SUMMARY REPORT

PEER REVIEW OF "PRELIMINARY REPORT: INTERLABORATORY VARIABILITY STUDY OF EPA SHORT-TERM CHRONIC AND ACUTE WHOLE EFFLUENT TOXICITY TEST METHODS" (WET STUDY REPORT)

Prepared for:

U.S. Environmental Protection Agency
Office of Water
Office of Science and Technology
Health and Ecological Criteria Division
1200 Pennsylvania Ave., NW
Washington, D.C. 20460

Prepared by:

Versar, Inc.
6850 Versar Center
Springfield, Virginia 22151

March 2001

TABLE OF CONTENTS

1.0 INTRODUCTION

In January 1993, responding to recommendations in the report, "Safeguarding the Future: Credible Science, Credible Decisions," Administrator William Reilly issued an Agency-wide policy for peer review. Administrator Carol Browner confirmed and reissued the policy on June 7, 1994 and instituted an Agency-wide implementation program. The principle underlying the Peer Review Policy is that all major scientific and technical work products should be peer reviewed.

Peer review is a process for enhancing a scientific or technical work document so that the decision or position taken by the Agency, based on the technical document, has a sound, credible basis. The goal of the Agency's Peer Review Policy is to ensure that scientific and technical work products receive appropriate levels of critical scrutiny from scientific and technical experts as part of the overall decision making process. Generally, this technical review precedes the customary, more broadly based public review of the total decision.

The EPA Whole Effluent Toxicity (WET) Interlaboratory Variability Study was conducted to meet requirements of a July 1998 settlement agreement between EPA and Edison Electric Institute, Western Coalition of Arid States, *et al.* This settlement agreement resolved a judicial challenge to the 1995 rule that approved WET test methods for use in National Pollutant Discharge Elimination System (NPDES) monitoring. In addition to requiring that EPA conduct the WET Interlaboratory Variability Study, the settlement agreement required the peer review of study results.

The Engineering and Analysis Division (EAD) within EPA's Office of Water was responsible for conducting the WET Study and producing the study report. EAD's core responsibility within the Office of Water is the development of effluent limitations guidelines and standards under the Clean Water Act. EAD also provides analytical services, via EPA contract laboratories, to support effluent limitation guideline development. Due to this experience in analytical testing, EAD was tasked with the responsibility of the WET Interlaboratory Variability Study.

The purpose of this peer review was to evaluate the scientific credibility of the work product titled, "Preliminary Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods, Volume 1 & Volume 2: Appendix" (the WET Study Report). This summary reports presents the peer review comments of three expert reviewers.

1.1 Peer Reviewers

This draft document was reviewed by a panel of three expert peer reviewers: X, Y, and Z. These panelist were selected because of their expertise in various areas relevant to this document. The reviewers have the educational background and the experience (combined total of 50+ years) to provide a review of significant quality. They have a combination of skills that include: aquatic toxicology, whole effluent toxicity testing, statistics, laboratory toxicity testing, ecotoxicology, monitoring study design, methods development, environmental quality criteria development, and have authored numerous publications in peer-reviewed environmental journals.

1.2 Peer Review Comment Format

The comments and recommendations from all three reviewers have been combined and organized as follows:

- General comments;
- Charge to the reviewer;
- Specific comments by document page number referenced by reviewer;
- Miscellaneous comments; and
- Additional references recommended by the reviewers.

2.0 CHARGE TO THE PEER REVIEWERS

Background

EPA's WET testing policies and regulations are intended to support goals of the Clean Water Act (CWA), specifically, to provide for the protection and propagation of fish, shellfish, and wildlife. CWA section 101(a)(3) states, "it is the national policy that the discharges of toxic pollutants in toxic amounts be prohibited." EPA has pursued this objective through the water quality standards program and the National Pollutant Discharge Elimination System (NPDES) permitting program. These programs have adopted an integrated strategy of water quality-based toxics control that includes chemical-specific controls, whole effluent toxicity controls, and biological criteria and biosurvey approaches. When technology-based permit limits are insufficient to achieve water quality standards, Federal regulations at 40 CFR 122.44 (d)(1) require the permitting authority to include water quality-based effluent limits in the NPDES permit. EPA's adoption of water quality-based permitting that integrates chemical and biological monitoring to protect receiving water quality was a major step forward for toxics control.

NPDES permit limits for WET are required (40 CFR 122.44(d)(1)(iv)) if a discharge causes, or has a reasonable potential to cause or contribute to an instream excursion above a numeric water quality criterion for WET (i.e., reasonable potential). NPDES permit limits for WET also are required 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), unless the permitting authority has identified the parameters causing toxicity and placed limits in the permit to control those parameters appropriately (40 CFR 122.44(d)(1)(v)).

Monitoring of compliance with WET permit limits is accomplished by the routine testing of permitted discharges using EPA-approved WET test methods (40 CFR 136). Acute and short-term chronic WET tests estimate the toxicity of wastewaters to protect aquatic life. These tests measure the aggregate toxic effects of effluents or receiving waters to standardized, freshwater or marine plants, vertebrates, or invertebrates. The result of a single WET test could be used to assess compliance with a permit limit for WET (expressed in terms of acute or chronic toxic units). For example, an end-of-pipe permit limit of 0.3 TU_a may be established when the State water quality standards have no mixing zone allowance, or an end-of-pipe permit limit of 1 TU_c may be established when there is little or no dilution instream.

Objectives of the WET Study

The three primary goals of the WET Study were to:

- Estimate interlaboratory precision for each method in the WET Study
- Provide data on the rate at which participating laboratories successfully completed tests initiated (successful test completion rate)
- Provide data on the rate at which the tests indicate the presence of toxicity when measuring non-toxic samples (false positive rate)

The term "interlaboratory precision" means the degree of mutual agreement among WET test results from various laboratories testing the same sample. EPA estimated interlaboratory precision for the WET test methods by calculating a coefficient of variation (CV) for LC50 and IC25 endpoints. Interlaboratory precision of NOEC endpoints was described by the range and distribution frequency of NOEC values. EPA determined the successful test completion rate as the percentage of initiated and properly terminated tests that meet the test acceptability criteria as specified in the WET method manuals. The false positive rate was

determined as the percentage of successfully completed tests that indicated toxicity was present in reagent water samples, known as blanks.

WET Study Overview

The WET Study was designed to quantify the interlaboratory precision of 12 of the 17 promulgated WET test methods. The five acute and seven short-term chronic WET methods listed below were performed by participant laboratories in accordance with the promulgated test procedures (USEPA, 1993; USEPA, 1994a; USEPA, 1994b).

- Method 1000: *Pimephales promelas* (fathead minnow) Larval Survival and Growth Test
- Method 1002: *Ceriodaphnia dubia* (cladoceran) Survival and Reproduction Test
- Method 1003: *Selenastrum capricornutum* (green alga) Growth Test (with and without EDTA)
- Method 1004: *Cyprinodon variegatus* (sheepshead minnow), Larval Survival and Growth Test
- Method 1006: *Menidia beryllina* (inland silverside), Larval Survival and Growth Test
- Method 1007: *Mysidopsis bahia* (mysid shrimp), Survival, Growth, and Fecundity Test
- Method 1009: *Champia parvula* (red macroalga), Reproduction Test
- *Ceriodaphnia dubia* Acute Test
- *Pimephales promelas* Acute Test
- *Cyprinodon variegatus* Acute Test
- *Menidia beryllina* Acute Test
- *Holmesimysis costata* Acute Test (using test procedures to measure acute toxicity on *Mysidopsis bahia*)

In designing the WET Study, EPA consulted American Society of Testing Materials guidance for determining the precision of a test method (ASTM, 1997). The "base" WET Study design consisted of a minimum of 9 and a maximum of 20 participant laboratories. Additional laboratories (above 20) participated on a more limited basis as part of an "extended" WET Study design. The laboratories selected for participation typified laboratories that routinely conduct WET testing for permittees. Prior to selection, laboratories were required to submit data to show that they met pre-qualification requirements, which included: historical records of acceptable control charts using reference toxicants, documentation of laboratory standard operating procedures (SOPs) for test conduct and quality control, documented experience in conducting tests and successfully meeting test acceptability criteria, and documentation of healthy and reliable test organism cultures and/or sources.

For each WET test method, a referee laboratory collected, prepared and shipped blind test samples to participant laboratories for testing. Laboratories participating in the base study design received four blind test samples that included some combination of blank samples, reference toxicant samples, receiving water samples, and industrial and/or municipal wastewater effluent samples. Laboratories participating in the extended study design received three blind test samples that included some combination of blank samples and reference toxicant test samples. Data generated by all qualified participating laboratories were independently reviewed, compiled in a study database, and statistically analyzed to determine interlaboratory precision, successful test completion rates, and false positive rates for each test method.

Peer Reviewer Instructions

The primary function of the peer reviewer is to judge whether the WET Study Report clearly and accurately assesses method precision, successful test completion rates, and false positive rates for the WET test methods evaluated, and whether the results are scientifically acceptable within the context of the intended regulatory use. Peer reviewers are not asked to comment on the design of the WET Study, only the implementation of

the study plan and the results of the study. Peer review of the WET Study plan was conducted in October 1998 through November 1998 according to EPA peer review policy. EPA considered the peer review comments and published a final version of the WET Study plan on June 11, 1999.

To enhance the quality and appropriateness of the preliminary WET Study report, peer reviewers are requested to address the three aspects of the report listed below.

1) Document's Responsiveness: Peer reviewers are asked to comment on the responsiveness of the document in meeting the requirements of the WET settlement agreement (see section on additional and supporting information) and following the WET Study plan (Appendix A of the WET Study Report). More specifically, reviewers are asked to address the following questions:

   - Is the WET Study Report responsive in evaluating WET test method precision, successful test completion rates, and false positive rates?
   - Does the WET Study Report provide reference to the data and equations used to calculate the coefficient of variation (or other applicable estimate of precision) for each test endpoint?
   - Does the WET Study Report provide a chart of any new coefficients of variation for the various WET methods derived from the results of the Interlaboratory Variability Studies?

2) Document's Scientific Merit: Peer reviewers are asked to comment on the scientific soundness of approaches used to implement the study and review, analyze, and report results. More specifically, reviewers are asked to address the following questions:

   - Was the study plan appropriately implemented to achieve the objectives of determining method precision, successful test completion rates, and false positive rates?
   - Was the inclusion/exclusion of data appropriate for the determination of precision?
   - Was the method for calculating precision scientifically acceptable?
   - Are the results scientifically acceptable within the context of the intended regulatory use? (See the background for determining the context of the regulatory use.)

3) Document Quality and Clarity: Peer reviewers are asked to critique the document for its clarity given its potential scientific and technical applications. More specifically, reviewers are asked to address the following questions:

   - Is the WET Study Report clear in stating the objectives, describing study implementation, and presenting study results?
   - Are data summarized and presented in a technically clear and appropriate manner?

3.0 COMMENTS

3.1 General Comments

Reviewer X

The study was comprehensive and credible. There were a number of mistakes made, but that is not unexpected in a study of this size and complexity and reasonable remedies were applied. Overall, I was impressed by the level of detail and effort involved in this work though I have a number of comments. Whether it actually answered the primary issues it was supposed to is another question, and one that touches on the study design which is not subject to this particular review. Unfortunately, though the WET Study Plan was provided, a copy of the WET Settlement Agreement was not provided, thus the relevance of my comments to that Agreement is uncertain. My comments regarding "bigger picture" issues such as the relevance and implications of WET tests are provided in Section 5. of this review.

One thing that does surprise me is that the testing was apparently all "nominal" rather than actual. I would have expected that concentrations of the reference toxicants and the spiked toxicants would have been determined in at least some (preferably all) laboratories rather than being inferred. I see this as a failure of the Study Design.

It is not clear why KCl was used for freshwater and Cu for marine waters, particularly when there was a switch to KCl for some of the marine tests after Cu started to precipitate out. It is also not clear why a dilution factor of 0.5 was stipulated to set the test concentrations since this is much greater than would normally occur in "real life".

Several things seem very clear from the report. For instance: (1) the *Holmesimysis* test is not a good one for routine usage; (2) *Selenastrum* testing should always be done with EDTA other than in investigative, not regulatory, usage; (3) the Mysid chronic test is suspect.

Reviewer Y

Versar, Inc guidance defines the purpose of this review as enhancing "the quality and credibility of Agency decisions by ensuring that the scientific and technical work products underlying EPA's decisions receive appropriate levels of peer review by independent scientific and technical experts. To this end, the reviewers should judge whether the WET Study Report clearly and accurately assesses method precision, successful test completion rates, and false positive rates for the WET test methods, and whether the results are scientifically acceptable within the context of the intended regulatory use. Relative to their intended use, monitoring with WET methods within the NPDES permitting process supports the Clean Water Act (CWA). The CWA aims to protect and allow propagation of fish, shellfish, and wildlife by prohibiting the discharge of toxicants in toxic amounts into receiving waters.

This WET study stems from a specific litigation settlement in which the EPA agreed to assess three aspects of 12 of 17 existing WET methods. Based on the results, the EPA may ratify or withdraw the various methods. Participating laboratories were selected to represent those that would routinely use WET methods. Much effort was made to ensure that at least nine and maximally 20 laboratories produced results for each test method and associated effect metric. However, only seven laboratories participated in the sheepshead minnow acute and chronic tests. Interlaboratory evaluations were not possible for two (*Champia* growth and *Holmesimysis* acute tests) of the 12 WET methods because insufficient numbers of laboratories participated. *Holmesimysis* test results were considered preliminary because animals were collected from the field (Page 45). Also there was deviation from the SOP (Page 45), precipitation of the copper spike (Page 48), and high

control mortality (Page 45). The tests evaluated in the Study were the following (species, test type, effect metric):

Cladoceran (*Ceriodaphnia dubia*), acute lethality(LC50)
Cladoceran (*C. dubia*), survival (LC50, NOEC) and reproduction (IC25, NOEC)
Fathead minnow (*Pimephales promelas*) acute lethality (LC50)
Fathead minnow (*P. promelas*) larval survival (LC50, NOEC) and growth (IC25, NOEC)
Green alga (*Selenastrum caprocornutum*) growth (IC25, IC50, NOEC)
Mysid (*Mysidopsis bahia*) survival (LC50,NOEC), growth and fecundity (IC25, NOEC)
Sheepshead minnow (*Cyprinodon variegatus*) acute lethality (LC50)
Sheepshead minnow (*C. variegatus*) larval survival (LC50,NOEC) and growth (IC25,NOEC)
Inland silverside (*Menidia beryllina*) acute lethality (LC50)
Inland silverside (*M. beryllina*) larval survival (LC50, NOEC) and growth (IC25, NOEC)
Red macroalga (*Champia parvula*) reproduction test (IC25, NOEC)
Mysid (*Holmesimysis costata*) acute lethality (LC50 using *Mysidopsis bahai* methods))

Laboratory selection was done with the intent of getting results representative of those generated by laboratories routinely conducting WET methods. "Participant laboratories must represent a cross-section of the laboratory community qualified to conduct WET tests using proper test procedures and QA/QC provisions detailed in the method manuals" (Page 5 of Vol. 2, Appendix). Laboratories included EPA, university, municipal, state, industrial, and private laboratories. A subset of laboratories was selected by EPA using formal criteria. Other laboratories not selected by EPA could and did participate in the interlaboratory testing after they also demonstrated conformity to these criteria. Selection included not only representation as described above, but also selection of laboratories "that possess the capacity and capabilities, experience and proficiency, and quality assurance and quality control necessary to meet the needs of the study" (Pages 6-7 of Vol. 2, Appendix). All had to be capable of responding to the defined schedule of sample mailing, testing, and reporting. The nine low-bid, qualified laboratories were selected by the EPA for the corresponding WET method with 11 more laboratories being randomly selected from a pool of non-EPA sponsored laboratories.

Three qualities of the report were to be addressed in the review process: (1) the report's responsiveness to the requirements of the WET settlement agreement, (2) the report's scientific merit, i.e., scientific and technical merit relative to the intended regulatory use, and (3) the report's quality and clarity. The products of the study to be judged relative to these qualities were the following:

· Estimate interlaboratory precision of each method (listed above) in the WET Study. For the LC50 and IC25 metrics, the coefficient of variation (CV = 100[SD/Mean]) was used to assess precision of the testing laboratories. Range and frequency distribution were used to reflect precision for the NOEC which, by definition, can only take on a discrete value of one of the concentrations used in a specified sample dilution series. The CV for sets of endpoints were estimated with attention paid to observations below or above the range of the test methods. If endpoint values were determined to be greater than 100% or less than the lowest sample dilution, the reported ">100%" or "<Lowest Percentage" values were replaced with an actual percentage, e.g., ">100%" becomes 100% or "<6.25%" becomes 6.25% in calculations of CV. Precision was not estimated if censored values constituted 20% or more of the observations. Twenty percent or greater censoring existed for the blank sample results for all tests, the *Ceriodaphnia* chronic and silverside acute test methods applied to the reference toxicant sample, and all samples for the mysid chronic fecundity endpoint. If only

one estimate was available, that single value was reported. Analysis of variance was performed to determine intralaboratory ("within laboratory"), interlaboratory ("between laboratory"), and total variance of the WET methods. Outlier testing was done on results "to determine if any additional test results should be excluded from the analysis of precision" (page 92 of Vol. 1).

