Draft Policy for Toxicity Assessment and Control

Response to Peer Review Comments

The external peer review of the draft Policy for Toxicity Assessment and Control (draft Policy) has concluded. The participating reviewers are listed below, along with their comments (in bold) and State Water Board staff (staff) responses. Staff would like to thank the peer reviewers for their insightful analyses of the draft Policy.

Participating Peer Reviewers

- Gerald A. LeBlanc, Ph.D.
  Professor of Environmental and Molecular Toxicology
  Department of Environmental and Molecular Toxicology
  North Carolina State University

- Michael C. Newman, Ph.D.
  A. Marshall Acuff Jr. Professor of Marine Science
  Department of Environmental and Aquatic Animal Health
  Virginia Institute of Marine Science
  College of William and Mary

Gerald A. LeBlanc, Ph.D. Review

1. Inclusion of the Test of Significant Toxicity in the draft Policy for Toxicity Assessment and Control. The test of significant toxicity (TST) provides an efficient, cost-effective means of evaluating instream waste concentrations (IWC) for toxicity. The approach is statistically sound, reduces burden associated with the assays, and, by structuring the assay around a hypothesis of significant toxicity, provides incentive for precision in assay performance.

Response: Staff agrees with this assessment of the TST approach.

2. Use of effect level reporting in compliance determination. The use of maximum daily effluent limitations (MDEL) and average monthly effluent limitations (AMEL), as described in the draft document, is an effective and appropriate approach to confirming toxicity associated with the IWC and triggering remediation activity. It is not evident why response levels of double those set for the TST are used in establishing compliance with these limitations. It seems that exceeding the TST, a statistically sound measure, should trigger exceedance of the MDEL.

Response: The decision to set the MDELs at double the percent effects of the established regulatory management decisions for chronic and acute toxicity was policy-based. Prior versions of the draft Policy directly applied the percent effects of the established regulatory management decisions as effluent limitations, but stakeholder concerns regarding the sensitive nature of toxicity tests prompted revisions.

Furthermore, it is not clear why exceeding the MDEL (<0.4 effect level) requires a minimum of one follow-up test; while a lesser effect (>0.4<0.8) requires at least two follow-up tests. I would seem that any "fail" test should require at least two follow-up tests to ensure that the AMEL is not exceeded.

Response: The purpose of the follow-up test is not to verify toxicity. Rather, this test is for dischargers to demonstrate the effectiveness of any corrective action they have taken to eliminate toxicity. This option exists to prevent the implementation of an accelerated monitoring schedule when a toxic event is shown to be an isolated case of equipment failure or operator error.

<u>3. Comparative approaches to toxicity analyses.</u> Three design approaches to toxicity assessment are described: the NOEC approach, the point estimate approach, and the TST approach. The three approaches provide decidedly different descriptors of toxicity as is well described in the Draft Staff Report. Selection of an appropriate approach should be driven by the information needed from the toxicity assessment.

The NOEC approach seeks to define, or bracket, the threshold effect concentration of the toxicant as being between the lowest observed effect concentration (LOEC) and the no observed effect concentration (NOEC). This approach seeks to define the concentration of the toxicant below which the defined effect is not expected to occur.

The point estimate approach utilized a concentration-response curve, as defined though the testing of multiple concentrations of the toxicant, to interpolate a defined point on the curve (e.g., LC50 as the concentration that is expected to be lethal to 50% of the exposed organisms). The 50% response level is typically used as the endpoint in this approach as it has the greatest statistical strength. However, other endpoints also are used. For example, the EC05 (concentration of the toxicant that is expected at affect 5% of the exposed organisms) can be used as an estimator of the threshold effect concentration. The point estimate approach is most typically used to quantify the relative toxicity of a material, but also can be used to estimate threshold effect concentrations.

The TST approach seeks simply to determine whether a defined exposure level of the test material causes a response that significantly deviates from controls. It provides no insight into the threshold effect concentration of the material, nor does it define the relative toxicity of the test material. The purpose of routine and accelerated monitoring of wastewater is to test the hypothesis that the waster does exhibit measurable toxicity. All three test designs can be used to test this hypothesis. However, the TST provides the most direct, cost-effective, yet statistically sound means for establishing a lack of toxicity associated with the IWC.

Response: Staff agrees with this summary. It should be noted, however, that the multi-concentration toxicity tests required during accelerated monitoring may bracket the threshold effect concentration of an effluent.

