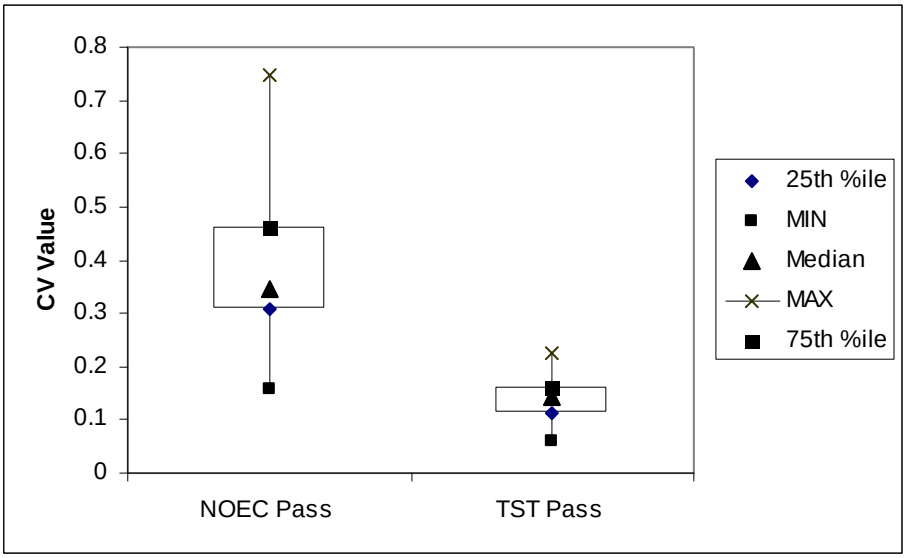


Figure 4-1. Range of CV values observed in chronic *C. dubia* ambient toxicity tests for samples that were found to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*). California's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

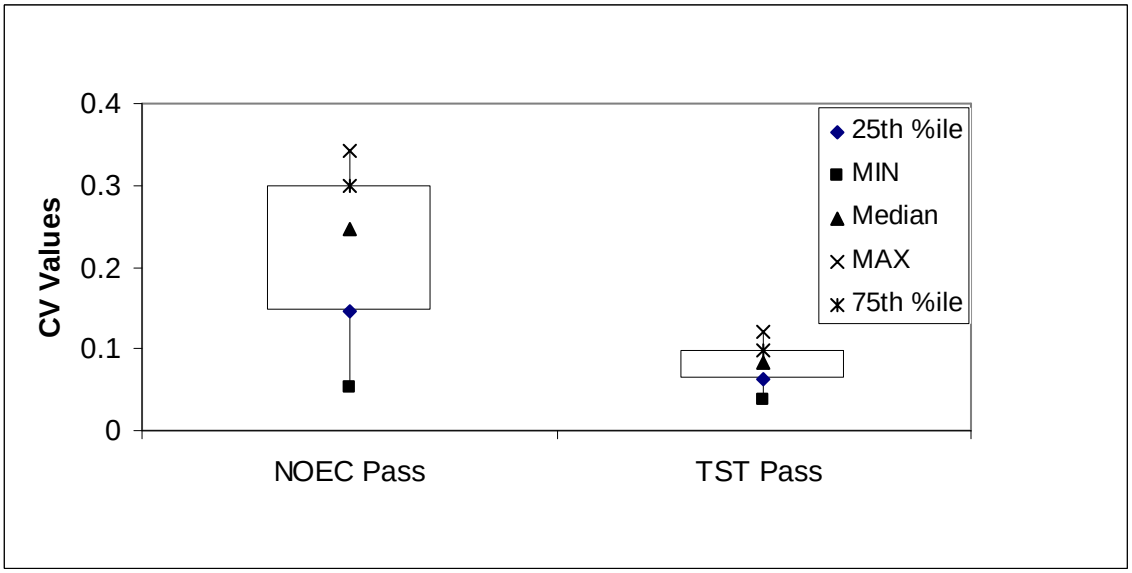


Figure 4-2. Range of CV values observed in chronic *P. promelas* ambient toxicity tests for samples that were declared to be non-toxic using the traditional t-test but toxic using the TST approach (*NOEC Pass*) and for those samples declared toxic using t-test but not the TST approach (*TST Pass*). California's 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 ($p < 0.01$)	# (%) 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).