· Provide data on the rate at which participating laboratories successfully completed tests. The completion rate was expressed as a percentage of all participating laboratories after some adjustment of the tally. Adjustment was associated with the understanding that successful completion was to be estimated for "initiated and properly terminated tests that meet the test acceptability criteria as specified in the WET method manuals."

· Provide data on the rate of false positive tests, i.e., tests indicating the presence of toxicity when measuring non-toxic samples. Blank samples were used for this purpose and percentage of blank samples for which toxicity was reported (regardless of magnitude of the reported toxicity) was used as a metric of the false positive rate. This false positive error rate was estimated as the percentage of successfully completed tests that falsely indicated toxicity in the blank samples. This error rate was not compared to anticipated error rates at the conventional a employed in the NOEC-associated hypothesis tests (0.05).

Reviewer Z

This reviewer was impressed by the general clarity of the report, the lack of typographical or spelling errors, and the attention to detail throughout the document. This must have been an extremely difficult report to write, given all the data and all the summary tables. It was equally difficult to review. Considering that this is a preliminary report without much peer review, the document is surprisingly polished. On the other hand, the attention to detail and dedication to providing all the necessary detail in several instances appeared to dilute the main message. For example, the crux of the report is delivered in the very last section without any conclusions to provide a context for the results. It appears that EPA may wait until all the peer review comments are received and evaluated before drawing any conclusions, but it may have been helpful if the reviewers had been able to review EPA's intent at drawing conclusions. Having reviewed the conclusions, the peer reviewers may have been able to provide additional guidance. Given that the report was the result of a legal settlement and that any premature conclusions may have significant ramifications with respect to the lawsuit, the reviewer understands that EPA may have been reluctant to draw any conclusions whatsoever. Finally, the reviewer is extremely troubled by the results. The high CV values for several of the tests and the inability of several of the most competent labs in the country to successfully complete some of these tests does not does not provide much support for the continued use of those particular tests for their intended purpose. In fact, the results seem to support many of the claims promulgated by the lawsuit. Having said that, EPA clearly used credible and appropriate science in the evaluation of each test method for the WET testing program. The report clearly and accurately assesses interlaboratory method precision, successful test completion rates, and false positive rates for the WET test methods evaluated. The results are scientifically acceptable, but within the context of intended regulatory use, the documented variability in several tests may preclude their intended regulatory use.

This reviewer believes that the body of the report is too long and the Appendix is too short. The report would be easier to read and the main points easier to follow if several sections were edited to include only the most pertinent information in the body of the report and move the highly-detailed information that is not necessary for understanding the results into the Appendix. This could include sections such as how the laboratories were selected and how the samples were sent, preliminary testing, etc. With all of the detail that is presented, the crux of the message is lost. This problem is exacerbated by the absence of a conclusions section. The reader is left to imagine the environmental and the regulatory significance of the results in this study. As it

stands, the body of the report is 177 pages long and the Appendix approximately 80 pages long. As a target for editing, EPA might consider removing between 50 and 100 pages from the body of the report and shifting them to the Appendix. At a minimum, the report and the appendix should be about the same length or approximately 125 pages apiece. The report would be even easier to read and comprehend if as much as 100 pages of text and tables were removed from the body of the report and added to the appendix. Alternatively, EPA could choose to make a longer Executive Summary of 5 to 10 or even 25 pages to get across the main message. While all of these editorial messages seem somewhat drastic, the reviewer firmly believes that most of those who read this document in its present form will miss the forest through the trees; i.e., the main points will be lost in all the detail that is presented. It is necessary to include those details, but they do not have to be presented in the body of the report. Finally, the reviewer found Appendix A much easier to read than the body of the report and much of this Appendix was too repetitive. EPA should seriously consider exchanging the introductory material in the body of the report for the introductory material in Appendix A and only include the most detailed information in Appendix A. Appendix B seems acceptable in its present form.

This reviewer is also concerned that the report is less than it could have been because of the implied pressure or legal agreements to include or exclude certain concepts and certain details in this report. It appears that some of these issues have detracted from the overall flow of report by subtraction, or forced a particular approach because of specific pressures or agreements to include or exclude them. Examples are as follows: 1) Not including the data from the referee laboratories in the statistical results; 2) Including the minute details regarding any and all procedures; and 3) Excluding potentially useful information on statistical analyses of mean, min, max, etc from the referee laboratory testing. There is important information included in these referee laboratory results that were not evaluated in any way. Before categorically rejecting them, they should have been evaluated to determine if the trends were really any different from that of the participant laboratories. If the data from the referee laboratories were found to be significantly different from the participant laboratories they could be rejected at that point. If they were not different, the data could have been included with an asterisk and presented both ways. This would have been a more scientific and a more informative approach. The reviewer is concerned that there may have been other issues such as this that are not immediately obvious but may have detracted from the report in more subtle ways. In summary, the reviewer believes that EPA was forced to use a particular approach because of the lawsuit and elected not to use other approaches because they were not required to do so.

This reviewer is also concerned about a generic government policy that forces the selection of low bidders for conducting bioassays in the final selection process. While it is not clear how this may have affected the final results, it gives the appearance of removing randomness from the selection process, particularly when the referee laboratory results were deliberately excluded. The reviewer understands the low bidder process but it certainly could not have increased the likelihood of a random selection process. However, on several sections of the report the reviewer had the distinct feeling that the low-bidder process may have precluded the ability of EPA to include the wide spectrum of capabilities from laboratories across the country that routinely conduct these tests and satisfy the demand of the litigants that "average" results be evaluated in this study. While it could be argued that the purpose of this study was to omit the "best" laboratories and include only "average" laboratories, the low-bidder approach seems to favor the selection of the "worst" laboratories. It could also be argued that only the "best" laboratories could stay in business with low prices, it has also been the experience of this reviewer than most laboratories with low prices are cutting corners somewhere where it is not apparent, but where it could adversely affect the results. In other words, the reviewer is concerned that the variability and false positive rates reported here may have been over-estimated, while test completion rate may have been under-estimated from that of the truly average labs. Another non-random aspect of the selection process includes laboratories that were subsidized by outside funding sources. This could have added an additional bias. In other words, these labs were not selected at random either. It is not clear how this could have affected the results, but the bias remains. If there was a tendency to favor selection of the

"best" laboratories, this could have made the results appear better than they actually were. Alternatively, if the "worst" laboratories were selected because of some particular affiliation, this could have made the results poorer than they actually were. In either case, this is another source of bias.

As an example, the following sections (and others) should be edited to transfer more information to the Appendix, keeping only the pertinent information in the body of the report:

Section 3.3

Section 4.0 (Preliminary Testing) goes from P21 - 49. While this section on preliminary testing contains important information, it really takes away from the message of the report. Since these data were not used in calculating any of the three parameters to be assessed in this study (variability, test completion rates, false positive rates), it is not clear why the reader is subjected to almost 30 pages of supplementary information. This is a good candidate section for the appendix. The other problem with this section is that it ends abruptly and is quickly followed with Section 5.0, Sample Preparation. Without even taking a breath the reader is launched from preliminary testing to final testing with no mention of a transition. It appears as if this section on preliminary testing was added at the last moment, almost as an afterthought. The flow of the report would be much better without this section, or a reduced version. Even then, more of a transition from lab selection to sample preparation may be necessary. Within Section 4.0 there is also a slight transition problem in going between the different parts (1-4), particularly going to Part 4 titled "Definitive Testing." Even though this part is clearly listed under Section 4.0 title "Preliminary Testing," the reader could get the impression that "Definitive Testing" is really a transition from "preliminary testing to final testing." This is particularly true as the reader goes from P17 - 20 without another "Preliminary Testing" heading. The phrase "Definitive Testing" is misleading in this context. Regardless of whether or not EPA decides to include this information in the body of the report, it may be more clear if "Preliminary Testing" or "Preliminary" is included as a subtitle in each of the Parts (1-4). The other benefit of moving Section 4.0 to the Appendix is that it would remove any ambiguity whether this was a significant part of the Interlaboratory Variability Study. The question of ambiguity is related to the inclusion of data from the referee laboratories in the calculation of variability, successful test completion, and false positives as I previously discussed. The argument raised is that there could be something useful in these preliminary tests with both referee and participant laboratories to provide a different perspective on the Final Results. The reader has also expressed concern that data from the referee laboratory was not included at the insistence of the litigants. EPA makes a good case for not including these data because the laboratories were not naive about the identity of the samples. The argument for including the data from the preliminary testing is the same one made for including it in the final testing; i.e., analyze the data both ways, with and without the referee laboratories and with and without preliminary testing. The reviewer believes that this would provide more insight into variability, test completion and false positives than simply omitting it. It might also provide some information on whether specific issues such as holding time had any influence on variability, test completion, and false positives. Additional ambiguity is added in the last sentence of Section 4.1.4: "During interlaboratory testing, referee laboratories again shipped samples round-trip back to themselves and conducted testing simultaneously with participant laboratories." The reviewer believes that this sentence is intended to identify that the same shipping procedures were used during the preliminary testing as the final testing and that these procedures were intended to have the samples be treated as equally as possible among all labs, including the referee labs. Unfortunately, that is not what the sentence actually says. First, although preliminary testing is mentioned in the first sentence of this section, the main heading is definitive testing which is ambiguous in itself as described above. Second, the comparison is supposed be made between "definitive testing" and "interlaboratory testing" in that the samples were treated equally. In this context it is difficult to distinguish between "definitive testing" and "interlaboratory testing." Third, "definitive testing" implies final, decisive, or conclusive and does not seem to fit with "preliminary testing." "Interlaboratory testing" implies nothing with respect to preliminary or final

testing and should be clarified, particularly since the preliminary testing did include an element of interlaboratory testing even though EPA decided not to use the data in that way (which the reviewer also objects to). This reviewer believes that different descriptors should be used to clearly identify what was done for the "preliminary testing" and what was done for the "final testing" or perhaps switch the phrase "definitive" to apply to interlaboratory testing such as "final definitive interlaboratory testing" and something like "preliminary testing trials."

The following tables should be edited, combined, or removed to shorten the body of the report and transfer more information to the Appendix:

Table 2.4; Table 3.3; Tables 4.1 through 4.22 (All the tables in section 4.0)
All the tables summarizing laboratory test conditions do not need to appear in the body of the report and in Appendix A. Depending on how EPA decides to reformat the body of the report, if at all, these tables could go in either the body of the report or Appendix A, not both.

If possible, some tables should be converted to graphs to make them easier to read. This could be another method for shortening the report. If the crux of the tables can be shown in a figure, the detailed data that went into that figure could be included in the Appendix.

Terms to be included in a glossary, list of acronyms, and species list:

e.g., spiking - sample to which a material has been added for experimental purposes (8th Edition ASTM standard definitions p 492)

Glossary

accessibility
accuracy
acute
aeration
airbill
alga
algae
algal suspension
aliquot
alkalinity
ambient laboratory
illumination
ampule
artificial sea salts
background testing
base study design
between-lab
bioassay grade
bioassessment
biological oxygen demand
blank
blind sample
brood
brood board
bulk effluent
chemical oxygen demand
chronic
cladoceran
cleaned and rinsed
control chart
"Cool white"
cooler
culture
data qualifier flag
database
definitive testing
discharge outfall
dissolved oxygen
effluent
endpoint
episode
Erlenmeyer flask
excursion
extended study design
false positive
FedEx
Forty Fathoms
h statistic
hardness
holding time
homogenize
hypothesis testing
industrial effluent
interlaboratory
k statistic
larval
life stage
low bidder
Millipore Milli-Q
minimum significant difference
moderately hard
municipal wastewater
mysid
narrative water quality
nauplii
neonate
numeric water quality
outlier

participant laboratory
pH
photoperiod
plastic carboy
plastic container
point estimate
precision
prequalification
quality assurance
quality control
random selection
range-finding
reagent
reagent grade
real-world
receiving water
reconstituted
referee laboratory
reference toxicant
river water
rounding
sample code
significance level
spiking
standard
standard dilution factor
state certification
static non-renewal
static-renewal
submersible pump
synthetic freshwater
synthetic seawater
total suspended solids
total organic carbon
total dissolved solids
total ammonia
total residual chlorine
tracking
traffic report
treatment plant
valid test
within-lab
within-laboratory precision
within-laboratory variance
Ziploc bag

Acronyms

ASTM
CV
DMRQA
EDTA
IC25
KCl
LC50
MHSF
NOAEC
NOEC
NPDES
PROC MEANS
PROC MIXED
SAS
SCC
SETAC
SOP
SOW
WET
YCT

Species

Artemia
Ceriodaphnia dubia
Champia parvula
Cyrpinodon variegatus
Holmesimysis costata
Menidia beryllina
Mysidopsis bahia
Pimephales promelas
Selenastrum capricornutum

3.2 Response to Charge

1. Document's Responsiveness

Reviewer X

· Is the WET Study Report responsive in evaluating WET test method precision, successful test completion rates, and false positive rates?

Yes, in terms of the WET Study Plan but see other comments, below. The evaluations were restrictive in terms of both the toxicants evaluated (Cu and KCl) and in terms of the concentrations used. The test concentrations were set and standardized between laboratories, which would not normally occur in "real life". Variations in test concentrations between laboratories would have greatly increased NOEC variability (Chapman, 2000).

However, the major focus should be not on CVs but rather on the difference between the min and max for acceptable tests. This is what is truly important in a regulatory sense. Regulators look at values, too often in terms of "bright lines". Examination of the data tables reveals that the difference between min and max can, depending on test, etc. sometimes be 8-fold to >10-fold and is often between 2 and 4-fold. This is recognized in noting the concentration range between NOEC values, which again can span several concentrations. My "take" on the min and max differences and the NOEC differences is that there is much more variability occurring in a regulatory sense than in apparent from simply examining CVs. I would judge the tests in terms of how they do against a factor of 2 guideline (min and max within a factor of 2 and NOEC values do not exceed 2 concentration ranges). Greater variability than this is, in my opinion, a real problem for hard regulatory use of these tests. Quoting "percentage of values within one concentration of the median" is misleading and not useful.

Another issue is whether tests identified as invalid in this variability testing would have been so identified in "real life". See for instance, last paragraph on page 102. Similarly, recalculations are good but would these have occurred in "real life"? I doubt it – if a lab provided their calculations and signed off on them, in most cases they would have been taken at face value. Thus the actual level of false positives in "real life" as defined by this study can be expected to be higher.

· Does the WET Study Report provide reference to the data and equations used to calculate the coefficient of variation (or other applicable estimate of precision) for each test endpoint?

Not in all cases. For instance, it took me a while to figure out what the total CV values were in Table 9.6 and following tables. I am concerned that an average was taken of a percentage. I was always taught in math that one goes back to the original data, rather than taking averages of averages. And in some cases I still cannot figure out from the text (which I did review several times), how total precision values were derived. For instance, see Table 9.33. I understand within-lab, between-lab and average values. But I do not understand where the total values came from. It may be my simple mind, but I suspect I will not be the only one confused. Also, beware of too many significant figures. For instance, Table 9.3.2 IC50 values in some cases have too many significant figures (e.g., 3.51 should be 3.5).

· Does the WET Study Report provide a chart of any new coefficients of variation for the various WET methods derived from the results of the Interlaboratory Variability Studies?

A chart is not provided, but CVs are. What is meant by a "chart" in the question above?

Reviewer Y

The document is responsive to the requirements of the settlement agreement relative to WET method precision, and rates of false positive tests and test completion. Despite the clear intent, thoughtfulness, and focused efforts of the EPA, the Survey results do not allow assessment of two WET methods (Champia parvula reproduction and Holmesimysis costata acute lethality) due to insufficient laboratory participation. Failure of copper spiking for some tests compromises, but does not exclude, discussion of precision, and rates of completion and false positives for those tests. Although a minimum of nine laboratories was required for each test, only seven laboratories conducted the sheepshead minnow acute testing. This reduction in number of laboratories likely did not compromise conclusions.