<u>4. Utility of the proposed accelerated monitoring schedule.</u> The proposed accelerated monitoring schedule, as described in the Draft Document, meets both federal requirements and state data needs. No concerns are noted.

Comment noted.

5. The Big Picture. **The draft policy follows closely US EPA guidelines and no significant scientific concerns are noted. This reviewer is satisfied that the guideline will prove effective in the sagacious monitoring of wastewater for toxicity. The following are some general items that could improve clarity of the document.**

Part I: Definitions, O. Replicate. **Replicates are used to measure or quantify variability, but not to control variability.**

Response: Staff agrees with this suggestion and will revise the definition accordingly.

Throughout the document, "response" and "effect" seem to be used interchangeable. The terms are not interchangeable since a "response" is associated with the test organism; while, "effect" is associated with the wastewater (e.g., the organisms respond to the effects of the wastewater). This misuse is most glaring in the equation on page 5 where response measures appear to be used to quantify an effect level. The units on both sides of the equation (response and effect) should be the same.

Response: Staff reviewed a number of journal articles and other scholarly writings and indeed, the terms are often used as described by the peer reviewer. However, alternative usages are also common. In the interest of greatest clarity, staff has chosen to use the two terms in a manner consistent with U.S. EPA guidance documents on toxicity testing, most specifically the National Pollutant Discharge Elimination System Test of Significant Toxicity Implementation Document. Therefore, no changes will be made to the draft Policy.

Part III.A.1 **If all species exhibit no response to the IWC when establishing the most sensitive species for use in wastewater toxicity evaluations, is the discharger required to continue to use three species when evaluating reasonable potential?**

Response: Test species sensitivity screenings and reasonable potential analyses are conducted using the same set of toxicity tests (except in the case of Major POTWs where reasonable potential is assumed). All test species sensitivity screenings and reasonable potential analyses for chronic toxicity require the use of one vertebrate, one invertebrate, and one plant, while acute toxicity requires the use of one vertebrate and one invertebrate. The relative percent effect at the IWC for each of the organisms will be used to assess the most sensitive species in the event that each test results in a "pass."

How does a discharger deal with a receiving water (with no waste water discharge) that is inhospitable to one or more of the species evaluated?

Response: So long as the receiving water has designated aquatic life uses, effluent monitoring, including toxicity testing, is typically required. U.S. EPA's toxicity test methods include the option (often preferred in most cases) of using synthetic laboratory culture water for dilution rather than upstream water, particularly if the upstream water is toxic to the test species.

It would be helpful if all abbreviations used in the document are defined in Part I.

Response: Staff agrees and will define every term assigned an acronym in Part I of the draft Policy.

Michael C. Newman, Ph.D. Review

1. Inclusion of the Test of Significant Toxicity in the draft Policy for Toxicity Assessment and Control. The proposed methods are a major advance from the currently compromised NOEC approach to a useful and statistically defensible one. Despite its widespread regulatory use, the current NOEC/LOEC approach is fundamentally indefensible as explained throughout our literature (e.g., Chapman et al. 1996, Hoekstra and van Ewijk 1993, Jager 2012, Kooijman 1996, Laskowski 1985, Newman 2008, Van der Hoeven 1997, Warne and Van Dam 2008).[1] In contrast, a well-reasoned, ratio-based equivalence method (such as, the TST) that sets *a priori* α, β, and a non-zero effect size (ES,b) is sound and appropriate. Beyond its ability to fulfill the Water Board's immediate needs, it generates results amenable to future interpretation by Bayesian or meta-analysis methods.

The initial use of two treatments instead of a series of treatments has substantial merit because several treatments would provide little additional insight. The TST uses fewer organisms and costs less. Or, with the same number of organisms and expense, the TST could allow stronger inferences than the conventional approach.

Response: Staff agrees with the assessment of the TST approach.

The TST, as presented in the draft documents, incorporates a Welch's t-test (Welch 1938), that is, it accommodates unequal variances. This appears different from equivalence tests explored by others for this purpose (e.g., Dixon 1998, Denton et al. 2011, Diamond et al. 2011, Erickson and McDonald 1995, and Shukla et al. 2000). These other papers performed calculations with an estimated common variance. Welch's t is reasonable for the regulatory purposes of the TST although it has the very minor disadvantage of making df calculations more tedious. Unlike two-tailed equivalence test methods described in Dixon (1998), the proposed TST is a one-tailed test with the response mean being lower for the IWC sample than for the control/reference. One-tailed testing seems reasonable given the application although some responses could be positive. The power will be slightly better for the one-tailed test *if the assumption of a decrease in response is correct*.