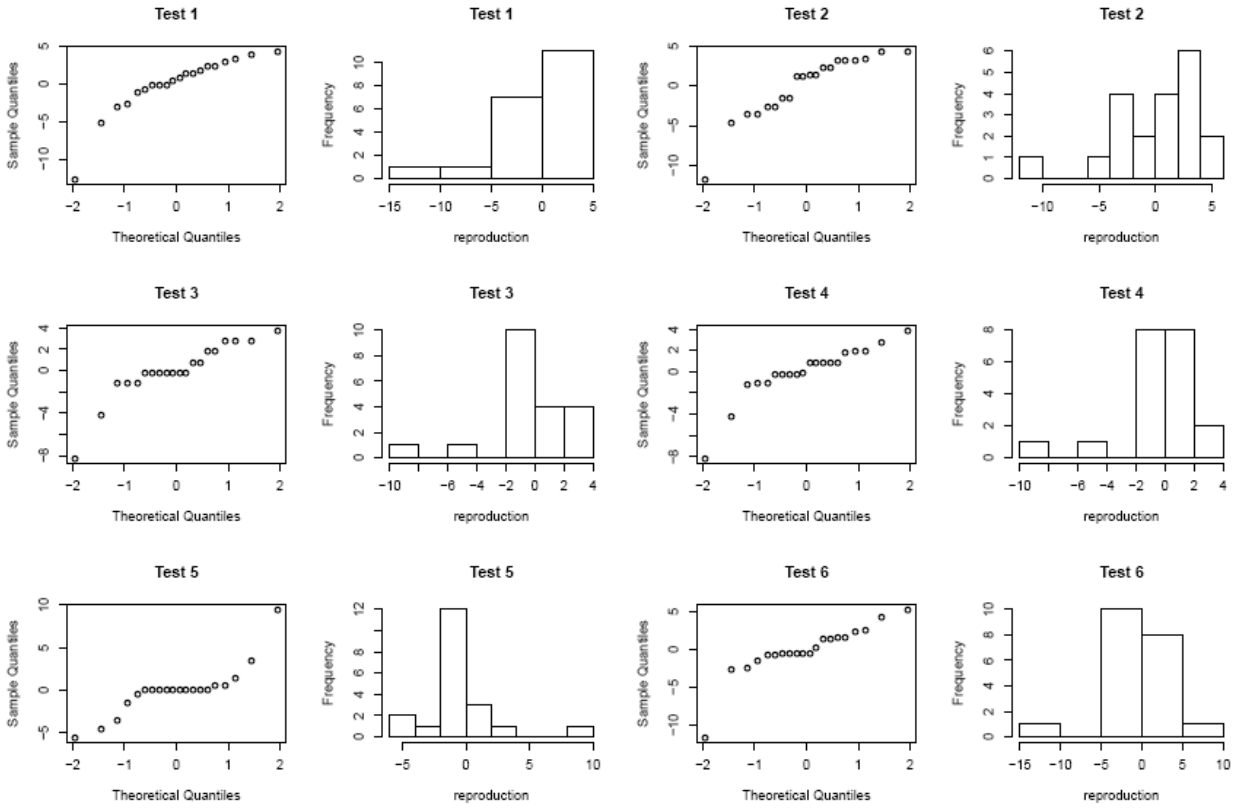


Figure A-1. Probability plots and histograms of examples of *Ceriodaphinia* chronic reproduction test data showing non-normal distribution and especially leptokurtic distribution.

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.