To improve this Survey report, specific comments and suggestions are made below. These detailed comments do not suggest inadequacy of the Survey. They are suggestions for enhancing interpretation and optimizing presentation of results.

Reviewer Z

The document directly addresses the three major elements of the settlement agreement: a) false positive rates; b) successful test completion rates; and c) coefficients of variation for each of 12 WET test methods. Furthermore, EPA did a very good job of addressing each of these three issues with scientific rigor and provided sufficient detail to explain the experimental design and the test results. In other words, the WET Study Report was extremely responsive in evaluating WET test method precision, successful test completion rates, and false positive rates. The main problem is that the body of the report contains too much detail for the reader to easily follow the train of thought from background to procedures to results. The lack of conclusions also seems to leave the report unfinished even though EPA did everything that was required.

- Is the WET Study Report responsive in evaluating WET test method precision, successful test completion rates, and false positive rates? YES

- Does the WET Study Report provide reference to the data and equations used to calculate the coefficient of variation (or other applicable estimate of precision) for each test endpoint? YES

- Does the WET Study Report provide a chart of any new coefficients of variation for the various WET methods derived from the results of the Interlaboratory Variability Studies? YES

2. Document's Scientific Merit

Reviewer X

- Was the Study Plan appropriately implemented to achieve the objectives of determining method precision, successful completion rates, and false positive rates?

Yes, within the limits of the Study Plan, which was not subject to this review.

- Was the inclusion/exclusion of data appropriate for the determination of precision, successful test completion rates, and false positive rates?

With the exception of the exclusion of referee laboratory data (see comments in Section 3. of this review), it was appropriate.

- Were the methods for calculating precision, successful test completion rates, and false positive rates scientifically acceptable?

Yes, except where percentages were taken of percentages or the methodology for the totals was not clear (see comments elsewhere in this review).

- Are the results scientifically acceptable within the context of the intended regulatory use?

Yes and no. The data are there, though they need clarifications as noted in this review. However, I am not convinced that the Study Plan allowed for direct comparisons with regulatory use. For example, test concentrations were regimented and had larger than normal gradations, and false positives were not evaluated in terms of ecological significance but rather in terms of testing only. These tests are applied, too often, as decisive when (see Section 5.0 of this review, below) they are far from such.

Reviewer Y

The study generated data sets useful for defining precision and rates for ten of the 12 WET methods. Results from the Champia parvula reproduction and Holmesimysis costata acute lethality methods are preliminary and associated conclusions are not definitive. Those tests relaying heavily on copper spiking were compromised by copper precipitation and associated conclusions should be tempered by this fact.

The intent was to estimate precision and rates for the WET methods under routine conditions and with samples representing those normally tested with WET methods. The predominant reference toxicant (KCl) was one that likely produces more consistent results than would complex toxicant mixtures that make up many effluents. The selection of KCl was reasonable and it is not being discussed here as an error. The difference between KCl-related and actual toxic effluent-related precision should be discussed briefly and a degree of temperance applied in speculating from these results to all effluents assayed with WET methods. Precision based on KCl spikes may reflect the best precision to be expected.

Does selection of laboratories for participation go beyond the normal process that would be used by a permittee to select a laboratory to do WET testing of their effluent (see comments below)? The difference is probably insignificant but it should be discussed. Brief discussion is needed about how representative the selected laboratories were of the population of laboratories for which conclusions are being made.

Modification of calculations used to estimate precision for censored data sets would enhance the accuracy of conclusions. The methods applied to precision estimation for censored data are demonstrably biased (see below) but easily corrected. Also, assessment of conclusions would be greatly enhanced if 95% confidence limits for IC25 and LC50 values were provided in tables. Finally, variance estimates are needed for rates of test completion and false positive results.

Reconsideration of the use of outlier tests could make the results more reflective of the precision expected among laboratories implementing the WET methods. If I understand the outlier testing correctly, the valid results from the various laboratories were subjected to outlier testing to discard any unusual data. Although only 15 of 698 observations were discarded as a consequence, it is impossible to tell how this atypical exclusion influenced the precision estimates. Regardless of the number of outliers discarded, this step would not have occurred among laboratories applying the WET methods. During normal testing, a laboratory would not have knowledge of results from many replicate tests of an effluent sample.
Relative to the intended regulatory use of the data (see paragraph 1 of this review), the results will be technically acceptable and valuable after the minor changes mentioned in this review are implemented.

Reviewer Z

The approaches used to plan the study, implement the study, and review, analyze, and report results were scientifically sound. A very rigorous approach was used in planning the study. It is surprising that given the attention paid to every small detail that there were careless errors by the laboratories such as transposing sample numbers. While this could mean that the planning or implementation were not scientifically sound, the reviewer believes that it was human error caused by carelessness rather than scientifically unsound procedures. This could also have been related to the low-bidder process used for laboratory selection.

- Was the study plan appropriately implemented to achieve the objectives of determining method precision, successful test completion rates, and false positive rates? YES

- Was the inclusion/exclusion of data appropriate for the determination of precision, successful test completion rates, and false positive rates? YES/NO. While it would probably not affect the outcome of the results, the reviewer believes that referee laboratory results should have been included in a separate series of analyses for comparative purposes. EPA could choose to use or reject those results in the decision-making process, but the data should be included to show whether or not difference occurred.

- Were the method for calculating precision, successful test completion rates, and false positive rates scientifically acceptable? YES

- Are the results scientifically acceptable within the context of regulatory use? YES/NO. The results are scientifically acceptable within any context since the approach was scientifically rigorous. However, there is a distinction between scientifically acceptable in terms of accepting the results versus whether or not the results are acceptable for regulatory use. This is reminiscent of the following story: "The operation was a success, but the patient died!" The results should be accepted, but the results seem to show that some of these tests should not be used in the regulatory context because the successful completion rate is too low and the CV values are too high.

3. Document Quality and Clarity

Reviewer X

- Is the WET Study Report clear in stating the objectives, describing study implementation, and presenting study results?

Yes, with caveats noted in the specific comments (Section 2.3 of this report).

- Are data summarized and presented in a technically clear and appropriate manner?

Yes, with caveats noted in the specific comments (Section 2.3 of this report).

Reviewer Y

The document was clearly written with a few exceptions.

More specific discussion of statistical estimation methods applied for IC25, LC50, and NOEC calculation would enhance the document. Some undefined portion of the imprecision could emerge from differences in applied methods, e.g., different methods used for NOEC estimation.

The discussion of laboratory results during preparation of spiked samples is vague relative to characterizing the "acceptability" of the spiked sample prior to being sent to the participating laboratories. As one example, the last paragraph of page 21 discusses a sequence of results leading to the use of a certain spiking concentration. Yet, results from the survey suggested a much lower toxicity for the spike. In another example, effluent spiking results were judged "persistent and appropriate" on page 23, paragraph 2 when a 5.8% difference in toxicity was estimated on repeat testing but much more vague conclusions were drawn on page 37, paragraphs three and four about acceptability of copper spikes. Page 28, Section 4.2.2.2, paragraph one contains description of another sequence of steps leading to the establishment of a final effluent sample spike. The clearer presentation of the exact process of deciding what was or was not acceptable for spiked samples would help. Again, 95% confidence limits for results would help the reader to decide whether or not toxicity differences between repeated tests were trivial or unusual.

Quantify the Precision for the Twelve WET Methods

Ten WET methods could be assessed. The results for these ten methods will be useful for assessing method precision after the recalculation of CV's associated with censored data sets.

Quantify the Successful Completion Rates for the Twelve WET Methods

Ten WET methods could be assessed. The production of uncertainty estimates is needed to complete description of results for successful completion rates.

Quantify the False Positive Rates for the Twelve WET Methods

Ten WET methods could be assessed. The production of uncertainty estimates is needed for the false positive rates. Where possible, the results should be compared to the number of false positives expected based on random chance alone. For example, expected numbers of false positive results could easily be estimated from the NOEC-associated hypothesis tests that use experiment-wise error rates of 0.05.

Scientific Acceptability Relative to the Intended Regulatory Use

In general, the study results will be acceptable for ten of the WET methods after the suggested minor changes have been made to the report.

Reviewer Z

This document is of very high quality and in general very clear. The main problem seems to be that the most important information is in Section 9.12, Results Summary, at the very end of the report. This problem is exacerbated by the fact that there are no conclusions or regulatory context given to the results. While it may have been EPA's intent to wait for the peer review comments, it might have been helpful for the peer reviewers to understand EPA's interpretation of the regulatory significance of the data. This may have been precluded by the court settlement or EPA's concern about introducing bias to the peer review comments. This reviewer is also concerned that litigation sensitivity may have introduced bias to the science in other areas as well. For example, EPA carefully avoided including data from the referee laboratories in any of the summary statistics. Since these laboratories had prior knowledge of the samples, it might seem reasonable to exclude these data for a completely unbiased scientific sample. Alternatively, this reviewer believes that important information was lost by excluding these data. It would have been more informative to present two

sets of summary data, one with the results from the referee laboratories and one without. If no differences were detected in results from referee versus participant laboratories, the results could have been pooled to provide a more representative sample. Alternatively, if the introduction of the referee laboratory data was significantly different, EPA could explain why the data were excluded.

It is not clear why these data were excluded. If EPA decided to exclude them for scientific reasons that is acceptable although this reviewer has provided an alternative for including or excluding these data in different analyses for information purposes. If, however, EPA was forced to do this as part of the litigation settlement, that is another matter. This reviewer is troubled by science being biased in this and other cases by the legal system. Since this could be an issue, EPA should clarify both in the Executive Summary and in the Background Information what they were required to do and what they were precluded from doing as part of this WET study. The fact that "The study was precipitated by litigation over the rulemaking that standardized and approved the WET test methods for use in NPDES monitoring" should be included in the Executive Summary. The fact that "In a settlement agreement with the litigants, EPA agreed to conduct the WET study and determine false positive rates, successful test completion rates, and coefficients of variation (CVs) for each of 12 WET test methods" should also be included in the Executive Summary. It is important to note that EPA was essentially forced to conduct this study and that the litigants may have biased the results in their own way by forcing a particular approach. This reviewer reviewed reports [on various web sites] with comments by the litigants on previous drafts of this document that indicate such a bias may have been forced upon EPA. Apparently the litigants commented that referee laboratory data should not be included in the summary calculations because of possible introduction of bias. Another bias might have been included by excluding these data.

This reviewer believes that it was appropriate and necessary to conduct this study and that important information was gained with respect to WET test variability, successful completion rates, and false positive rates. However, the legal system should not have dictated the science approach that was used to conduct the test or interpret the data. It seems likely that EPA was required to have its own attorneys review the study plan and debate with the other litigants what was acceptable and what was not acceptable. This in itself has compromised the study and EPA should acknowledge this fact in the document. EPA should make it very clear what they were required to do and explain, for example, the reasons for not including the referee laboratory results in the statistical analyses, including all the minute details of the study, and excluding simple calculations of mean, max, min, etc. in areas such as neonate production, number of broods and time for test completion. The reviewer believes that scientists at EPA have probably already made many of these calculations, if for no other reason than intellectual curiosity. Other readers of this document would like to see and have the right to see that information. In addition to the statistical analyses of test variability, successful completion rate and false positive rate, the other statistical comparisons in the Ceriodaphnia test mentioned above could be used in a preponderance-of-evidence approach to evaluate whether or not referee laboratory test results were really different from the participant laboratories. It is also surprising that EPA and the litigants did not reach some agreement in advance (before testing) on what level of variability, test completion rate, and false positive rate would be deemed acceptable and scientifically defensible for regulatory purposes. It appears that this lack of agreement will lead to additional controversy and acrimony and nothing will be resolved even after the peer review. The reviewer could offer a personal opinion with respect to the results of each test and whether or not the results justify its continued use for regulatory purposes but that is beyond the scope of the charge to peer reviewers. This reviewer has offered some informal comments on some test results but this issue should probably be reviewed and discussed by a panel of experts from SETAC or some other group that could provide specific guidance on this issue.

Given that this is a preliminary report, each section of the document is surprisingly clear with an absolute minimum of spelling and typographical errors although the transition among some sections could be

improved. Most of the editorial comments were relatively minor, although some reorganization might be helpful to shorten the document. There were some inconsistencies in the use of technical descriptive terms between sections and it appeared as though different authors wrote different sections, and this made the report somewhat more difficult to read. The important issue is that most if not all of the important information is included and presented in a clear and concise way within each section. The reviewer suggests that the material be reorganized so that the most important information be presented sooner rather than later and that the sections before results be condensed. The reviewer also suggests that the Executive Summary be expanded considerably with appropriate summary tables so the reader can see the final results by reading the Executive Summary. Conversely, the excellent table that appears in the Executive Summary should appear in the Results Summary (Section 9.12) and some of the tables in Appendix A be included in the pre-results sections. Repetitive tables in Appendix A should be deleted.

- Is the WET Study Report clear in stating the objectives, describing study implementation, and presenting study results? YES

- Are data summarized and presented in a technically clear and appropriate manner? YES

3.3 Specific Comments

Reviewer Z

Page xiii, Paragraph 1

Executive Summary. The opening paragraph is very well written. In particular, the 1st sentence clearly states what the report will present (results of the variability study) and the last paragraph clarifies the purpose of the study (interlaboratory variability, rate of successful test completion, rate of false positive incidence). It may be helpful to characterize the study as the "WET Variability Study" rather than just the "WET Study" because the latter is too vague. In practice, there could be several different studies of WET testing in the past, present and future but probably far fewer that focus on variability. Strictly speaking, since the study also included an assessment of successful test completion and rate of false positives, it could be argued that the study should be called something different. However, since the title characterizes it as an "Interlaboratory Variability Study" and rates of successful test completion and false positives could be considered part of a variability assessment, this shorter version of referring to the study as the "WET Variability Study" should be acceptable an abbreviated title for quick reference and yet more descriptive than just "WET Study". Alternatively, an acronym could be used in place of the "WET Variability Study", the "WVS" which is somewhat difficult to read or the "WET VS" or even the WET Variability, Completion, False Positive Study", the "WET VCFP".

Reviewer Z

Page xiii, Paragraph 2

It is interesting to learn that interlaboratory data were not obtained for two tests (red macroalga reproduction test and mysid acute test) due to insufficient participant laboratory support. However, the reader immediately wonders about the significance of this result. Perhaps the significance of the result will be explained later, but it might be helpful to add a sentence about what this means to the program or that the significance will be addressed later in the report. Some context in terms of the Executive Summary should be added here.

Reviewer Z

Page xiii-xiv and Summary Table

It is generally not helpful to begin a paragraph with a "no technical information sentence" like the location and contents of Table 1. This is particularly true in the Executive Summary where all attempts are made to be concise. In fact, this sentence may not even belong in the executive summary. It seems as though the main purpose of the Executive Summary is to summarize technical information contained in the report, not to identify where that information is found. If the reader is interested in locating this information the table of contents should suffice for that specific information. Please note that all the preceding comments were the reviewer's were based on the first read of the document and before realizing that Table 1 was part of the Executive Summary. Having gone through the entire document, and even though tables are generally not included in the Executive Summary, the reviewer now believes that this table is not only appropriate but necessary in the Executive Summary. The table is certainly a good way to summarize the results of the WET variability study. In fact, this reviewer now believes that the entire Executive Summary should be re-formatted to include that table on the first page so the reader does not have to read through the entire Executive Summary to learn the crux of the report or to flip to the next page after reading the sentence about Table 1 on the first page. This information is important and a summary table is the best way to get the message across. It should be the focal point of the first page and the other paragraphs used to set the scene and explain what was done. This reviewer suggests placing the last paragraph at the top of the page with Table 1 underneath and rewording the last paragraph to show the table in parentheses after the important summary information. The details regarding which specific tests were used, the number of labs participating, and the command and control system are supplemental details that are less important than the test results. In fact, the reader can see at a glance in the table which test methods were assessed. Therefore, all of the

supplemental information could be placed on the following page with no loss of impact to the reader. Given that so much detail is provided in this report, this reviewer also believes that EPA would be justified in extending the Executive Summary to include more of this detail. One item that seems particularly conspicuous by its absence is that "The study was precipitated by litigation over the rulemaking that standardized and approved the WET test methods for use in NPDES monitoring..." In a settlement agreement with the litigants, EPA agreed to conduct the WET study and determine false positive rates, successful test completion rates, and coefficients of variation (CVs) for each of 12 WET test methods. Based on the results of the WET study and peer review comments, EPA will ratify or withdraw each of the WET test methods." This seems like extremely important information that should be included in the Executive Summary. There are similar details regarding the conduct of the test that have an overall significance to the study that should also be included. For example, the SCC consultant and the four referee laboratories should be identified and their selection process described in the Executive Summary.