Response: The reviewer states that the TST documents appear to incorporate a different equivalence test than what has been explored by others for the same purpose. While the statement is true for some of the references cited by the reviewer (e.g. Erickson and McDonald 1995), it is not true for the Denton et al. 2011 or Diamond et al. 2011 publications. Both of these publications relied upon Welch's t-test as did the U.S. EPA TST documents. The work conducted by U.S. EPA, and used in the Denton et al. and Diamond et al. papers evaluated toxicity data having unequal variances, and a special appendix was included in U.S. EPA's California Toxicity Testing Methods Analyzed Using the Test of Significant Toxicity (TST) Approach document describing this evaluation, demonstrating the robustness of Welch's t-test

[1] Indeed, conventional null hypothesis significance testing (NHST) that is the foundation of the NOEC/LOEC method is being judged with increasing frequency in many sciences today to be fundamentally wrong. As examples, see Altman 2004, Anderson et al. 2000, Fidler et al. 2004,2006, Gigerenzer 2004, Hurlbert and Lombard 2009, Ioannidis 2005, Kruschke 2011, Lecoutre et al. 2001, McCloskey 1995, Sterne and Davey Smith 2001, Trafimow and Rice 2009, Wacholder et al. 2004, Ziliak and McCloskey 2004.

and TST with unequal variances. In addition, the Test Drive Analysis of the TST, conducted by the State Water Board, also evaluated toxicity data with unequal variances. The papers cited by Denton et al. and Diamond et al. are consistent with U.S. EPA's TST documents and California's proposed use of the TST approach. We agree with this reviewer's observation that one-tailed testing, as recommended in U.S. EPA's TST documents, and proposed in the draft Policy has better power than a two-tailed test and that the assumption of a decrease in response is correct for toxicity testing.

Compliance of NPDES Waster/Point Source WDR Discharger:
The equivalence testing method with specified error rates and effect sizes will be much more useful and easily implemented than the previous NHST-based NOEC method. The decision sequence as testing continues through time is reasonable and unlikely to create difficulties for dischargers. As the State and regulated community become more experienced with the approach and its results, meta-analysis of the equivalence testing results could potentially enhance decision-making. Details for such techniques can be found in Chow and Liu (1997, 2000) and Lopes and Neves (2010) although those discussions focus on human health applications. In the 2000 Chow and Liu chapter, they focus on meta-analysis of results from two-period, two-drug cross-over experimental designs. Although the TST approach does not involve such experiments, Section 13.3.2 and 13.3.3 might provide insight useful in discussions of the future for the TST approach. Perhaps such meta-analyses could eventually be used as quality control charts are used now for chemical analyses?

Response: Staff appreciates this recommendation for future consideration, as this reviewer states that all such suggestions are intended for "[…] future evaluation, not to identify flaws in the draft TST approach."

Analytical Function for Storm Water and Channelized Dischargers:
The details in the draft report including Appendix D present no obvious difficulties. Implementation of TST for this group will be a distinct improvement relative to existing methods.

Response: Comment noted.

<u>2. Use of effect level reporting in compliance determination.</u> These steps seem reasonable and clear. The proposed AMEL seems better suited, in my opinion, than the alternative.

Response: Comment noted.

Further enhancements might be explored as regulators and the regulated community gain more experience with the approach and its results. Again, a straightforward meta-analysis could be explored for combining results of the original and mandated repeated test results. Also, a Bayesian vantage on the TST (see Dixon (1998), pages 283-284) might enhance the approach and allow formal integration of testing results through the mandated sequence of tests. Earlier test(s) could be used to generate prior probabilities (e.g., Min and Zellner 1993, Sutton 2001) for the repeated tests.

Response: Staff appreciates this recommendation for future consideration, as this reviewer states that all such suggestions are intended for "[…] future evaluation, not to identify flaws in the draft TST approach."

3. Comparative approaches to toxicity analyses. A *properly conducted* hypothesis test is very useful to the Water Board for deciding whether or not to take action from evidence-based "toxic"/"not toxic" outcomes.