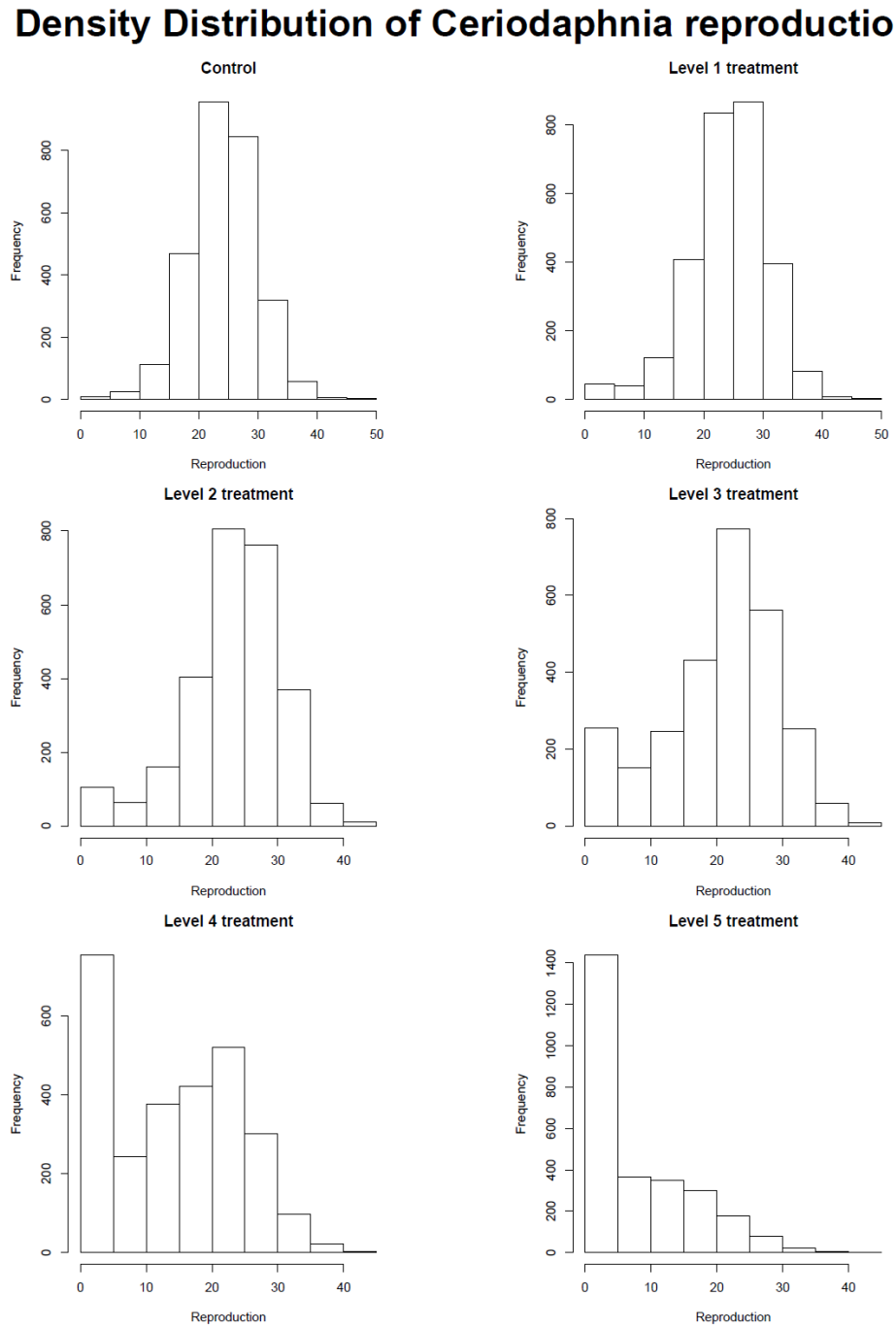


Figure A-2. Histogram of observed *Ceriodaphnia* reproduction at different level of effluent concentrations based on 281 multiple concentration tests.