Reviewer Y

Page xiv, Table 1.

Please provide n and, where possible, estimates of uncertainty about these calculated percentages. The number of significant figures seems to be inconsistent here and in other places in the report. Finally, should ranges be reported for estimated NOEC values as part of the summary table?

Reviewer Z

Page 1, Paragraph 1

Section 1.0. EPA has done a good job of defining WET in the opening sentence of the 1st paragraph and outlining the tests that were promulgated to for measuring WET. However, there appears to be a slight disconnect between moving from the 1st paragraph to the 2nd paragraph on Regulatory Background.

Reviewer Z

Page 1, Paragraph 1

Section 1.1. The first sentence of the first paragraph should clearly describe how WET testing supports the Clean Water Act. The connection between WET and the Clean Water Act is too vague and the inclusion in the list of three approaches used by EPA for "water quality based toxic control" is too ambiguous. EPA should clearly state that WET testing is an integral part of the three integrated approaches and explain how they collectively support the enforcement of the Clean Water Act.

Reviewer Z

Page 1, Paragraph 2

The 2nd sentence of the 2nd paragraph includes the following "Some states have included numeric criteria for WET, while others have relied on narrative criteria." This reviewer and other readers would probably immediately ask the question, "Which states and how many on a percentage basis use which method and why?" It seems as though one or two sentences here would help clarify the relative percentage for each approach or if there is a regional or volume (of effluent) factor that helps determine which approach is used where. The term "excursion" is used twice. Once in each of the last two sentences of this paragraph. While the term "excursion" can be used to indicate "deviation," "diversion," or "digression," in practice "excursion" is generally used with qualifiers such as "excursions above" or "excursions below" some criteria. This reviewer believes that it would be better to use more direct descriptors such as "deviation," "diversion," or "digression" to remove any doubt about the meaning. Alternatively, the text could identify "excursions above or below" the water quality standard. See a similar misuse of this term on P85 Section 8.2.2. 1st sentence.

Reviewer Z

Page 2, Paragraph 1

Section 1.2. Same comment as on P1 regarding the title of this section and the parenthetical expression in the 1st sentence. It may be helpful to characterize the study as the "WET Variability Study" rather than just the "WET Study" because the latter is rather vague. The last several sentences of the 1st paragraph identify that this WET Variability Study was promulgated by a lawsuit and that the results of the study and peer review comments will help determine which test methods will be ratified or withdrawn. This seems like a very important piece of information which should be included in the Executive Summary. More detail should be given regarding the duration of the study rather than just "1999-2000." Since the actual testing dates are given later in the text it might be helpful to give the exact dates for the beginning and termination of testing. This will remove any doubt on behalf of the reader regarding the magnitude of this effort. As stated here, the study could have been a short study from December 1999 to January 2000 or a very long one from January 1999 to December 2000. Providing the exact dates will remove any doubt. It will also help place the study in better context when the reader reaches those sections of the text.

Reviewer Z

Page 2, Paragraph 2

Section 1.3. EPA has done a good job of documenting the chronological history of documents used to clarify various aspects of WET testing. Including these details is important in a document of this type. It is not clear, however, if all of these documents were used by the laboratories conducting the tests. Some clarification or context is necessary to describe how these documents were used. As a finer point of discussion however, it is vague to refer to this study as a WET study. This report examines the variability in "WET" as defined in pages 1-2. Therefore, this reviewer suggests using throughout the text as the WET Variability Study." This is also consistent with the title which accurately characterizes the study as a variability study of test methods. That is what this report is all about and that is how the study should be characterized in a shorter form throughout the text.

Reviewer X

Page 2, Paragraph 4

USEPA, 2000b should be 2000a and, next paragraph, that 2000a reference should be 2000b. This is not to imply that the wrong references are cited but rather that the first time a reference with multiple designations (a,b,c) is cited, one should go alphabetically.

Reviewer Y

Page 3, Paragraph 1

Section 2.1. The precision is estimated using a CV that requires the assumption of a normal distribution of observations. Where the observations normal for all IC25 and LC50 sets of observations?

Reviewer Z

Page 3, Paragraph 1

Section 2.1. EPA has done a good job in outlining the objectives of the study, describing how the study plan was distributed for public and peer review, and how the plan was revised to incorporate public and peer review comments. It would be difficult for anyone to argue that EPA should have conducted this process any differently. However, much of the wording in the introductory portion of each section is repetitive and somewhat colloquial; e.g., "generate data to characterize." The reviewer suggests modifying the introductory sentence as follows: "In conducting the WET Test Variability Study, EPA's primary objectives were to characterize: 1) interlaboratory variability...; 2) the rate at which participating laboratories...; and 3) the rate at which WET tests indicate...." This reviewer regards "generate data" as colloquial, unnecessary, and too repetitive throughout the list in the 1st paragraph.

Reviewer Z

Page 3, Paragraph 2

Section 2.1 Bullet Items. As in the Executive Summary, this reviewer is struck by the fact that only one laboratory could be procured for the Champia chronic test method and only two participant laboratories for the Holmesimysis acute test method. Why did this happen and what is the significance of this result? Somewhere in this report EPA needs to clarify why this occurred and what it means to the testing program. In other words, does it really matter what the variability in the test is if few if any labs routinely conduct this test? These statements need to be placed in the context of the WET testing program.

Reviewer Z

Pages 3 and 4

Section 2.2.1. The SCC consultant and the four referee laboratories should be identified in the Executive Summary. This reviewer believes that EPA should explain how the four referee laboratories were selected and provide a definition for a referee laboratory. For example, were the four referee laboratories selected because they were under contract to EPA, did they bid on this job and was cost a significant factor in selection, or were they selected because they were determined to be the best for this job based on some type of ranking system? Since these labs play a significant role in the outcome of the study, reviewers will want to know more about how they were selected. It may also have some bearing on how credible the results are with respect to the litigation mentioned previously. Beyond the referee laboratories, how were the participant laboratories selected? Could any volunteer lab participate or was there some threshold for acceptance. Table 2.1 itemizes prequalification and selection but does not identify the qualifications. On the whole, however, it is very informative and a nice way to present the information. It is also somewhat incongruous to find the list of referee laboratories by name in the Acknowledgment on Pxii, mentioned as referee laboratories in the Executive Summary, a list of test methods on P5, but the actual list of referee laboratories and the associated test methods for which they were responsible not identified until P14 in Table 3.2. This reviewer actually placed flags on Pxiii for the list of referee laboratories, P5 to keep track of the test methods, and P14 to link referee laboratories with test methods as the rest of the report was reviewed. This probably means that this information should have been consolidated and presented earlier to avoid confusion in the early portion of the text as well.

Reviewer Z

Page 4, Paragraph 2

Section 2.2.2. The first sentence in this section identifies that; "EPA evaluated 12 of the 17 promulgated WET test methods in the WET study. These included two acute freshwater methods, three chronic freshwater methods, three acute marine methods, and four chronic marine methods." This is a good start but none of the tables or succeeding text clearly identify which test belongs in which category. This is a recurring problem throughout the report. The reader should not have to search through the document to deduce which test belongs to which category, it should be clearly identified in a table as this reviewer suggests by adding another column to several tables near the beginning of the report. Actually, there are no tables that clearly identify the categories of testing "freshwater or marine" or the round in which they are included (1, 2, or 3). This is most clearly identified in Table 1 of the Appendix. In fact, the entire Section 1 in the Appendix (Introduction and Background) is easier to read and understand than comparable sections in the main report. Several issues that are identified as problems throughout the main report are clarified early in the Appendix in approximately half the space. These include issues such as identification of marine and freshwater tests, linking the tests with the appropriate category, identification of a "base study design," and clear identification of the testing done in each round. This reviewer believes that EPA should consider replacing the 91 pages of text leading up to the results with the 36 pages that are included in Appendix A. The intended purpose of Appendix A is not clear other than a shortened, easier to read version of what appears in the main body of the report. The reviewer believes that it is not necessary to duplicate the material in the text with the shorter

version in the Appendix. EPA should use Appendix A in place of most of the material in the body of the report. This would save about 50 pages of unnecessary text. If required, EPA could include some of the detailed information in the Appendix instead of what is there now but the report would certainly be easier to read with this replacement. It is also not clear why the summary of test conditions are repeated in the tables of Appendix A. What is the purpose of Appendix A? The Appendix should only include those details necessary to understand the finer points of the study without burdening the body of the report. Much of this material is just repetition and is not necessary.

Reviewer Z

Pages 5 and 6

Section 2.2.3. This is a good description of how and why the laboratories were selected but a capsulized version of this section should be included in the Executive Summary and in an introductory paragraph in this section. This is a recurring problem throughout this document; i.e., the individual sections are very well written, but this reviewer does not believe that they have been adequately summarized in previous sections of the document to alert the reader as to what is coming later (in more detail). In summary, the old adage of organization would be helpful here: "1) Tell 'em what you're gonna tell 'em; 2) Tell 'em; and 3) Tell 'em what you told 'em." This is also relevant to Table 2.2. This table is very informative, but the reader should be reminded about relevant freshwater and saltwater categories in this section as a prelude of what is to come later. The reviewer did not find the word "seawater" until page 32. There, it is in reference to the composition of reference toxicants. EPA should also be consistent with other documents that use the term "saltwater" rather than seawater. Obviously, seawater is appropriate for a description of reference toxicant composition, but there should be major categories in each section that separate freshwater from saltwater tests. This could also be accomplished in the table with a header along the side and perhaps shading to designate these major categories of tests as a reminder to the reader. The other generic problem with this report is that some of the sections include too much detail to the detriment of making crucial points about the results. This reviewer is well aware that this information must be provided somewhere and EPA was probably under pressure from the litigants to include more detail rather than less. However, much of this material belongs in an Appendix.

Reviewer X

Page 6, Table 2.3.

Footnote: state, not states.

Reviewer Z

Page 7, Paragraph 1

Section 2.2.4 For each test method, four test sample types were prepared in bulk by the referee laboratory, divided, and distributed to participant for testing." The referee laboratories are listed in the Acknowledgment but it is not clear what each one did. Did each referee laboratory have an area of speciality and prepare only those samples for the participating laboratories? Was each referee laboratory assigned to a particular area of the country and prepare samples for each region? Or was it handled differently? Not only is this interesting, but it has a significant bearing on the statistical design and the validity of the results. The referee laboratories should be listed here and the activities of each clearly described. Then, specific duties should be referred to in the sections where they appear. There is a real disconnect in this section between what was done, by whom, and when. This should all be tied together and very clearly explained in this section. This information is not provided until the Table on P14.

Reviewer Z

Page 7, Paragraph 2

Section 2.2.4 It is not clear what is meant by "base study design" which is mentioned at least twice in this paragraph. More information is presented later to clarify this point, but it should be clarified here or earlier if mentioned previously. See also P14 Section 3.5. At a minimum the reader should be referred to Figure 3.1 to show how "base design" fits into the overall scheme. Even this is inadequate since "base design" is not clearly defined here or elsewhere. The same is true of "extended study design." More detail is needed in this paragraph.

Reviewer X

Page 7, Paragraph 3

Line 1: No comma after month and before year where the day of the month is not provided

Reviewer Z

Page 7, Paragraph 3

Section 2.2.5. It is not clear why samples were tested with and without EDTA. What is the purpose of EDTA and how are the results to be interpreted based on their respective responses if different?

Reviewer Z

Page 7, Paragraph 3

Section 2.2.5. This is a good description of sample distribution relative timing of the tests. Other activities should be described in similar detail. Furthermore, this single sentence describing the length of the study should be included in the Executive Summary and in the Introduction. It might also be helpful if a short-hand description of the time period were also included; i.e., approximately 6 months. Strictly speaking, even giving the months of testing is inadequate since the tests apparently began as early as 28 September 1999 and ended as late as 4 April 2000. Since the beginning of the study was at the end of the first month and the end of the study was at the beginning of the last month, there were really only 6 full months of interlaboratory testing. This means so much more than simply saying that the study was conducted between 1999 - 2000 that the extra detail is necessary. EPA may also choose to include the preliminary testing conducted by the referee laboratories. This would extend the time period but the text currently does not describe this very well and there are no tables for clarification. Regardless of EPA's choice, more information needs to be provided in the Executive Summary and the Introduction so the reader understands how long the various activities extended.

Reviewer Z

Page 8, Table 2.4.

As with the previous table that could be used to distinguish between freshwater and saltwater tests, several of the tables in this section have sufficient space to identify which labs were responsible for each of these activities. Alternatively, this is the type of table that could be included in an Appendix or information included as part of another table. As it stands, the majority of space in this table is taken up with a list of the test methods.

Reviewer Z

Page 9, Paragraph 1

Section 3.0. This section is very well written and has much of the detail that the reviewer was expecting earlier with respect to the selection of referee and participant laboratories. This section should remain here, but it should also be referred to earlier in the text and summarized in the Executive Summary so the reader knows that it is coming. However, Section 3.0 is quite long and detailed. It may be more appropriate to summarize the information here and present full details in an Appendix.

Reviewer Z
Pages 10 -13
Section 3.3. This section is extremely well written and describes the process in detail. It is a necessary section and very helpful to the reader. However, there should be some indication earlier in the text that this information is coming, perhaps a brief synopsis in the Executive Summary.

Reviewer X
Page 11, Paragraph 6
Line 4: were, not was [data are the plural of datum].

Reviewer X
Page 11, Paragraph 6
What is meant by "(particularly for less common test methods)"? Surely what is meant is "more common test methods"?

Reviewer X
Page 13, Paragraph 3
Section 3.4. For such an expensive and important study selecting the "lowest bidder" may not have been smart given some of the problems that resulted with some referee labs. For the referee labs what was required was the most competent, not the cheapest laboratory.

Reviewer Z
Page 14, Table 3.2.
This is the table the reviewer has been waiting for. This summary is needed much earlier in the text and perhaps even in the Executive Summary. One of the reasons this is so mysterious is that the four referee laboratories are identified in the Acknowledgment and the tests are identified in Table 1 of the Executive Summary. However, the reader must wait until Page 14 of the text to determine which labs were responsible for which tests. This information could easily be included in Table 1 as a footnote, in a separate column, or a variety of ways. This would clearly establish testing responsibilities in the first few pages. It could even be included in the Acknowledgment. It could also be included in Table 2.2 or Table 2.3 or Table 2.4. Table 2.4 might even be a good place to include the date when the contracts were awarded to place a temporal context for the testing dates included in that table. In any event, including this information in Table 3.2 on Page 14 is too late to get this information. It may be appropriate to include the table here for redundancy, but the information is needed much earlier.

Reviewer Z
Page 14, Paragraph 1
Section 3.5. It is not clear what is meant by "base study design." More information is presented later to clarify this point, but it should be clarified here or earlier if mentioned previously. See also P7 Section 2.2.4 2nd Paragraph.

Reviewer Z
Page 15, Figure 3.1.
This figure is very informative and should probably come earlier as well. The section is very well written.

Reviewer X

Page 16, Figure 3.3.

Clarify why certain tests were only done by the referee laboratory. Why not simply omit these data since they are not useful to the overall study goals.

Reviewer Z

Page 16, Figure 3.3.

Same as above [as for Figure 3.1], but the reader also should know where the final list of these labs will appear. It may be appropriate to have this list in an appendix or at the end of the report, but it should be included somewhere and this is where the reader should be told where to find that information. Even if EPA chooses not to include the list of labs here, they should at least identify where the information may be obtained by calling a number, website, or email address. Alternatively, this table could also go in an Appendix.

Reviewer Z

Page 17, Paragraph 1

Section 4.1. This section is very well organized, very well written, and presents the necessary detail to understand what is being done. However, there are a few minor grammatical questions. "..real-world sample matrices (i.e., effluent, receiving water...)" seems unduly cumbersome and indirect. Everyone is familiar with effluents and receiving water as possible issues to be evaluated for testing suitability. EPA may want to consider reversing the parenthetical expression to read, ."..suitability of various effluents and receiving water (i.e., real-world sample matrices)." In this scenario, effluents and receiving water are identified immediately as the specific matrices being tested for suitability and the fact that they are "real-world" matrices is only explanatory for emphasis rather than the other way around. EPA could also choose to add this modifier later in the sentence or add a sentence to explain the significance of real-world. It appears to this reviewer that it is more important to identify effluents and receiving water than the fact that they are real-world effluents and receiving water. In the simplest terms, the sentence would read almost as well without real world matrices, but not without effluents and receiving water. It is also interesting that in this section on preliminary testing that there is no mention of the "base study design." One would have to assume that any preliminary testing would have to be a part of the "base study design."