The NHST-based NOEC is an improperly conducted hypothesis test (see references cited on page 1 for details). Within the general NHST convention, rejection of the null hypothesis that there is no effect does not logically lead to the conclusion that there is an effect. Making such a conclusion involves a fundamental misinterpretation of NHST. The p-value from a NHST is the probability of getting the results or more extreme results if the null hypothesis were true, i.e., p(Results|H_0=True). It is not the probability that the null hypothesis is true given the results of the test, that is, not p(H_0=True|Results). This can be easily shown with the Bayes-LePlace Theorem,

$$\mathbf{p(H_0 = True\ Results} = \frac{\mathbf{p\ \ H_0 = True\ \ p(Results|H_0 = True)}}{\mathbf{p(Results)}}$$

Clearly, p(Results|H_0=True) is not p(H_0=True|Results). Nor, despite longstanding NOEC convention, is 1 minus the p-value a general estimate of the probability that the sample is toxic, i.e., not p(H_{Toxic}=True|Results$^+$). It should be clear from the above reasoning that 1 - p(Results|H_0=True) is not p(H_{Toxic}=True|Results). Further, Trafimow and Rice (2009) demonstrated the weakness of any argument in support of the HNST convention by stating that the correlation between 1 - p(Results|H_0=True) and p(H_{Toxic}=True|Results) is good enough for practical purposes. Such a correlation argument is demonstrably false.

The probability of the NHST alternative hypothesis being true after a "significant outcome" (H_{Toxic} = True|Significant Outcome) is called the Positive Predictive Value (PPV). *A priori* error rates (α and β) and effect size (ES) must be specified in order to estimate PPV from a significant NHST outcome. Also needed in many calculations of PPV is an estimate of *a prior* probability. The NHST-based NOEC approach only sets the least important error rate (α) *a priori* and uses experimental design to place vague and unquantified limits on the most important error rate (β). By default, the ES is 0. This nil ES is supported by custom only, not best professional judgment. In reality, any two populations will be judged different (ES=0) if enough samples are taken from them. The NHST-based NOEC served the needs of early 1970s regulators but should be replaced by logically defensible methods.

In contrast to the conventional NHST-based NOEC approach, the described TST is an equivalence-based hypothesis test that is conducted correctly. Emphasis is on the most important error rate and professional judgment is used to establish an ES before the test begins. There is simply no ambiguity that the TST is superior to the invalid NOEC method.

There is also no hesitancy about whether the proposed TST or the point estimation method is best for the described WET purposes. The equivalence testing-based TST has advantages in certain instances and the estimation method has advantages in others. The TST is clearly superior to point estimation *for the stated purposes* in the draft document. The value of point estimation methods comes into play if the issue progresses from the question of "is it toxic?" to one of "how toxic is it?" The estimation methods would facilitate assessment of how much temporal variability in toxicity is present in discharge or storm water, and also conduct of a formal TRE.

Response: Comment noted.

<u>4. Utility of the proposed accelerated monitoring schedule.</u> I have little experience with this aspect of the regulatory process. However, the schedule appears to be reasonable and consistent with other regulations.

Response: Comment noted.

If the associated data are too variable to produce an adequate estimate from a model, the simultaneous confidence interval method of Delignette-Muller et al. (2011) might be a useful approach instead of the conventional (invalid) NOEC methods.

Response: Comment noted.

Also meta-analysis of the results from the sequence of four tests might provide additional insight and facilitate better estimation of toxicity. The general approach can be illustrated using the forest plots described by Borenstein et al. (2009), Cumming (2012) and many others. A fabricated set of ECx values and a fixed effect model are used as an example here. Assume that a sequence of four ECx values are obtained:

ECx	Variance of ECx (from regression)
2.5	1.0
1.0	1.0
2.0	0.8
2.4	1.2

The estimated ECx (M) for the combined four tests can be calculated,

$$\mathbf{M} = \frac{\mathbf{W_i M_i}}{\mathbf{W_i}}$$

where M_i = the ECx for the ith of four tests. Weighting (W_i) for each ECx is based on its associated variance, i.e., $W_i = 1/V_i$. The variance of M (V_M) is estimated to be $1/(\Sigma W_i)$. The 95% CI for M in this straightforward example would be $[\mathbf{M} - \mathbf{1.96}\ \ \overline{\mathbf{V_M}}\ \mathbf{to}\ \mathbf{M} + \mathbf{1.96}\ \ \overline{\mathbf{V_M}}]$.

Various meta-analysis software packages do these and more difficult calculations, and produce easily interpreted forest plots. Two examples of forest plots are provided below in which the individual M_i results (left) or the sequential cumulative M (right) are plotted for the four tests.[2] The individual tests results are displayed in the left plot above the combined ECx estimate (diamond). The 95% confidence limits for the ECx estimate for the combined results are at the left and right tips of the diamond in both forest plots. The combined ECx and its confidence interval are 1.96 [0.989 to 2.929]. From the left plot, the confidence interval for the second test's ECx (M_2) overlaps 0; however, the confidence interval for the combined test results does not. A statistically significant effect would be judged to be present despite the results of the second test. The right cumulative plot indicates that the third and fourth tests were not required to get a combined ECx estimate that was acceptably precise and demonstrably different from 0.

