PEER REVIEW – TEST OF SIGNIFICANT TOXICITY (TST)

Michael C. Newman, A, Marshall Acuff Jr Professor
College of Willian & Mary Virginia Institute of Marine Science

Findings to be Addressed by Peer Reviewers

The statutory mandate for external scientific peer review (Health and Safety Code Section 57004) states that the reviewer's responsibility is to determine whether the scientific portion of the proposed rule is based upon sound scientific knowledge, methods, and practices. State Water Resources Control Board (State Water Board) staff requests that you make this determination for each of the following topics that constitute the scientific portion of the proposed standard.

GENERAL RESPONSE

All of the suggestions detailed herein are intended to stimulate exploration of future enhancements of the TST approach, not to identify flaws in the draft TST approach. The TST approach and associated details are sound as proposed and better meet the goals of the Water Board than current methods.

1. ***Inclusion of the Test of Significant Toxicity in the draft Policy for Toxicity Assessment and Control: The U.S. Environmental Protection Agency (U.S. EPA) developed the Test of Significant Toxicity (TST) to address concerns regarding the use of the current hypothesis test method, the No Observed Effect Concentration (NOEC). Derived from the bioequivalence approach used by the Food and Drug Administration and countless researchers, this modified hypothesis test requires the use of only two dilutions: the in-stream waste concentration (IWC) and a laboratory control. Unlike the NOEC approach, the TST incorporates percent-based effect thresholds, (b values) that define unacceptable levels of chronic or acute toxicity in an IWC sample. In addition, the TST utilizes a restated null hypothesis that assumes toxicity, thereby placing the responsibility of proving otherwise on the discharger. Most importantly, however, the TST establishes a false negative error rate that is absent from the current NOEC approach (see the draft Staff Report and the TST Implementation document for additional information about the TST).***

   Evaluate the conceptual soundness of the draft Policy for Toxicity Assessment and Control's (draft Policy) null hypothesis-based objectives, and the provision requiring applicable dischargers to use the TST for all toxicity data analyses. Consider the TST as a means of compliance determination for NPDES wastewater and point source WDR dischargers, as well as the analytical function it will serve for storm water and channelized dischargers.

RESPONSE

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.

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*.

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

[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.

the TST approach. Perhaps such meta-analyses could eventually be used as quality control charts are used now for chemical analyses?

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.

2. ***Use of effect level reporting in compliance determination: The TST was designed in a manner that prevents an IWC sample from being incorrectly declared to be "toxic" no more than five percent of the time whenever the effect level of the test organisms is at or below 10 percent. As such, the sensitivity of a test's design or the influence of within-test variability can occasionally produce a result of "fail" (i.e. toxicity detection) when the effect level of the sample is above 10 percent, but below the regulatory management decisions (RMD) of 0.75 for chronic, and 0.80 for acute (i.e. an effect level of 0.25 for chronic, and 0.20 for acute). In most cases, these discrepancies can be mitigated by retesting with additional replicates (see the TST Test Drive for examples).***

   To address this issue, State Water Board staff has included both maximum daily effluent limitations (MDEL) and average monthly effluent limitations (AMEL) in the revised draft Policy. The proposed MDELs are set at effect levels equivalent to double the RMDs for acute and chronic toxicity (0.40 and 0.50 respectively). Discharge samples that "fail" below an MDEL will be directed to conduct two additional toxicity tests in order to determine compliance with the AMEL. If either of these subsequent toxicity tests results in a "fail," the discharger will be in exceedance of the AMEL and required to implement an accelerated monitoring schedule. Assess the effectiveness of these effluent limitations in reducing discrepancies that arise from toxicity detections below the RMDs. Determine whether or not the proposed AMEL is better suited to reduce these discrepancies than that of a monthly limitation set at the RMDs and measured by the average effect level of three toxicity tests.

   RESPONSE
   These steps seem reasonable and clear. The proposed AMEL seems better suited, in my opinion, than the alternative.

   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.

3. ***Comparative approaches to toxicity analyses: As previously explained, State Water Board staff believes the TST to be an improvement over the hypothesis testing approach currently used for toxicity monitoring. Unlike the NOEC approach, the TST rewards laboratory precision, accounts for false negatives, and incorporates an effect threshold that clearly specifies the level of biological impact. Additionally, the two-concentration test design of the TST costs approximately 50 percent less than the five-concentration tests required by the NOEC.***

   Staff also believes the TST to be more appropriate for the draft Policy than point estimate approaches such as the Spearman-Karber method or Probit. While point estimates offer some benefits over hypothesis testing, such as the ability to interpolate effect levels and utilize non monotonic data, staff feels that this statistical approach will introduce unnecessary complications to toxicity data analyses. For example, bias can be introduced through ill-fitting models and data smoothing techniques, while the Spearman-Karber, Trimmed Spearman-Karber, and Graphical methods are incapable of calculating endpoints below a 50% effect level (see the draft Staff Report for additional information).