Reviewer X

Page 17, Paragraph 3

Why was copper the only contaminant measured in the marine tests? In all samples? Presumably this is related to the use of Cu for spiking in marine tests, but there are no data re concentrations after spiking. Such data should be provided. Also, in some cases Cu was changed to KCl for marine samples – this should have been measured for consistency.

Reviewer Z

Page 17, Paragraph 3

Section 4.1.1. See previous comment [p. 17, para. 3 comment] regarding the significance of "real-world matrices" and "effluent and receiving water." A number of physical and chemical analyses are listed in this section that are not defined. These terms should also be included in a glossary within this report: alkalinity, hardness, pH, total residual chlorine, total ammonia, dissolved oxygen, total dissolved solids, total suspended solids, total organic carbon, biological oxygen demand, and chemical oxygen demand. In most other sections of this report, the test medium is referred to seawater and the test methods referred to as saltwater. The reviewer believes that this is one of the few times where the word "marine" is used. This is not necessarily an inappropriate usage, but EPA should make the conscious decision to use specific descriptors in certain ways and use them consistently throughout the text. Other EPA documents refer to Freshwater Acute and

Saltwater Chronic in terms of describing both test methods and ambient water quality criteria. The word marine is rarely used and it seems somewhat conspicuous in this context. Most readers would probably also want to know why salinity and copper were measured in these "marine," "saltwater," or "seawater" tests.

Reviewer Z

Page 18, Table 4.1.

This table is necessary, informative, and presented in the appropriate place, but it is also a candidate for the Appendix. Every box in the matrix is checked except four! Footnotes are provided for every box in the Selenastrum row, three boxes in the Champia chronic, and three boxes in the Holmesimysis acute.

Reviewer Z

Page 19, Paragraph 1

Bulleted Items: Accessibility, Historic testing and experience, Characterization, Consistency. EPA seems overly attached to using the phrase "real-world samples" as it appears in every section of bulleted items, including the introductory phrase. As mentioned previously, the main point to be made is that various effluents and receiving waters were evaluated for accessibility. This point should be made once or twice for emphasis, but not as the main point of each sentence throughout the sentence. The fact that they are real-world samples is not as important as the fact that various effluents and receiving waters were evaluated. Since effluents and receiving water are clearly identified, the reader can even assume that they come from the real world. Where else would they possibly come from? It seems as though one would have to say that they were artificial effluents and artificial receiving water to be anything but real world. It is acceptable to be redundant for emphasis, but in this case where the usage is flawed from the start, the redundancy is unnecessary and annoying. The reader needs to hear that effluents and receiving waters were evaluated for these elements, not real-world samples.

Reviewer Z

Page 19, Paragraph 2

Section 4.1.2. "Spiking of the reference toxicant sample was targeted to produce effect concentrations of 50% sample" sounds awkward. It is also not clear what is being spiked and what spiking really means in this context. The two problems appear to be "spiking of" and "produce effect concentrations." What is really being produced is a measurable effect and the sample is simply being spiked. The reviewer suggests something like the following: "Spiking the reference toxicant sample was targeted to produce a measurable effect when diluted by 50%." Spiking needs to be defined and explained more clearly so the reader can better understand what is being spiked, what is being used for the spiking, and what is the intended purpose. A general explanatory sentence or two is required at the very beginning of this paragraph to set the scene for the detailed information that will follow. There should also be some explanation of why different spiking agents were used for all freshwater and some saltwater test methods. There is also a generic problem inherent in this section and in all previous sections in that the saltwater and freshwater test methods have yet to be clearly identified in the previous sections or in any of the tables. The reviewer has suggested that this could be easily accomplished by including those divisions in one of the tables that lists the methods (e.g., Table 2.2) so that the reader will know the different categories before coming to this section. The word "saltwater" is not even used in this paragraph. What follows is a more specific example of this problem: "KCl was used as the spiking agent for freshwater methods, sheepshead acute and chronic methods, and the mysid chronic test method; $CuSO_4$ was used as the spiking agent for silverside acute and chronic test methods, Champia chronic, and Holmesimysis acute test methods." This sentence clearly states that KCl was used for freshwater methods, but to remove any ambiguity it should say "all freshwater test methods." To set the scene for more detail that will come later the sentence should go on to say "and some saltwater test methods." This might be handled best by identifying the saltwater methods in a parenthetical expression as follows: "KCl was used as the spiking agent for all freshwater test methods and some saltwater test methods (i.e., sheepshead acute and

chronic and mysid Mysidopsis bahia chronic). Copper sulfate ($CuSO_4$) was used in the other saltwater test methods (i.e., silverside acute and chronic, Champia chronic, and the mysid Holmesimysis acute). The other problem here, which also occurs in other sections of the text, is that there is no clear distinction between the mysid chronic test method (referring to Mysidopsis bahia) and the Holmesimysis acute test method. This ambiguity is also seen in most of the tables. It appears as though the reader is expected to remember that the chronic mysid test will always refer to Mysidopsis bahia and the acute mysid test will always refer to Holmesimysis costata. The problem is exacerbated the using the common generic name of mysid when referring to the chronic test and the scientific name of mysid when referring to the acute test. This ambiguity could be removed completely by the following: 1) Adding freshwater and saltwater divisions to Table 2.2 and others that lists these tests; 2) Placing the two mysid tests next to each other in the table so that it is easy to see at a glance that there are two mysid tests; 3) Always referring to each mysid by the common name and scientific name in each table and throughout the text; and 4) Add one or two sentences identifying early on that there are two mysid tests but one is chronic and one is acute. These changes would be relatively easy to make and remove much, if not all, of the ambiguity regarding mysid test methods. It might also help to have the test methods listed in order of complexity within the freshwater test methods and the saltwater test methods. For example the freshwater group would list the algal test method followed by the cladoceran test methods and then the fish test methods. Using the same convention, the saltwater group would list the algal test method followed by the two mysid test methods and then the fish test methods.

Reviewer Z

Page 20, Paragraph 2

Section 4.1.3. ."..persistence of toxicity in the real-world samples..." and ."..volume of the spiked real-world samples..." are probably the worst infringement of using "real-world samples" as a substitute for effluent and receiving water. It might be acceptable to use "real-world samples" again as a parenthetical modifier, but this is not the major issue. The real issue is that "the persistence of toxicity in effluents and receiving water was determined in Part 3 of the preliminary testing." In fact, the sentence reads better by changing the subject from real-world samples to effluents and receiving water and reversing the order of the sentence as written. The sentence describing the "referee laboratory" shipping samples "round-trip" to the "referee laboratory" is slightly ambiguous since there were a total of four different laboratories apparently shipping different samples. The ambiguity could be removed with something like the following: "Each sample type that was used in interlaboratory testing was collected, prepared, packaged and shipped by the referee laboratory responsible for that particular sample exactly as described for interlaboratory testing (see Section 6). Referee laboratories shipped each of these samples round-trip back to themselves so each sample would theoretically receive similar handling during the shipping process." It must be acknowledged, however, that since final destinations were different distances from the central shipping area and also probably had different temperatures during that time of year, it seems likely that conditions for each shipment were probably somewhat different. Nevertheless, EPA did its best to account for handling differences, but more detail should be provided on the purpose of "round-trip shipping." This sentence is worded much better in the last sentence of this section on P21: "During interlaboratory testing, referee laboratories again shipped samples round-trip back to themselves and conducted testing simultaneously with participant laboratories." However, this sentence does not identify the purpose of this round-trip shipping either.

Reviewer X

Page 20, Paragraph 3

Clarify what is meant by "round-trip". Did the sample go across the country or merely to the sorting area for the carrier and then return?

Reviewer Z

Pages 21-48

Section 4.2. The supplementary information provided in this section is far too long and detailed to be included in this report. The information belongs in an Appendix. Even in an Appendix, each section is far too redundant in terms of describing each sample type. The wording is almost identical in each section with the exception of the results from each sample. The section could be shortened considerably by the following: 1) Using a subheading of "Preliminary Testing Results - Freshwater; 2) Giving the detailed description of each sample type that were identical across freshwater tests in an introductory paragraph; 3) Adding a summary table to show which sample was used for which test; and 4) Providing only summary results under each subsection. It is possible with this consolidation that the section might fit in the body of the report, but this reviewer still believes that it belongs in an Appendix.

Reviewer Y

Page 21, Paragraph 2

Section 4.2.1.1. The difference in spiking level and interlaboratory
test results makes the accuracy of this *Ceriodaphnia* chronic test difficult to assess.

Reviewer Z

Page 21, Paragraph 2

Section 4.2.1. Preliminary Testing Results should be divided into Freshwater and Saltwater for the benefit of the reader and those not familiar with the different categories of testing. Although the reference toxicant sample is clearly identified as "moderately hard synthetic freshwater," Ceriodaphnia acute and chronic test methods should be listed under a generic heading of "Freshwater Testing" as a reminder to the reader and those unfamiliar with particular groups of tests. This will also help to reinforce the descriptions of reference toxicant sample type, effluent sample type, and receiving water sample type which follow. The fact that these were all freshwater samples is not clearly identified in each section and a major heading would help clarify this as well as the test method as being associated with freshwater. For example, after identifying freshwater as the test medium in the first sentence of Reference toxicant sample type (4.2.1.1), the word freshwater does not appear again in this section. This is another reason for including "Freshwater Testing"as a major category. The reviewer also suggests adding the word "freshwater" to the first sentence of the last paragraph on this page for emphasis. EPA may also want to reverse the order of this sentence as in the previous suggestion as follows: "The Ceriodaphnia freshwater chronic test method produced IC25 values of 323 mg/L during Part 2 of the preliminary testing." As suggested in previous sections, the most important information should be included at the beginning of the sentence. The reader believes that "Part 2 of the preliminary testing" is really the explanatory modifier in this sentence and that the other information is more important and should come first. Potassium chloride is abbreviated and identified in parentheses for the first time on in Section 4.2.1.1 in the first sentence. However, the abbreviated form KCl has already appeared at least three times on pages 19-20. The full name and the abbreviation should appear the first time it is referenced and thereafter the abbreviated chemical name is sufficient. KCl should also appear in the glossary.

Reviewer X

Page 22 Table 4.2

NOAEC should be NOEC

Reviewer Y

Page 22, Table 4.2.

Page 22, Tables 4.2 and similar tables throughout the document. It is very difficult to assess the quality of estimates of LC50 or IC25 without seeing the associated 95% confidence limits. It is difficult to assess variation within and among laboratories without this information being available. Although their inclusion

will make the tables more difficult to read, reporting of these limits is essential. For example, LC50 values are discussed during different stages (Parts) of the study on Page 23 (section 4.2.1.2., Paragraph 2). It is difficult to make judgements about how minimal differences were in LC50 estimate before/after storage unless the uncertainty associated with each LC50 is presented. This is a generic oversight throughout the document. Although the motivation for omitting these metrics of uncertainty was to make the report as clear as possible, the omitted information is needed to make crucial judgements.

Similarly, NOEC values without MSD values are difficult to assess. Assuming that laboratories have the ability to apply a few methods to the WET data, the power of statistical testing among laboratories may vary, i.e., tests vary in power and will contribute to NOEC assignment. (The MSD is reported to the EPA as suggested by g_2 of Table 8.3 (Page 87) so it could easily be included in this report.)

Reviewer Z

Page 22 Tables 4.2 and 4.3.

The tables are well written and formatted, and the data are presented in a clean and precise way. For emphasis, however, this reviewer suggests adding the word "Freshwater" somewhere in the table caption. The preferred format would be to have freshwater in capital letters at the beginning of the caption followed by a colon and then the rest of the descriptive material. Alternatively, the word "Freshwater" could be inserted before the word "acute" in Table 4.2 and before the world chronic in Table 4.3. NOAEC, NOEC, LC50 and IC25 should be defined in the text and in each table where they appear.

Reviewer X

Page 23, Paragraph 1

were, not was.

Reviewer Z

Page 23, Paragraphs 1-3

P23 Section 4.2.1.2. The word "freshwater" does not appear anywhere in the three paragraphs of this section. The section is very well written but needs to be placed in the appropriate context. This reviewer suggests adding ."..a freshwater municipal wastewater treatment plant effluent..." and perhaps including "freshwater" in the title as well. It is somewhat awkward to read all of these details about the municipal wastewater plant effluent without providing the name or location, but this is probably proprietary or the plant asked for anonymity. Nevertheless, it would be easier to read this section if the name and location were provided.

Reviewer X

Page 24, Table 4.4.

Why so many missing parameters? Same question applies to other similar tables.

Reviewer Z

Page 24, Table 4.4.

This table is easy to read and conveys the information in a clear and concise manner. This reviewer is extremely troubled, however, with the wide range of measurements for all of the parameters measured only at the beginning and end of the sampling period; i.e., total dissolved solids; total suspended solids; total organic carbon; biological oxygen demand and chemical oxygen demand. While this is one of the few tables that actually identifies the category of "freshwater test methods," there is no explanation in the text as to why the aforementioned samples were only taken at the beginning and end of the sampling period. It can only be assumed that the samples were not taken because EPA felt that they were unnecessary due to limited variability or because of budgetary constraints. In either case, the reason should be identified here. It appears

that this section may have inadvertently omitted since the reviewer could not even find Table 4.4 identified in the text. Finally, there should be some explanation as to the toxicological and environmental significance of this rather large variability among these parameters, particularly since the purpose of this work is to identify variability in acute and chronic WET test methods. While the TOC, BOD, and COD appear to be well within natural variability at approximately 20%, total dissolved solids and total suspended solids at 33% and >64%, respectively seem very high and approaching levels that could adversely affect toxicity test results and contribute to the variability in results being measured as part of this study.

Reviewer Z

Page 25, Paragraphs 1-3

Section 4.2.1.3. As in previous sections, the word "freshwater" does not appear in any of the three paragraphs on this page the reader is left with the inadequate ."..natural surface water...." This could easily be improved by ."..natural surface freshwater..." or freshwater surface samples..." Again, EPA seems to insistent on "real-world" and "natural" samples as the main cause when the real issues appear to be "effluents," "receiving water," and "reference toxicants. " It would also be helpful if the referee laboratory collecting the samples was identified to tell a more complete story. The reader should not have to flip back to the acknowledgment, see where each laboratory is located and then deduce that EA Engineering in Sparks, MD collected the receiving water samples. Adding this information provides additional continuity and helps provide a context for the sample collection process. It is not clear why this information should not be included here.

Reviewer Z

Page 26, Table 4.5.

Table 4.5 accurately indicates that the water chemistry values are associated with receiving water samples for the freshwater test methods. This table is also identified in the text but technically speaking, the data are worse than for the previous table because the same parameters that showed considerable variability in the previous table; i.e., TOC, BOD, COD, TDS, and TSS were only measured at the beginning of the test. This is worse because this study was supposed to measure WET testing variability and the variability in significant parameters associated with receiving water were not even measured. This increases the uncertainty in the assessment of testing variability. It is not clear how much of the measured variability is really attributable to the testing methods and how much might be due to the variability in both effluent and receiving water physical-chemical parameters.

Reviewer Z

Page 27, Paragraph 1

Section 4.2.2. As suggested previously, the section heading should include the word "freshwater" and probably in all caps at the beginning of the heading for emphasis.

Section 4.2.2.1. The first sentence in this section correctly identifies "moderately hard synthetic freshwater" as a clue to the fact that these are freshwater tests but the only reason this description appears is that it is necessary to explain how the reference toxicant sample type was prepared. In other sections where "real-world" effluent and receiving waters appear, the term freshwater is not used to describe them. The reviewer now sees that perhaps the original intent of EPA was to separate "real-world" samples from "synthetic" samples through the use of these descriptors. Unfortunately, this message gets lost in attempting to interpret the presence of "effluent" and "receiving water" samples. If EPA would really like to retain these descriptors, more of an explanation is required, and much earlier in the report. This reviewer still believes that it is a mistake to rely on those descriptors alone and that it would be better to use "effluent" and "receiving water" as the primary categories. It is OK to use "real-world" and "synthetic" as modifiers for emphasis but more of an explanation is still necessary to better understand the significance of these terms in the EPA context.

Reviewer X
Page 27, Table 4.6.
For this and other tables, there is value in the comparison between Part 4 and IL as these are true measures of variability since the same sample was tested. These data should be emphasized at the end along with the overall variability data. For instance, Table 4.8 indicates a great deal of variability that should be emphasized. (91,2,5-8): I do not agree with the exclusion of the referee laboratories from the final data analyses – data should have been analyzed with and without the referee laboratories.

Reviewer Z
Page 27, Table 4.6.
The table caption should include the term "freshwater."