[2] The example is based on data with symmetrical confidence intervals. This is often not the case for ECx estimates and additional computations would be needed in such cases.

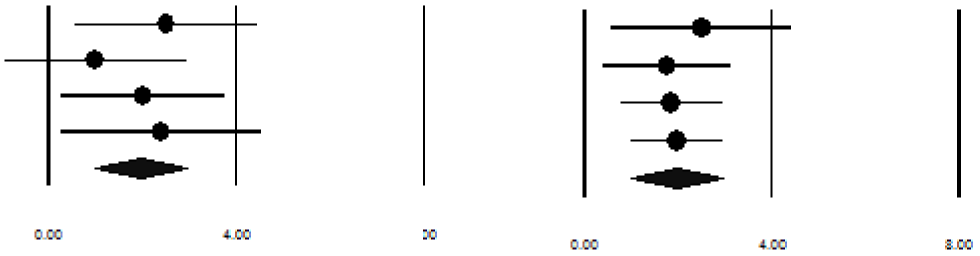

Figure 1. Illustration of meta-analysis for a series of ECx values using a fixed effects model and four fabricated ECx values and their variances. Horizontal bars are 95% confidence intervals. Results are presented for each individual test (left) and also cumulatively for the sequence of four tests (right). The topmost estimate and 95% CI are those for the first test and the results for the last (fourth) test are those immediately above each diamond. Calculations and plots were generated with Comprehensive Meta Analysis V 2.2.064 software (Info@meta-Analysis.com). Although p-values can be calculated, they are unnecessary after confidence intervals are generated and interpreted. A p-value for H_0: ECx>0 was recalculated as tests were added (right). P-values decreased (top to bottom) from 0.012, to 0.013, to 0.001, to <0.001.

The fabricated illustration provided above involves symmetrical confidence intervals, a condition not met for many ECx estimates. Different approaches can be used such as one involving transformation to logarithms or perhaps permitting a limited degree of deviation from symmetry (Figure 2). Still other approaches are possible, especially if more information than the ECx and its 95% confidence interval are available.

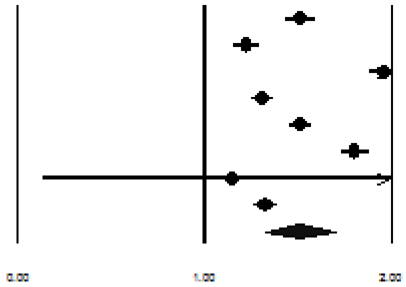

Figure 2. Simple illustration of Log_{10} transformation of LC50 and its confidence limit to generate approximately symmetrical limits. The 48h Mysid shrimp LC50 data are shown for eight oil dispersants (top to bottom): Corexit 9500A, Disperist SPC 1000, JD-2000, Nokomis 3-AA, Nokomis 3-F4, Saf-Ron Gold, Sea Brat 4, and ZI-400 (from Table 4, Hemmer et al. 2011). Estimates for one (Sea Brat 4) had a wide confidence interval so those results were given minimal weight in computations of the overall dispersant LC50. The composite Log_{10} LC50 of oil dispersants from the (random model) meta-analysis was 1.506 [1.319, 1.694] so the overall LC50 was 32.1 μl/L [20.8, 49.4].

This oil dispersant example was provided also to illustrate how one might assess the heterogeneity among LC50 values. The question could be asked during meta-analysis, “Is there substantial heterogeneity among these estimates or does the difference among estimates just reflect sampling error?” Comparison of the meta-analysis df to a Q statistic allows such questions to be tested (Cumming 2012). In this example, Q = 326.1 and df = 7. The difference between Q and df is significant (p<<0.05), indicating

substantial heterogeneity among dispersants. This same approach might be useful for assessing heterogeneity in the proposed accelerated monitoring schedule results estimated for four samples taken through time.

Response: Staff appreciates this recommendation for future consideration, as this reviewer states that all such suggestions are intended for “[…] future evaluation, not to identify flaws in the draft TST approach.”

<u>5. The Big Picture.</u> The draft TST is a major advance that is acceptable as described: no changes are needed. All of the suggestions above and here are put forward as potential issues for future exploration.