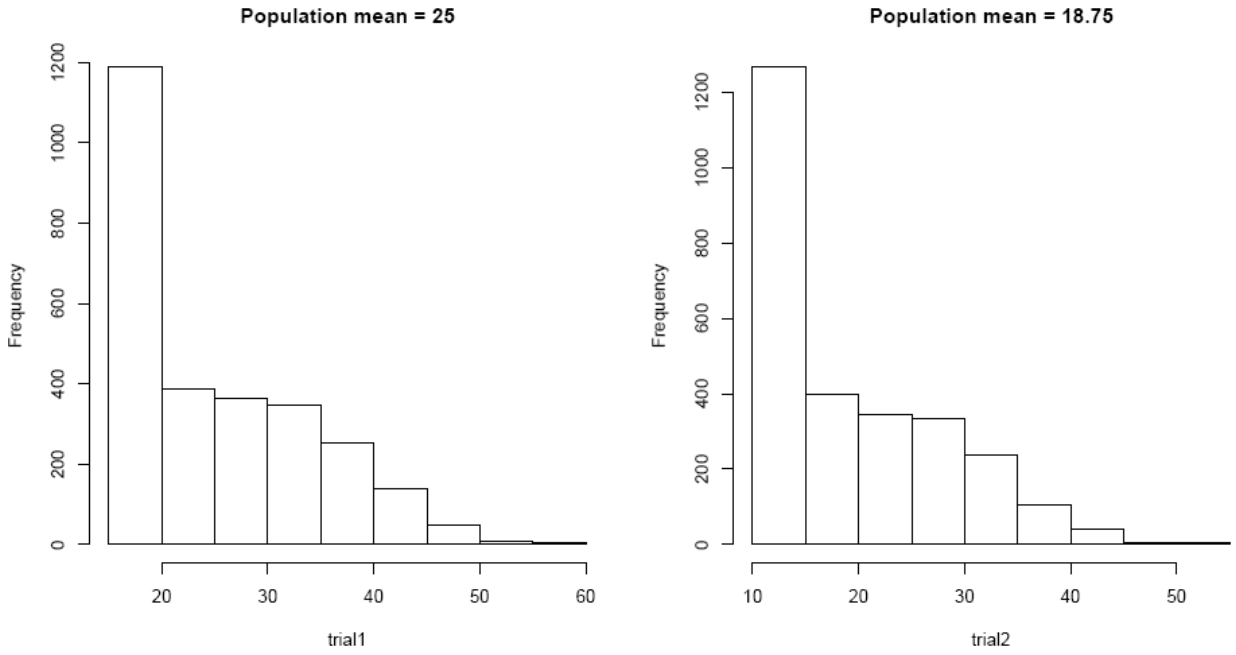


Figure A-3. Simulated frequency distributions of Ceriodaphnia reproduction data with two populations having non-normal data and different means. Both populations have a standard deviation of 7.7.

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

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

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*.