Reviewer Z
Page 28, Table 4.7.
The table caption should include the term "freshwater." The use of the term "fathead" to describe the chronic testing sounds a little colloquial although the reviewer has used this term frequently and it appears often in other documents. The reviewer has also seen the abbreviation FHM for fathead minnow. This does not appear appropriate for an EPA document however, and is somewhat consistent with other sections with titles that use the genus name "Ceriodaphnia" and the full common name "fathead minnow." Strictly speaking, each section title and each table caption should include both the common name and the scientific name as shown originally in Table 2.2; e.g., Fathead minnow, Pimephales promelas.

Reviewer X
Page 28, Paragraph 1
Line 10: How much less than 90%?

Reviewer Z
Page 28, Paragraph 1
Section 4.2.2.2. This section title should include the term "freshwater."

Reviewer Z
Page 29, Paragraph 2
Section 4.2.2.3. This section title should include the term "freshwater."

Reviewer Z
Page 29, Paragraph 4
P29 Section 4.2.3. This section title should include the term "freshwater." The reviewer suggests the following: "Preliminary Testing for the Selenastrum Freshwater Chronic Test Method."
Section 4.2.3.1. This section title should include the term "freshwater."

Reviewer X
Page 30, Paragraph 2
A 17% change is a lot higher than previously noted, yet the sample is still deemed "reasonably persistent and appropriate for use in the interlaboratory study". What was the cutoff at which variability would have been too high for a sample to be used?

Reviewer Y

Page 30, Paragraph 2

Section 4.2.3.2. "The toxicity of the sample was determined to be **reasonably persistent and appropriate for use** in the interlaboratory study." Here and elsewhere, it is difficult to follow the process for deciding what was or was not acceptable as a spiked sample. The conclusion that the spike was "reasonably persistent" (based on a 17% difference on repeated testing) is difficult to compare to statements in the report about the other spiking results. For example (Page 40, paragraph two), a temporal difference of 3.5% was described as "persistent" toxicity. With this format for presenting results, it is difficult to understand what exactly is the difference between "persistent" and "reasonably persistent and appropriate for use." What would be unacceptable or inappropriate? The 48% change reported in paragraph one on page 39 was assumed to be insufficiently stable but the stability of the modified spike was not assessed. The modified, spiked sample was used in the exercise, i.e., was judged appropriate for use without conclusions about its persistence. Page 40 reports a 30.2% difference in IC25 for a spiked sample, leading to the conclusion that that was an acceptable spike for the interlaboratory sample. Clearer and more formal presentation of spiking decisions would be very helpful.

Reviewer Z

Page 30, Paragraph 2

Section 4.2.3.2. This section title should include the term "freshwater."

Reviewer Z

Page 31, Table 4.8.

This table caption should include the term "freshwater" and the common name of Selenastrum. The reviewer suggests the following: "Results from green alga (Selenastrum capricornutum) freshwater chronic growth preliminary testing" or "Results from the freshwater Selenastrum chronic preliminary testing."

Reviewer Z

Page 31, Paragraph 1

Section 4.2.3.3. This section title should include the term "freshwater."

Reviewer Z

Page 32, Paragraph 1

Strictly speaking, it is more appropriate to designate the concentration of KCl as xxx mg KCL/L rather than xxx mg/L KCL. EPA may want to consider making this change throughout the text.

Reviewer Z

Page 32, Paragraph 2

Section 4.2.4. This section title should include "SALTWATER" as a major heading for this section beginning with mysids. Fortunately, section 4.2.4.1 includes "seawater" because it is another "synthetic" preparation and the term saltwater is necessary.

Reviewer Z

Page 32, Paragraph 3

Section 4.2.4.2. This section title should include the term "saltwater."

Reviewer Y

Page 33, Table 4.9.

Although precision is not being assessed in this study, it would be useful to have nominal or predicted effects levels in this and similar (e.g., Table 4.12) tables.

Reviewer Z

Page 33, Table 4.9.

This table should include the term "saltwater" as a major heading and identify the scientific name of the test species such as "SALTWATER: Results from mysid (Mysidopsis bahia) chronic preliminary testing" or "Saltwater results from mysid (Mysidopsis bahia) chronic preliminary testing" or "Results from mysid (Mysidopsis bahia) saltwater chronic preliminary testing."

Reviewer Z

Page 33, Paragraph 2

Section 4.2.4.3. Interestingly, this section does identify the receiving water sample type as "natural seawater." Although the seawater medium is correctly identified, the term "natural" is used as a modifier instead of the term "real-world" which was used in the introductory material on pages 17-19 of Section 4.1 Preliminary testing plan. This discrepancy gives the appearance that these sections were written by different people with a different view of what is important to distinguish these samples and which descriptors should be used. It is important that EPA determine which attributes are most important to distinguish these sample matrices and how they should be described. As mentioned previously in other sections, this reviewer believes that it is important to: 1) distinguish between the freshwater and saltwater tests as a generic division between the test methods described in this report; 2) use the "real-world" designation for emphasis and to distinguish between "real-world" and "synthetic" sample preparation; and 3) distinguish between "natural" and "real-world" samples if there is a difference; and 4) be consistent between the use of "real-world" and "natural." EPA may have deliberately chosen not to make the distinction between "freshwater" and "saltwater" test methods because it was determined that it was unnecessary or because of possible confusion between the type of test and the type of water used in the effluents. The reviewer believes that it is more important to make this distinction between the types of test for ease in reading and to help the reader conceptualize similarities and differences between the major categories of testing. Furthermore, the other differences in sample preparations such as "real-world," "natural," and "synthetic" should be carefully chosen to select the meaning that EPA chooses to convey. This should not be done in an arbitrary fashion and left to the discretion of the reader to interpret what is meant in the text. It should be stated in a clear and concise manner if possible. If need be, however, it may be necessary for the text to be somewhat more verbose if additional clarification is necessary to avoid uncertainty and ambiguity. For example, why was "real-world" used to describe the samples in the beginning of this section and "natural" used at the end? Was this deliberate or arbitrary? These descriptors must be carefully chosen for clarity and consistency to convey the appropriate meaning.

Reviewer X

Page 34, Paragraph 1

A 30% change (see comment o) above) seems excessive yet there is no comment as to persistence and appropriateness. Spiking was increased but stability was not reassessed? Same comment (35,3,7): a 27% change. Same comment re (39,1,3): a 48% change! Same comment re (44,3): a 22% change and a factor of 2X difference in IC25 values.

Reviewer Z
Page 34, Table 4.10.
This table should include the term "saltwater" and the scientific names of the test species Mysidiopsis bahia and Cyprinodon variegatus.

Reviewer Z
Page 34, Paragraph 2
Section 4.2.5. As in the other section titles the term "saltwater" should appear in the title and the scientific name should also be provided; e.g., "SALTWATER: Preliminary testing for Sheepshead Minnow (Cyprinodon variegatus) Acute and Chronic Test Methods."

Reviewer Z
Page 35, Table 4.11.
This section should include the term "saltwater" and the scientific name of the test species; e.g., "SALTWATER: Results from sheepshead minnow (Cyprinodon variegatus) acute preliminary testing. The full common name should be given for each test species rather than the colloquial "sheepshead." Another possibility for table captions for each of these sections is to divide them into the major categories as follows: "PRELIMINARY SALTWATER ACUTE TESTING: sheepshead minnow (Cyprinodon variegatus) results." Another comment regarding redundancy. The reviewer understands that EPA would prefer to keep this document as clear and concise as possible and avoid redundancy wherever possible. Unfortunately, each table and figure should stand alone and the reader should have all the information necessary to read and understand the tables and figures without referring to the text. This necessitates including the full common and scientific names in each table as well as identification as "Preliminary Saltwater Testing."

Reviewer Z
Page 35, Paragraph 2
Section 4.2.5.2. The section title should include the term "saltwater."

Reviewer Z
Page 35, Table 4.12.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER CHRONIC TESTING: sheepshead minnow (Cyprinodon variegatus) results."

Reviewer Z
Page 35, Paragraph 2
Section 4.2.5.3. The section title should include the term "saltwater."

Reviewer Y
Page 37, Paragraph 3
Section 4.2.6.1 Also Paragraph 1 on Page 39 and Paragraph 2 on Page 40. Unfortunately, the stability and validity of this copper spike cannot be defined based on these results.

Reviewer Z
Page 37, Paragraph 3
Section 4.2.6. The section title should include the term "saltwater" and the scientific name of the test species; e.g., "PRELIMINARY SALTWATER TESTING: Silverside Minnow (Menidia beryllina) Acute and Chronic

Test Methods." The common name should also be consistent throughout the test. In Table 2.2 Menidia beryllina is referred to as the Inland silverside.

Section 4.2.6.1. The section title should include the term "saltwater."
Section 4.2.6.2. The section title should include the term "saltwater."

Reviewer Z
Page 38, Table 4.13.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER ACUTE TESTING: inland silverside (Menidia beryllina) results."

Reviewer Z
Page 38, Table 4.14.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER CHRONIC TESTING: inland silverside (Menidia beryllina) results."

Reviewer Z
Page 39, Table 4.15.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER CHRONIC TESTING: water chemistry results for inland silverside (Menidia beryllina) acute and chronic test methods."

Reviewer Z
Page 40, Paragraph 1
Section 4.2.6.3. The section title should include the term "saltwater." Natural seawater samples and spiking are very clearly described in this section and it is clear that these are seawater preparations. Nevertheless the reviewer still suggests including seawater in each of the subheadings. If EPA chooses to include "natural" as a modifier in these titles that would be OK too.

Reviewer X
Page 41, Paragraph 1
There is no such thing as "single-laboratory validation"!!!! Same comment (for page 45,paragraph 1,Lines 6-8).

Reviewer Z
Page 41, Paragraph 1
P41 Section 4.2.7. The section title should include the term "saltwater" and the common name of the test species; e.g., "PRELIMINARY SALTWATER TESTING: red macroalga (Champia parvula) Test Method."

Reviewer Z
Page 41, Table 4.16.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING: water chemistry results for inland silverside (Menidia beryllina) test methods."

Reviewer Z
Page 42, Paragraph 1
Section 4.2.7.1. The section title should include the term "saltwater." The use of natural seawater is properly described in this section. This reviewer initially suggested using the term "saltwater" in reference to test methods because EPA has used this term in other documents to describe toxicity test results and criteria (e.g., ambient water quality criteria). There is a subtle distinction between the use of the term "saltwater" and "seawater" and EPA has used the correct term here and elsewhere in describing "natural" versus "synthetic" seawater. This brings up the other issue regarding "real-world" versus "natural." EPA should review these issues and then determine the appropriate usage for each section and still maintain the consistency necessary for this document.

Reviewer Z
Page 42, Table 4.17.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING RESULTS: red macroalga (Champia parvula) reproduction test." It may also be helpful to indicate that this is a reproduction test. Again, each table and each figure should stand alone without having to refer back to the text for the generic test type.

Reviewer Z
Page 43, Paragraph 1
Section 4.2.7.2. The section title should include the term "saltwater."

Reviewer X
Page 43, Paragraph 2
Define "relatively consistent" given the very large differences noted in the rest of the paragraph. The effluent was not "relatively consistent" by any reasonable definition I can think of!

Reviewer Z
Page 43, Table 4.18.
The table caption should include the term saltwater and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING: water chemistry results of the effluent sample source for red macroalga (Champia parvula) chronic test method."

Reviewer Z
Page 43, Paragraph 3
Section 4.2.7.3. The section title should include the term "saltwater."

Reviewer X
Page 44, Paragraph 2
Why was an explanation not sought? This level of variability is incredible to say the least. This needs to be investigated further.

Reviewer X
Page 44, Paragraph 3
I am confused by the logic in cases where the IC50 was less variable than the IC25 and was thus used for determining the correct spiking level. This seems to me to be a case of ignoring a possible problem rather than trying to resolve it. I note in this regard that IC25 values are important in regulatory use, often more important than IC50 values.

Reviewer Z
Page 44, Table 4.19.
The table caption should include the term "saltwater" and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING: water chemistry results of the receiving water sample source for red macroalga (Champia parvula) chronic test method."

Reviewer Z
Page 45, Paragraph 1
Section 4.2.8. The section title should include the term "saltwater", the appropriate common name and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING: mysid (Holmesimysis costata) acute test method." As mentioned previously, there seems to be a disconnect between descriptions of referee lab responsibilities and what they actually did for each test. The omission of the name of the referee lab actually conducting this phase of the study leads to a discontinuity in the thought process of the reader. Each section should name the referee lab and tell what they did to help complete the story. In this case, if it was MEC that did the work, why not say so? Is there some reason for not identifying the lab responsible? This information is given on page 14 so why not here for clarity. Each of the sections should include the name of the referee lab responsible for that portion of the work.

Reviewer Z
Page 45, Table 4.20.
The table caption should include the term "saltwater" and the appropriate scientific name; e.g., "PRELIMINARY SALTWATER TESTING: Results from mysid (Holmesimysis costata) acute testing."

Reviewer Z
Page 45, Paragraph 3
Section 4.2.8.1. The section title should include the term "saltwater."

Reviewer Z
Page 46, Table 4.21.
The table caption should include the term "saltwater" and the appropriate common and scientific names; e.g., "PRELIMINARY SALTWATER TESTING: water chemistry results of the receiving water sample source for mysid (Holmesimysis costata) acute test method."

Reviewer Z
Page 46, Paragraph 2
Section 4.2.8.2. The section title should include the term "saltwater." In this section and others within this report, EPA seems to go out of its way not to identify the referee lab conducting the work. It would be relatively easy in this section, as in others to simply identify the lab in parentheses after identifying it as a referee laboratory. If this is done, it also helps complete the story for the reader and place the information in appropriate context. For example, it helps to explain why MEC collects its seawater in San Francisco Bay and why Ogden Environmental collects its seawater from Scripps. This reviewer cannot understand why this information is not included. It seems like a logical way to present the information and complete the story for the reader.

Reviewer Z
Page 47, Table 4.22

The table caption should include the term "saltwater" and the appropriate common and scientific names; e.g., "PRELIMINARY SALTWATER TESTING: water chemistry results of the receiving water sample source for mysid (Holmesimysis costata) acute test method."

Reviewer Z

Page 47, Paragraph 1

Section 4.3 The reviewer suggests adding "freshwater", and the appropriate common and scientific names as modifiers in the first sentence of the first paragraph to identify the Selenastrum test category and the kind of test it is; e.g. "For the freshwater green alga (Selenastrum capricornutum) chronic test method using growth as an endpoint..." This additional detail accomplishes several things. 1) It differentiates between a saltwater and a freshwater test. 2) It provides the common name for those who do not realize that Selenastrum is a red macroalga species. 3) It gives the species name for clarity. 4) It indicates that this is a growth test. This latter point in itself, is enough to explain the problem in the preliminary testing, although EPA should probably explain it more directly. Most investigators realize that there is more variability in an algal growth test than in survival endpoints. This is not surprising and EPA probably has additional data and references to support this explanation. Since the purpose of this study is to evaluate variability in the test methods this seems to be an important point to be made here. In other words, rather than simply saying that this was a problem encountered in the preliminary testing, this reviewer believes that the "problems" go to the heart of the variability in that particular test.

Reviewer Z

Page 47, Paragraph 2

Section 4.3. The reviewer suggests adding "saltwater" and the appropriate common and scientific names as modifiers in the first sentence of the first paragraph to identify the mysid and sheepshead general test category and the specific kind of test it is; e.g., "For the saltwater mysid (Mysidopsis bahia) chronic and sheepshead minnow (Cyprinodon variegatus) saltwater acute and chronic test methods...." These designators are even more important in this section than in other sections because there are two mysid tests being evaluated and at first glance it is not clear which one has created the problem. It could be argued that those familiar with these tests would immediately understand that it had to be Mysidopsis bahia because the Holmesimysis costata test method is for acute testing only. Nevertheless, it is not clear that readers of this report will be limited to those intimately familiar with the test protocols, particularly if other acceptable chronic protocols are developed for Holmesimysis costata. This uncertainty could also be helped by reformatting Table 2.2 to accentuate the differences and similarities among the tests. For example, this reviewer has already suggested shading and grouping the freshwater and saltwater tests to distinguish them at a glance. The reviewer also suggests a separate column to separate and to draw attention to the endpoints and their similarities and differences. It would also be helpful to group similar species together. For example, if the two mysids were placed in successive rows at the end of the table, it would draw attention to the fact that there were two mysids being tested and that one was acute and the other chronic. Separated by the distance they are this point is not obvious from a quick glance at the table. As in the previous paragraph, the referee laboratory should be identified to provide the appropriate context for the tests and to help the reader understand what each lab was doing. Also as in the previous section, it appears as though the problems encountered are really a reflection of the variability in test methods being addressed in this study. Therefore, this reviewer believes that these issues should be discussed in terms of their significance to variability in test methods rather than problems in accomplishing the work. This applies if it is variability in survival or growth rates or variability in the performance of reference toxicants. If any of those factors affect the test variability, they should be discussed in this context. It is somewhat surprising that there were these difficulties with $CuSO_4$ (is copper sulfate identified previously in the text? If not, it should be.) if the referee laboratories were as experienced as suggested in the original description. This document should explain more about the problems in using $CuSO_4$ as a reference toxicant and why it was originally selected. Furthermore, switching to KCl does not seem adequately justified either. The reader is troubled by passing these issues off as "problems" when they appear to be related to test variability. More justification should also be provided regarding the variability in Holmesimysis populations over temporal and spatial scales. If this was really an anomaly, it should be described as an anomaly. If this was an example of inappropriate assumptions about the availability of this species in sufficient numbers EPA is obligated to explain what really happened with respect to the decision-making process and uncertainties going into this study. Finally, there should be some explanation for the low control survival. If these laboratories were chosen for their experience and expertise in these particular test methods, more data should be presented with respect to control survival they have measured over the years and a determination on whether or not this was an anomaly or a regular occurrence. This also has a bearing on test variability and the purpose of this study.