   Assess the efficacy of the TST in light of the NOEC and point estimate approaches. Consider the benefits and drawbacks of the three approaches when applied to routine monitoring, accelerated monitoring, and Toxicity Reduction Evaluations (TRE).

   RESPONSE
   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., $p(\text{Results}|H_0=\text{True})$. It is not the probability that the null hypothesis is true given the results of the test, that is, not $p(H_0=\text{True}|\text{Results})$. This can be easily shown with the Bayes-LePlace Theorem,

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

   Clearly, $p(\text{Results}|H_0=\text{True})$ is not $p(H_0=\text{True}|\text{Results})$. 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.

4. ***Utility of the proposed accelerated monitoring schedule: The draft Policy proposes the implementation of an accelerated monitoring schedule for dischargers that exceed the chronic or acute effluent limitations. This schedule, adopted from U.S. EPA's Toxicity Training Tool guidebook, would consist of four, five-concentration toxicity tests, conducted at approximately two-week intervals, over an eight-week period (see the draft Staff Report for additional information). The use of five concentrations during accelerated monitoring serves to satisfy the federally-required test conditions that are incorporated by reference in 40 Code of Federal Regulations section 136.3. In addition, staff is of the opinion that multiple- concentration analyses can***

prove beneficial to dischargers that are required to conduct a TRE after an exceedance occurs during accelerated monitoring.

Evaluate the appropriateness of this accelerated monitoring schedule. Consider its effectiveness in characterizing effluent magnitude and the individual probability of declaring a sample as a "fail" below the proposed effluent limitations.

RESPONSE
I have little experience with this aspect of the regulatory process. However, the schedule appears to be reasonable and consistent with other regulations.

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.

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,

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

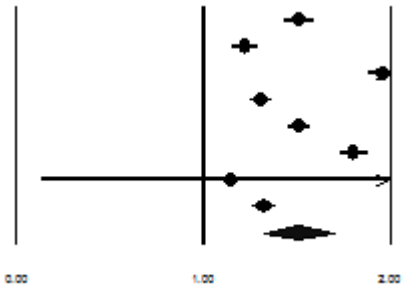
where M_i = the ECx for the i^{th} of four tests. Weighting (W_i) for each ECx is based on its associated variance, i.e., $W_i = 1/V_i$. The variance of M (V_M) is estimated to be $1/(\Sigma W_i)$. The 95% CI for M in this straightforward example would be $[M - 1.96\ \overline{V_M}\ to\ M + 1.96\ \overline{V_M}]$.

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.

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.



[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 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.

5. The big picture: Reviewers are not limited to addressing only the specific topics presented above, and are asked to contemplate the following questions as well:

a. In reading the draft Policy and Staff Report, are there any additional scientific topics that are part of the scientific basis of the proposed standard not described above? If so, please comment.

RESPONSE

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.

<u>Nonparametric Option</u>

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: $\mu_A \leq 0.75\mu_B$. 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 $\mu_A \geq 1.25\mu_B$. 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.

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).

b. Taken as a whole, is the scientific portion of the draft Policy based upon sound scientific knowledge, methods, and practices?

RESPONSE

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.

c. Reviewers should also note that some proposed provisions may rely significantly on professional judgment where available scientific data are not as extensive as desired to support the statute requirement for absolute scientific rigor. In these situations, the proposed course of action is favored over no action.

RESPONSE

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.)

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.

d. The preceding guidance will ensure that reviewers have an opportunity to comment on all aspects of the scientific basis of the draft Policy. At the same time, reviewers should also recognize that the State Water Board has a legal obligation to consider and respond to all feedback on the scientific portions of the draft Policy. Because of this obligation, reviewers are encouraged to focus feedback on the scientific topics that are relevant to the central regulatory elements being proposed.

RESPONSE

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.

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.

REFERENCES

Altman M. 2004. Statistical significance, path dependency, and the culture of journal publication. *J Socio-Econ* 33:651-663.

Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J Wildlife Manage* 64(4):912-923.