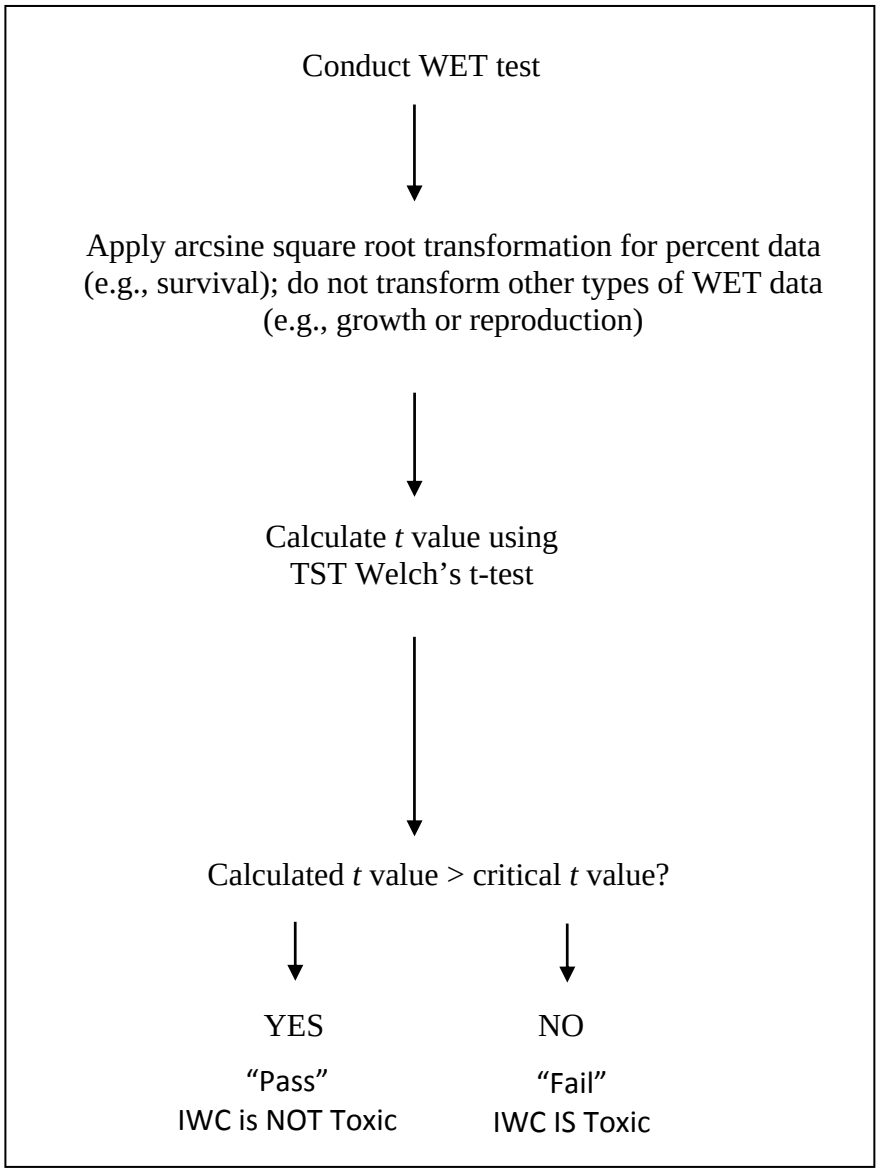


Figure B-1. Statistical flowchart for analyzing valid WET data using the TST approach for control and the IWC, receiving water, or stormwater.

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

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

JUN 18 2010

OFFICE OF WATER

SUBJECT: Final National Pollutant Discharge Elimination System (NPDES) Test of Significant Toxicity (TST) Implementation Document

FROM: James Hanlon, Director
Office of Wastewater Management

TO: Water Division Directors, R1-10

The purpose of this memorandum is to transmit to you a copy of the final guidance document, "National Pollutant Discharge Elimination System Test of Significant Toxicity Implementation Document" (EPA 833-R-10-003). This document provides an additional recommended statistical approach for analyzing WET test data used for whole effluent toxicity (WET) reasonable potential determinations and NPDES permit compliance.

EPA developed the TST approach to provide an additional scientifically valid, statistical application for assessing WET hypothesis test data. The TST assesses the measurement of toxic impacts from effluent on specific test organisms' ability to survive, grow, and reproduce and is based on research and peer-reviewed publications. The TST examines whether there is a biologically significant difference defined as the measured difference which has a detrimental effect on aquatic organisms to thrive and survive when compared against the normal condition (i.e., a control). Using a WET test, this biologically significant difference is the comparison between an effluent's in-stream waste concentration (IWC), as specified in the permit, and the control. The TST recommendations advance the applied science of the NPDES WET Program by addressing both the false negative and false positive error rates which have been a concern for both permitting authorities and permittees. We believe the TST approach addresses these false negative and positive concerns and provides an incentive to NPDES permittees to provide valid, high quality WET test data to enhance NPDES WET reasonable potential and permit compliance determinations.

The TST document was externally peer reviewed according to EPA's requirements and after addressing the peer review comments the document was sent out to EPA Regions and States for their review. Comments received from EPA Regions and States were addressed and, where appropriate, revisions were incorporated into the final document.

The TST approach does not preclude the use of existing recommendations for assessing WET data provided in EPA's 1991 Water Quality-based Technical Support Document (TSD) which remain valid for use by EPA Regions and the States.

To compliment your understanding of the attached final TST document, we have scheduled a second webcast on Wednesday, July 14, 2010, from 1:00 to 2:00 P.M. (EST). This webcast will provide an introduction to TST, including an overview of its scope and context; how the TST should be implemented; advantages of the TST over other statistical approaches; and conceptual examples demonstrating the TST application. Please watch for an E-mail with additional details about how to participate in the webcast. If you have questions, please contact Laura Phillips (phillips.laura@epa.gov, 202-564-0741) of my staff.

Attachment (1)

Cc: Mark Pollins, WED/OCE/OECA
Debra Denton, R9
Regional Branch Chiefs, R1-10
EPA WET Coordinators, R1-10