Reviewer X

Page 49, Paragraph 1

Note here that this was not achieved in all cases.

Reviewer Z

Page 49, Paragraph 1

Section 5.0. This is a good introductory paragraph to introduce the subject. However, the last sentence in this introductory paragraph would also fit as a first sentence in the opening paragraph of section 4.0 on Preliminary Testing. The wording in this sentence is more clear and concise than many of the sections in Section 4.0 that attempt to describe Preliminary Testing: "Preliminary testing was conducted to validate the selection of real-world samples and spiking concentrations and confirm that each sample produced the

targeted effect and was appropriate for use in the WET (Variability) Study." One can also see the potential ambiguity in "sample preparation" as to whether this was sample preparation for the Preliminary Testing or sample preparation for the Final Definitive Interlaboratory Study. This issue should be clarified here and in other sections of the Final Report.

Reviewer Z
Page 49, Paragraph 2
Section 5.1 Title. This is the only section in the entire report to this point that correctly identifies a section as covering freshwater methods. It suggests to the reviewer that this section was written by someone other than those writing the previous sections, particularly the section on preliminary testing. Each of these sections also appears more clear and concise than the previous sections.

Reviewer Z
Page 49, Paragraphs 2 and 3
Section 5.1. The two paragraphs indicate that "plastic carboys" were used to hold the samples and these carboys were "cleaned and rinsed" prior to adding reagents. More detail is needed here to describe exactly how the carboys were "cleaned and rinsed" as well as defining the type of plastic carboys; e.g., polyethylene, polycarbonate, etc. Alternatively, EPA could cite another methods document as cited for reagents (Section 7 of the WET method manual), but it would be better to briefly describe the method and add one additional descriptor for the type of plastic.

Reviewer Z
Page 49, Paragraph 2
This section contains the formatting that this reviewer has suggested for previous sections and suggests that this section was written by someone else. As suggested for previous sections, the link between freshwater methods and the specific tests evaluated here is not really mentioned, described, or shown in tabular form until pages 49-50. This needs to be established much earlier. The text states that ."..the bulk blank sample was properly mixed and aerated for at least 24 hours prior to removing aliquots..." The reader is not given any information on what proper mixing is and how it is accomplished. It is also not clear why the samples are aerated. Is this part of the mixing process, designed to inhibit bacterial degradation, or is there some other reason? This point needs clarification because under the marine methods there is no mention of aeration. The reader may wonder why aeration is necessary in the freshwater methods and not in the marine methods. (See 1st paragraph P53).

Reviewer Z
Page 49, Paragraph 3
This paragraph is well formatted and provides the type of detail that should have been provided in previous sections. Each sentence clearly and concisely identifies "all freshwater test methods," "reference toxicant sample types," and explains differences among the tests. This summary language should appear much earlier in the text. The term "ampule" is used to describe the container used to hold the reference toxicant but "ampule" volume or container composition is not described. More detail is needed here.

Last 2 Sentences: The phrase "distribution to participant laboratories" is repeated in successive sentences and should be rewritten; e.g., "The bulk effluent sample for each test method was mixed thoroughly prior to spiking, following spiking, and prior to removing aliquots for distribution to participant laboratories. These bulk effluent samples were stored in the dark at <4°C until shipment to those laboratories."

Reviewer X
Page 50, Table 5.1

Clarify that Table 5.1 represents nominal not actual values.

Reviewer Z

Page 50, Table 5.1

This is probably the only table in the report to this point where the caption correctly separates and identifies "freshwater methods." The only suggestion for the caption from this reviewer is that the study should be referred to as the "WET Variability Study." Obviously there have been many studies on WET over the years, but this is probably one of few studies on WET variability. It would be preferable to include the common names and the full scientific names. However, if that is not possible, the shortened version could be justified on the basis of this being a summary table and if the other descriptors make the table more difficult to read. Finally, the reviewer believes it is necessary to identify the referee laboratory for each of these tests to complete the story being told. It would be relatively easy to add another column with an abbreviation for each lab, some of which already have an abbreviated name; e.g., EAE (EA Engineering), OEE (Ogden Environmental Energy), MEC (MEC), and ESI (EnviroSystems).

Reviewer X

Page 51, Paragraph 1 and Table 5.2

Line 2: are, not is. Table 5.2, same page: reconstituted, not recon-stituted.

Reviewer Z

Page 51, Paragraph 1

The text states that "Submersible pumps were used to circulate and homogenize sample between individual containers." The word "sample" should either be "samples" to indicate several samples in individual containers or "the sample" between individual containers. Strictly speaking if there were more than two containers the sentence should probably read "among" containers. The other issue has to do with the use of the term "homogenize," particularly in view of the phrase "properly mixed" in the previous two paragraphs. This reviewer assumes that submersible pumps were used for the larger samples because of the volume required. However, "properly mixed" and "homogenized" are similar and could be used interchangeably under many conditions. The report should distinguish between homogenizing and mixing and state the reason if it is related to sample volume. The same is true for the use of submersible pumps versus aeration. Aeration and submersible pumps are often used for mixing. More detail should be given so the reader can understand what is being done and why it is being done using the methods identified in the text.

Last 2 Sentences. The phrase "distribution to participant laboratories" is repeated in successive sentences and should be rewritten; e.g., "The bulk effluent sample for each test method was mixed thoroughly prior to spiking, following spiking, and prior to removing aliquots for distribution to participant laboratories. These bulk effluent samples were stored in the dark at <4°C until shipment to those laboratories."

Reviewer Z

Page 51, Table 5.2.

This table should include "freshwater" and "saltwater" test methods to clarify that it contains each as well as shading of one or the other category to quickly distinguish between the two categories as suggested in previous sections. It is not clear why Champia and Holmesimysis are not included in this list. The text should clarify why they are not present. The term "recon-stituted" is unnecessarily hyphenated in the first row of this table, probably a formatting remnant from a previous version.

Reviewer X
Page 52, Paragraph 1
Line 2: are, not is.

Reviewer Z
Page 52, Paragraph 1
P53 Section 5.2 Title. This is the only section in the entire report to this point that correctly identifies a section as covering saltwater methods. It suggests to the reader that this sections was written by someone other than those writing the previous sections. Each of these sections also appears more clear and concise than the previous sections. Interestingly, while previous sections have used "saltwater' and "seawater" to describe various elements of the study, the reviewer believes that this is only the second time that the word "marine" has been used (see also P17). Not that this is necessarily bad, but there should be some justification as to why it would be used here for the first time. The report really needs to be more consistent than that unless there is a specific reason for using the descriptor "marine" to distinguish this group of tests at this point in the report.

Reviewer Z
Page 52, Tables 5.3 and 5.4.
These tables should be shaded and coded to indicate which of the methods are freshwater and which are saltwater.

Reviewer Z
Page 53, Paragraph 2
Section 5.2.1 Title. Since "Marine Methods" is clearly indicated as a section heading, it may not be necessary to include "marine" in each of the subsections for additional clarification. The reader still feels that "saltwater" would be more appropriate since EPA seems to have used this designator more frequently in ambient water quality criteria documents. The reviewer still recommends that the complete common and scientific names be included in each section title for clarity; e.g., "Mysid (Mysidopsis bahia) Chronic and Sheepshead Minnow (Cyprinodon variegatus) Acute and Chronic Test Methods."

Reviewer Z
Page 53, Paragraphs 2 and 3
Section 5.2. The two paragraphs indicate that "plastic carboys" were used to hold the samples and these carboys were "cleaned and rinsed" prior to adding reagents. More detail is needed here to describe exactly how the carboys were "cleaned and rinsed" as well as defining the type of plastic carboys; e.g., polyethylene, polycarbonate, etc. Alternatively, EPA could cite another methods document as cited for reagents (Section 7 of the WET method manual), but it would be better to briefly describe the method and add one additional descriptor for the type of plastic.

Reviewer Z
Page 53, Table 5.5.
As suggested in several other places, the reviewer recommends referring to this work as the "WET Variability Study."

Reviewer Z
Page 55, Paragraph 1
The language used in this paragraph is more descriptive than in other sections in that it clearly identifies what was done for the freshwater methods described previously (properly identifying Section 4.2.1.2 for reference)

and then describing what was done for the saltwater tests. The only descriptor missing here is the word "saltwater" as a generic category for the tests listed in this section. As in the previous section, "homogenization" should be described in more detail relative to "properly mixed" and the type of plastic in the plastic carboys identified. The purpose of the submersible pumps should also be identified relative to "properly-mixed" in previous sections. The reviewer also believes that this is the first time that Forty Fathoms® (note: a registered trademark symbol should appear after all product names, assuming that Forth Fathoms is a registered trademark) artificial sea salts is described as "bioassay grade" although Forty Fathoms® artificial sea salts are mentioned several times in preceding sections (e.g., P34) they are not identified as "bioassay grade." If "bioassay grade" sea salts are used throughout, they should be identified as such in each section or identify "bioassay grade" sea salts in a glossary and only refer to them as Forty Fathoms® artificial sea salts throughout the text with the understanding that they are all the same sea salts. The phrase "distribution to participant laboratories" is repeated in successive sentences and should be rewritten; e.g., "The bulk effluent sample for each test method was mixed thoroughly prior to spiking, following spiking, and prior to removing aliquots for distribution to participant laboratories. These bulk effluent samples were stored in the dark at <4°C until shipment to those laboratories."

Reviewer Z

Page 55, Paragraph 2

The reviewer believes that there is not enough detail regarding the collection site in this paragraph, even with the citation for Section 4.2.4.3. Perhaps the sentence regarding timing of collection should be moved to Section 4.2.4.3. Alternatively, the remaining details in this paragraph could all be included in Section 4.2.4.3. The reader understand the dichotomy between providing enough detail in each section to make sense without being too repetitious versus giving the bulk of the information in previous sections and then constantly referring to those details with a citation. EPA has done a reasonably good job of balancing those factors in this report, but they may want to consider some formatting changes to group more of the necessary detail in generic sections provided previously versus more detail in the sections to follow.

Reviewer Z

Page 55, Paragraph 3

Section 5.2.2. The reviewer believes that this is the first time that "synthetic seawater" is mentioned without citing Forty Fathoms artificial sea salts. Perhaps this is OK, but EPA may want to check each of these sections where "artificial" appears to ensure that the proper information is being provided at the appropriate level of detail. "Packaging and distribution" appears to be another new phrase that appears for the first time in this section. Actually, this is a shorter and perhaps easier to understand "jargon" than language used to describe this procedure in other sections. It is not necessary for this language to be consistent throughout the text but the differences make the reviewer wonder why it was done this way in one section and a different way in another section. EPA should make the conscious decision for specific language in each section and understand why a particular language was chosen.

Reviewer X

Page 56, Paragraph 1

Last line: For how long were they stored in the dark? Ensure clarity here and elsewhere in the report.

Reviewer Z

Page 56, Paragraph 1

See comments on "synthetic seawater" above [p. 55, para. 3]. See comments on plastic containers above. See comments on ampules above. See comments on repetition of "distribution to participant laboratories" above [p. 55, para. 1].

Reviewer Z

Page 56, Paragraph 2

Regardless of whether or not plastic liners were used, the report should identify the type of buckets that were used; e.g., plastic or metal. Since the lids are identified as being plastic, it might be assumed that the buckets were plastic, but why not say so? It would also be helpful to identify the type of plastic. It is extremely important to identify the type of plastic liner; e.g., food grade, scientific grade, etc, because of the potential for certain plastics to adsorb or absorb chemicals. The description of homogenizing the samples is very clear and the reviewer believes that this is the only place in the text so far where this detail has been provided. This level of detail should be provided in the other sections as well. The text correctly identifies the Forty Fathoms artificial sea salts as "bioassay grade." This should also be identified in other sections. See comments on repetition of "distribution to participant laboratories" above [p. 55, para. 1].

Reviewer Z

Page 56, Paragraph 3

See comments on repetition of "distribution to participant laboratories" above [p. 55, para. 1]. These two sentences are repeated so many times in so many different sections that not only should the repetitious language be corrected, but it may not be necessary at all if the content of these two sentences is mentioned only once in one of the preliminary generic sections so the reader will know that all of the storage and distribution procedures were identical for a particular type of activity.

Reviewer Z

Page 57, Paragraph 1

Section 5.3. Problems identified for the Ceriodaphnia chronic test method and the silverside acute test seem to be part of the WET test variability this study was designed to assess. It seems inappropriate to address these as problems encountered in sample preparation because these issues are relevant to all WET testing. EPA should include a discussion of how these issues affect WET test variability in the context of the study.

Reviewer Z

Page 57, Paragraph 2

In this and other sections the document refers to the "ampule" as if it had the characteristics of the sample; e.g., "One ampule was prepared in deionized water...", "Testing results from the ampule prepared in deionized water...", etc. The ampule was not really prepared in deionized water, it was the sample that the ampule contained that was prepared in deionized water. The word "in" is also somewhat misused in that the way these sentences are written it sounds as if the ampule itself or the sample itself was prepared while being bathed in deionized water. It would probably be more correct to say that the sample was prepared "with" or "using" deionized water.

Reviewer Y

Page 58, Paragraph 1.

The statement is made that EPA-selected labs were randomly assigned to positions 1-9 and non-EPA sponsored labs were randomly assigned to 10-20. It appears that they were arbitrarily, not randomly, assigned positions 1- 9. This is a trivial point with no bearing on the validity of the study but the authors may want to change the wording to make this point clearer.

Reviewer Z

Page 58, Paragraph 1

Section 6.0. This title is very descriptive and easy to understand and appears for the first time on P55 Section 5.2.2. "Packaging and distribution" appears to be another new phrase that appears for the first time in that section. Actually, this is a shorter and perhaps easier to understand "jargon" than language used to describe this procedure in other sections. It is not necessary for this language to be consistent throughout the text, but

the differences make the reviewer wonder why it was done this way in one section and a different way in another section. EPA should make the conscious decision for specific language in each section and understand why a particular language was chosen.

Section 6.1. This section is well written and informative and presented with appropriate detail but "base study design" appears to be another new phrase that appears for the first time here or appeared so early and infrequently that it seemed new to this reviewer and other readers as well. This also leads the reviewer to believe that some re-organization may be necessary to keep some of these thoughts together.

Reviewer Z
Page 58, Paragraph 2
Section 6.2. This section provides the best description so far on what an "ampule" really consisted of but still does not identify what kinds of containers were used as "ampules." It is mentioned that the samples were shipped in the same container style and size for each test method but does not identify the style and size for each test or refer to a table where the information is given.

Reviewer Z
Page 59, Table 6.1.
P59 Table 6.1. As mentioned in most other tables that refer to the "WET Study", the reviewer suggests the following change, "WET Variability Study."

Reviewer Z
Page 60, Paragraph 2
Section 6.3. This section is well written, informative, and presented with appropriate detail. It is not clear, however, that it is necessary to describe a Ziploc bag as a waterproof enclosure using ("i.e."). It may be sufficient to call it a "Ziploc bag" and assume the reader knows that it is waterproof or simply call it a "waterproof Ziploc bag." In any case, since this is an EPA report the registered trademark symbol should probably accompany "Ziploc® bag."

Reviewer Z
Page 61, Table 6.2.
P61 Table 6.2. As mentioned in most other tables that refer to the "WET Study", the reviewer suggests the following change, "WET Variability Study."

Reviewer Z
Page 62, Table 6.3.
As mentioned in most other tables that refer to the "WET Study", the reviewer suggests the following change, "WET Variability Study."