Borenstein M, LV Hedges, JPT Higgins, HR Rothstein. 2009. *Introduction to Meta-Analysis*, John Wiley and Sons, Chichester, UK.
Chapman PM, Caldwell RS, Chapman PF. 1996. A warning: NOECs are inappropriate for regulatory use. *Environ Toxicol Chem* 15(2):77-79.
Chow S-C, Liu JP. 1997. Meta-analysis for bioequivalence review. *J Biopharm Stat* 7(1):97-111.
Chow S-C, Liu J-P. 2000. *Design and Analysis of Bioavailability and Bioequivalence Studies, 2nd Edition*. Marcel Dekker, Inc., New York, NY.
Cumming, G. 2012. *Understanding the New Statistics*. Routledge, New York, NY.
Delignette-Muller M-L., C Forefait, E Billoir, S. Charles. 2011. A new perspective on the Dunnett procedure: filling the gap between NOEC/LOEC and ECx concepts. *Env Toxicol Chem* 30(12):2888-2891.
Denton DL, J Diamond, L Zheng. 2011. Test of significant toxicity: a statistical application for assessing whether an effluent or site is truly toxic. *Env Toxicol Chem* 30(5):1117-1126.
Diamond J, D Denton, B Anderson, B Phillips. 2011. It is time for changes in the analysis of whole effluent toxicity data. *IEAM* DOI: 10.1002/ieam.278.
Dixon PM. 1998. Assessing effect and no effect with equivalence tests. In: Newman MC, CL Strojan (eds). *Risk Assessment: Logic and Measurement*, CRC/Ann Arbor Press, Boca Raton, FL, pp. 275-301.
Ellison AM. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol App* 6(4):1036-1046.
Erickson WP, LL McDonald. 1995. Test for bioequivalence of control media and test media in studies of toxicity. *Env Toxicol Chem* 14(7):1247-1256.
Fidler F, Cumming G, Burgman M, Thomason N. 2004. Statistical reform in medicine, psychology and ecology. *J Socio-Econ* 33:615-630.
Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv Biol* 20(5):1539-1544.
Gigerenzer G. 2004. Mindless statistics. *J Socio-Econ* 33:399-417.
Hoekstra JA, van Ewijk PH. 1993. The bounded effect concentration as an alternative to the NOEC. *Sci Tot Environ, Suppl* 1993: 705-711.
Ioannidis JPA. 2005. Why most published research findings are false. *PLoS Med* 2(8):e124.
Hemmer MJ, Barron MG, Greene RM. 2011. Comparative toxicity of eight oil disperants, Louisiana sweet crude oil (LSC), and chemically dispersed LSC to two aquatic test species. *Environ Toxicol Chem* 30(10):2244-252.
Hurlbert SH, Lombard CM. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann Zool Fennici* 46:311-349
Kooijman SALM. 1996. An alternative for NOEC exists, but the standard model has to be abandoned first. *Oikos* 75(2):310-316.
Kruschke JK. 2011. *Doing Bayesian Data Analysis.* Elsevier, Boston, MA.
Jager T. 2012. Bad habits die hard: the NOEC's persistence reflects poorly on ecotoxicology. *Environ Toxicol Chem* DOI 10.1002/etc.746.
Laskowski R. 1985. Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos* 73(1):140-144.

Lecoutre B, Lecoutre M-P, Poitevineau J. 2001. Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable? *Int Stat Rev* 69:399-417.
Lopez RA, Neves FAR. 2010. Meta-analysis for bioequivalence studies: interchangeability of generic drugs and similar containing hydrochlorothiazide is possible but not for those with enalapril maleate. *J Bras Nefrol* 32(2):173-181.
Mandallaz D, Mau J. 1981. Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* 37:213-222.
McCloskey DN. 1995. The insignificance of statistical significance. *Am Sci* 272:32-33.
Min C-K, Zellner A. 1993. Bayesian and non-Bayesian methods of combining models and forecasts with applications to forecasting international growth rates*. J Economet* 56(1-2):89-118.
Newman MC. 2008. "What exactly are you inferring?" A closer look at hypothesis testing. *Environ Toxicol Chem* 27(5):1013-1019
Shukla R, Q Wang, F Fulk, C Deng, D Denton. 2000. Bioequivalence approach for whole effluent toxicity testing. *Env Toxicol Chem* 19(1):169-174.
St J Warne M, Van Dam R. 2008. NOEC and LOEC data should no longer be generated or used. *Aust J Ecotox* 14:1-5.
Sterne JAC, Davey Smith G. 2001. Sifting the evidence – what's wrong with significance tests? *BMJ* 322:226-230.
Sutton AJ. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 10(4):277-303.
Trafimow D, Rice S. 2009. A test of the null hypothesis significance testing procedure correlation argument. *J Gen Psychol* 136(3):261-269.
Van der Hoeven N. 1997. How to measure no effect. III. Statistical aspects of NOEC, ECx and NEC estimates. *Environmetrics* 8:255-261.
Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 96(6):434-442.
Welch BL. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29:350-362.
Wolfson LJ, Kadane JB, Small MJ. 1996. Bayesian environmental policy decisions: two case studies. *Ecol App* 6(4):1056-1966.
Ziliak ST, McCloskey DN. 2004. Significance redux*. J Socio-Econ* 33:665-675.