Nonparametric Option

Philip Dixon (1998) describes a nonparametric (ratio-based) equivalence test that begins first by modifying one group and then doing a conventional nonparametric Wilcoxon test. He reasons that this approach of modifying one group first could be a problem with a parametric test because variances would likely not be equal after modification, but seems acceptable with a rank-based method. For a two-sided test, the first of the two subhypotheses (e.g., a ±25% equivalence region) is the following: $\boldsymbol{\mu_A \leq 0.75\mu_B}$. It is tested by multiplying all control values by 0.75 and doing a Wilcoxon rank sum test of equality. A large z score for a one-tailed test would produce a small (perhaps significant) p-value. The second subhypothesis is $\boldsymbol{\mu_A \geq 1.25\mu_B}$. A one-tailed Wilcoxon rank sum test would be done again but after the control values are multiplied by 1.25 this time. The hypothesis of nonequivalence is rejected if the subhypotheses are rejected. Perhaps such an approach could be modified in the future for the TST method in cases when assumptions of normality cannot be met despite transformations.

Response: Staff appreciates this recommendation for future consideration, as this reviewer states that all such suggestions are intended for “[…] future evaluation, not to identify flaws in the draft TST approach.”

Bayesian Context

Philip Dixon (1998) also discusses the Bayesian context for equivalence testing that might be useful to consider in future versions of the TST approach. The Bayesian context is quickly becoming the dominant one in applied statistics so it might be advantageous to begin now to gradually develop a Bayesian vantage for the TST approach. The Bayesian context seems especially useful for the decision-making being considered here.

A frequentist approach such as that of the present TST assumes a fixed, real quality such as the difference between two population means. P-values and confidence intervals are then generated with random observations from the two populations (i.e., control and IWC sample). In contrast, a Bayesian approach assumes a random difference between the two means. The observations are considered fixed instead. A prior distribution for the difference between means might be available but, more likely, a noninformative prior would be assumed. Observations are collected to update this prior that all differences are equally probable and to produce a random distribution. The posterior is used to estimate the upper and lower bounds of the 95% highest posterior density interval for the random difference. In this context, the true difference is believed to be contained within the upper and lower bounds with a 95% probability. Mandallaz and Mau (1981) discuss the Bayesian vantage to equivalence testing. Their discussion also

involves confidence intervals. Other authors who discuss the general advantage of a Bayesian decision-making framework are Wolfson et al. (1996) and Ellison (1996).

Response: Comment noted. Staff appreciates the suggestion

The draft policy is based on sound science, methods and practices. It is a substantial improvement relative to current methods. The State Water Board should be proud of this advance and, hopefully, it will serve as an example for other regulatory groups desiring to move beyond the NOEC.

Response: Comment noted.

Application of professional judgment in the presence of uncertainty is the rule, not the exception. All methods depend on professional judgment although comfort with custom might make this fact less obvious for established methods. For example, the NOEC is based on judgments about α (0.05 by custom), β (generally dictated by standard design) and effect size (0 by custom). (These professional judgments are now understood to be inadequate for environmental decision-making.)

Response: Comment noted.

The real question is whether a specific applied set of professional judgments is sound, open for scrutiny, and pragmatic. The TST procedures described in the draft document meet all of these criteria for good professional judgment.

Response: Staff agrees with this statement.

The Preamble to the Draft Policy states that "This policy shall be reevaluated by the State Water Resources Control Board (State Water Board) five years from its effective date." All of the suggestions detailed above are intended to facilitate such future reevaluation, not to identify flaws in the draft TST approach. The TST approach and associated details are sound as presently proposed and better meet the goals of the Water Board than current NOEC-based methods. Likely, they will involve fewer test individuals and lower costs. They also seem more useful than point estimation methods for the intended purposes.

Response: Staff agrees with this statement.

In the future, TST staff might wish to explore several of the issues mentioned above. (1) The nonparametric equivalence approach described by Dixon (1998) might provide a way ahead if transformations fail to produce normal data. (2) The Bayesian context often is no more complicated than the frequentist context but inferences tend to be easier to make and are more consistent with evidence-based decision-making. (3) Relative to the proposed accelerated monitoring schedule, the meta-analysis approach and Delignette-Muller et al.'s simultaneous confidence interval alternative to conventional Dunnett's testing might be useful to consider.

Response: Staff appreciates this recommendation for future consideration, as this reviewer states that all such suggestions are intended for "[…] future evaluation, not to identify flaws in the draft TST approach."