Reviewer Z
Page 63, Table 6.4.
As mentioned in most other tables that refer to the "WET Study", the reviewer suggests the following change, "WET Variability Study."

Reviewer Z
Page 64, Paragraph 1
Section 6.4. This section is a good description of problems encountered in sample distribution.

Reviewer X
Page 64, Paragraph 2
This comment re stability and persistence is disingenuous – not all samples showed similar levels of toxicity over time, as noted above.

Reviewer Y
Page 65, Paragraph 1
Laboratories knew weeks ahead that important samples were coming on a specific day. Representatives of the laboratories came together in a meeting to discuss the process. They likely optimized their culture preparation and were more focused than the ordinary laboratory would be for a routine sample. Perhaps this isn't true but the difference between the Survey's results and those anticipated during routine WET by many laboratories should be discussed.

Reviewer Z
Page 65, Paragraph 1
Section 7.0. This is a good summary section and correctly uses "reconstituted ampule samples" rather than "reconstituted ampules."

Reviewer Z
Page 65, Paragraph 4
Section 7.1(2). The study should be identified as the "WET Variability Study."

Reviewer Z
Page 65, Paragraph 5
Section 7.1(3). It may be more appropriate to hyphenate "Method-specific." Although test "commencement" is acceptable, test "initiation" may be a more appropriate usage. It might be appropriate to use the "@" symbol instead of "at" to be more direct and save a character space. It could even be appropriate to simply provide the temperature and deviation without "at" or "@."

Reviewer Z
Page 66, Paragraph 1
It seems as though temperature monitoring should have been a requirement rather than a recommendation. Specific instructions should be verified to ensure that the correct term is used in this section.

Reviewer Z
Page 66, Paragraph 4
(6). In this context the term "definitive" to describe the tests conducted seems insufficient. Either the term should be defined or described in more detail or refer to the WET methods manuals.

Reviewer X
Page 66, Paragraph 7
Line 3: Insert "the" before "method manual".

Reviewer Z
Page 66, Paragraph 7
(9). This section is well written with excellent detailed description. It even provides a notation that the Agency is going to change the methodology. This is very good.

Reviewer Z

Page 66, Paragraph 8

(10). This section is well written with excellent detailed description. There is even a notation that the methodology required was somewhat different than the WET methods. This is very good.

Reviewer Z

Page 67, Paragraph 1

(11). As mentioned previously, "base study design" should be defined here and elsewhere or give the most appropriate definition in sufficient detail elsewhere in the text and simply refer to that section here.

Reviewer Z

Page 67, Paragraph 2

(12). The phrases "daily observe" and "daily count" sound awkward and may be grammatically incorrect. The reviewer suggests changing the sentence structure to put the verbs at the end of each sentence; e.g., "Laboratories were required to observe mortality and remove dead organisms in each test daily", and ."..laboratories were required to count young and determine the number of broods at each count daily."

Reviewer Z

Page 67, Paragraph 3

(13). The phrase "extreme toxicity" is ambiguous and unclear and the parenthetical expression does not help very much. Furthermore, if complete mortality in all concentrations was a criterion for contacting SCC, then it should be stated as such without the need for the phrase "extreme toxicity." In practice however, the reviewer would have to assume that laboratories were required to contact SCC even if the survival was >0 in all concentrations; e.g., 5%. It is not clear what criterion was necessary for eliciting contact with SCC in control mortalities. Was it only complete mortality? The reviewer would have to assume that laboratories were required to contact SCC even if control survival was >0; e.g., 5%. If there were specific criteria they should be described here. Complete mortality seems too absolute for generic guidance. Perhaps it was if mortality approached 100%. Even then it seems as though the labs would have been given more specific guidance than "extreme toxicity."

Reviewer Z

Page 67, Paragraph 4

(14). The phrase "..fully document the reason for not completing the test in the final report" sounds like the test is part of the final report. This sentence should be restructured to read "...fully document in the final report the reason for not completing the test."

Reviewer Z

Page 67, Paragraph 5

(15). The phrase ."..response of control samples" sounds like the control samples were responding to something. In reality it is the organisms in the control samples that are responding and the text should be changed accordingly; e.g., ."..response or organisms in control samples."

Reviewer Z

Page 67, Paragraph 6

(16). It is not clear what is meant by "method manuals." Does this refer to WET methods manuals? If so, it should be so stated. In most other sections of the report WET methods manuals are referred to. If there are other methods manuals they should be clearly identified.

Reviewer Z

Page 67, Paragraph 8

(18). The first sentence states "Laboratories were required to submit hard copies of all data and statistical analyses." The next sentence states "All bench sheets and raw data, including sample tracking and chemistry analysis data were required." It is not clear how the first sentence is related to the second sentence. The second sentence, because it is separated from the first, implies that the requirements were different; e.g., original bench sheets etc. were required. If the second sentence is simply a subset of the first sentence they should be combined to remove any ambiguity as follows; "Laboratories were required to submit hard copies of all data and statistical analyses, including but not limited to all bench sheets, raw data, sample tracking, and chemical analysis data." It is usually more appropriate to refer analytical measurements as "chemical analysis" rather than "chemistry analysis."

Reviewer Z

Page 68, Paragraph 1

(19). The same question arises in this section in reference to the method manuals. If these manuals are the ones provided by EPA as part of the WET Variability Test, the only modifier that is necessary is the word "provided"; e.g., "Laboratories were required to analyze data in accordance with the statistical programs specified in the method manuals provided." The only point that needs clarification is whether or not there were specific manuals produced for purposes of this test or whether the laboratories only used existing manuals.

Reviewer X

Page 68, Paragraph 2

Line 2: were, not was.

Reviewer Z

Page 68, Paragraph 2

(20). Since the first sentence appears to correctly use the phrase "an LC50", the second sentence should be modified for consistency; e.g., "An NOEC and an LC50 for survival and an NOEC and an IC25..."

Reviewer Y

Page 68, Paragraph 3.

("…therefore, for the purposes of this study, a set of test condition variables were defined by EPA…") The suggestion here is that the flexibility afforded by the WET method protocols may have been restricted. Would this result in less variation than normally expected among testing laboratories?

Reviewer Z

Page 68, Paragraph 3

A question arises in this section with respect to the "method manuals." Although the document clearly identifies the difference between the WET methods and the "method manuals," the reader is still not given enough information here to identify the source or the contents of the "method manuals." In fact, this section describes everything in excellent detail and provides the rationale for the differences in methods without clarifying the "method manuals."

"method manuals" should be clarified in this section as well. ."..standardized for purposes of this " should be sufficient without "the" before purposes.

Reviewer Z

Page 69, Table 7.1.

This is an excellent summary table although the species and common names should be given in the caption. There are only two issues that appear to this reviewer. 1) While the summary indicates feeding of YCT and Selenastrum prior to testing and available food for newly-released young, it does not identify how much of each. A few more descriptive words here might be helpful. 2) The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The only other issue that seems relevant here is that a test acceptability criterion of 90% survival in the controls is listed here. If this is the case, it seems as though SCC should have been notified if survival went below 90% rather than the "extreme toxicity" mentioned on P67 Section 7.1(13).

Reviewer Z

Page 70, Table 7.2.

This is an excellent summary table although the species and common names should be given in the caption. The reviewer noticed that this summary includes the volume of YCT and algal suspension and the rate of feeding is mentioned. Interestingly, the algal species is not identified in this summary. If it is Selenastrum as identified in the acute test, it should be so stated. The only other issue that seems relevant here is that a test acceptability criterion of 80% survival in the controls, 15 or more young per surviving female in the control solutions, and 60% of surviving control organisms producing three broods is listed here. If this is the case, it seems as though SCC should have been notified if results went below these criteria rather than the "extreme toxicity" mentioned on P67 Section 7.1(13). The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used.

Reviewer Z

Page 71, Table 7.3.

This is an excellent summary table although the scientific name should be given in the caption. The volume of food and rate of feeding are specifically identified in this table for the fathead acute test method. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The only other issue that seems relevant here is that a test acceptability criterion of 90% survival in the controls is listed here. If this is the case, it seems as though SCC should have been notified if survival went below 90% rather than the "extreme toxicity" mentioned on P67 Section 7.1(13).

Reviewer Y

Page 72, Table 7.4.

Item 22.; Also Page 78, Table 7.10, Item 22. Given that one could have 20% mortality and still report results in the WET Survey, it important to have control mortalities reported and discussion of how calculations were done when control mortality was present.

Reviewer Z

Page 72, Table 7.4.

This is an excellent summary table although the scientific name should be given in the caption. The volume of food and rate of feeding are specifically identified in this table for the fathead chronic test method. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The only other issue that seems relevant here is that a test acceptability criterion of 80% survival in the controls is listed here. If this is the case, it seems as though SCC should have been notified if survival went below 80% rather than the "extreme toxicity" mentioned on P67 Section 7.1(13).

Reviewer Z

Page 73, Table 7.5.

This is an excellent summary table although the common name and the species name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used.

Reviewer Z

Page 74, Table 7.6.

This is an excellent summary table although the scientific name should be given in the caption. The volume of food and rate of feeding are not specifically identified in this table for the mysid chronic test method. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The only other issue that seems relevant here is that a test acceptability criterion of 80% survival in the controls is listed here. If this is the case, it seems as though SCC should have been notified if survival went below 80% rather than the "extreme toxicity" mentioned on P67 Section 7.1(13). This table very clearly identifies that "Bioassay Grade Forty Fathoms artificial seawater" was used in this test method. If this table had come earlier or this distinction made sooner in the text, a shortened version of the Forty Fathoms artificial seawater could be used in other parts of the text. This could also be a glossary item. In the table however the precise meaning of "Forty Fathoms artificial seawater" remains unclear. In most other sections of the text reference is made to preparing the test medium with Forty Fathoms artificial sea salts. Strictly speaking then, the test method did not really use Forty Fathoms artificial seawater but artificial seawater prepared with Forty Fathoms sea salts. The distinction is necessary because the reader could assume that Forty Fathoms artificial seawater comes pre-mixed and packaged for immediate use. This is a rather subtle distinction and most readers will probably understand completely. However, if other sections of the text describe the preparation in a particular way, the document should probably be consistent throughout to avoid confusion, particularly in the summary table.

Reviewer Z

Page 75, Table 7.7.

This is an excellent summary table although the scientific name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The volume of food and rate of feeding are specifically identified in this table for the sheepshead acute test method. See comments above on Bioassay Grade Forty Fathoms artificial seawater and notifying SCC if survival dropped below 90% in the controls [p. 74, Table 7.6].

Reviewer Z

Page 76, Table 7.8.

This is an excellent summary table although the scientific name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The volume of food and rate of feeding are specifically identified in this table for the sheepshead chronic test method. See comments above on Bioassay Grade Forty Fathoms artificial seawater and notifying SCC if survival dropped below 90% in the controls [p. 74, Table 7.6].

Reviewer Z

Page 77, Table 7.9.

P77 Table 7.9. This is an excellent summary table although the scientific name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The volume of food and rate of feeding are specifically identified in this table for the silverside acute test method. See comments above on Bioassay Grade Forty Fathoms artificial seawater and notifying SCC if survival dropped below 90% in the controls [p. 74, Table 7.6].

Reviewer Z

Page 78, Table 7.10.

This is an excellent summary table although the scientific name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. The volume of food and rate of feeding are specifically identified in this table for the silverside chronic test method. See comments above on Bioassay Grade Forty Fathoms artificial seawater and notifying SCC if survival dropped below 90% in the controls [p. 74, Table 7.6].

Reviewer Z

Page 79, Table 7.11.

P79 Table 7.11. This is an excellent summary table although the scientific name should be given in the caption. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. See above comments on control survival dropping below 90% [p. 74, Table 7.6].

Reviewer Z

Page 80, Table 7.12.

P80 Table 7.12. This is an excellent summary table although the species name and the common name should be given in the caption. It is somewhat unusual to have additional explanatory information in the caption, but in this particular case it draws the readers attention to the point and it is well made. This is probably a good example of where the style manual should be abandoned in favor of what is appropriate for easy readability and understanding. The descriptor "finalized" for interlaboratory testing schedule seems unnecessary. One can assume that the "non-finalized" testing schedule would not be used. See above comments on control survival dropping below 90% [p. 74, Table 7.6].

Reviewer Z

Page 81, Paragraph

Section 8.1. This section is well-written with just enough detail to outline deviations from the protocols and what was expected.

Reviewer Z

Page 82, Table 8.1.

In other sections of this report this reviewer has commented that each table should stand alone and give the full common and scientific names for each species. Since this table appears so late in the report and the reader has seen the combinations many times in the past it may be acceptable not to include all the information here. Alternatively, since there appears to be sufficient space for another column it might be useful to provide breakdown of participant and referee labs that went into each total so the reader has a better feel for the percentages.

Reviewer Z
Pages 82-84, Table 8.2.
This table appears to be a reasonable summary of the information provided from each lab. The first sentence on P84 appears to have a typographical error that may have been a combination of "note" and "neonate" combined to produce "notate." The reviewer believes that the sentence should read as follows; ."..count and note broods..." or ."..count and note number of neonates in each brood...." In either case, "notate" is probably not correct and should be fixed.

Reviewer Y
Page 84, Table 8.2.
There is no reporting of the specific computational techniques used to estimate LC50, IC25 or NOEC values. It is important to have this information because the available methods have different assumptions, e.g., probit estimation of LC50 assumes a log normal model, and the NOEC can be estimated with methods assuming or not assuming normality and homogeneous variance.

Reviewer Z
Page 84, Paragraphs 1 and 2
Section 8.2.1-8.2.2. The information on this page is clear, concise, and easy to read.

Reviewer Z
Page 85, Paragraphs 1
1st sentence: While the term"excursion" can be used to indicate "deviation," "diversion," or "digression," in practice "excursion" is generally used with qualifiers such as "excursions above" or "excursions below" some criteria. This reviewer believes that it would be better to use the original descriptors such as "deviation," "diversion," or "digression" to remove any doubt about the meaning. Alternatively, the text could identify "excursions above or below" some standard.

Reviewer Y
Page 85, Paragraph 2
This paragraph describes failure of reported results to meet criteria that resulted in their being omitted from use in estimation of percentage of [acceptable] test failures. More discussion of these failures to meet criteria would be helpful. This could be inserted into Section 9.1.1. on Page 91. Also, the results of outlier tests should be discussed more in that section. Would a normal application of a WET method by a laboratory have

Reviewer Z
Pages 85-87, Table 8.3.
P85-87 Table 8.3. This table appears to be complete and very detailed.

Reviewer X
Page 86, Table 8.3.
Why is e7 so specific to 2 hours?

Reviewer X
Page 87, Paragraph 1
Line 4: were, not was.

Reviewer Y

Page 87, Table 8.3.

Although I believe that I know what g_5 means, please provide some detail in the report. Some methods can cope with such data, e.g., Williams's test or the trimmed Spearman-Karber method.

Reviewer Y

Page 87, Section 8.2.3.

This section does not give enough detail to judge the variation associated with the computational techniques. It is focused entirely on describing data processing with one piece of software. I would prefer to know about the actual method applied for handling control mortality, generating point estimates (e.g., LC50 or IC25), and producing hypothesis testing-related estimates (NOEC).

Reviewer Z

Page 87, Section 8.2.3.

This section is very well written and provides sufficient detail to convey the message. It is also well-referenced.

Reviewer Z

Pages 88-90, Table 8.4.

This table does a good job of providing examples of how some sample results were affected by EPA guidance on concentration-response relationships. While it could be argued that this table is not necessary, this reviewer believes that it is helpful. However, it may be more appropriate in an Appendix. As a formatting issue, using the double box around the cell contents constricts the last cell on the bottom which makes the text more difficult to read. An extra line should be inserted to provide sufficient space below the characters.

Reviewer Z

Page 91, Paragraph 1

Section 9.1. This section is clear and concise.

Reviewer Z

Page 91, Paragraph 2

Section 9.1.1. The reviewer wonders why the referee laboratory testing was excluded from determinations of test completion rates, false positive rates, and precision for the methods. While the documents cites knowledge of the identity of samples before testing as the reason for exclusion, this reviewer believes that potentially useful information is being ignored. While it may be justified to treat the data differently, it might be useful to calculate the determinations of test completion rates, false positive rates, and precision for the methods to see if there really was a difference. The data could be presented with and without the referee lab results included to show how advance knowledge may have biased or not biased the results. The data might also suggest that the referee laboratories were "better" at conducting the tests. In either case, the reader would probably like to know whether there were any real differences or not. It seems as though it is always better to have more information than less and then use the appropriate caveats and analysis to explain the results.

Reviewer Z

Page 91, Paragraph 3

Section 9.1.2. It is not clear why "Participant laboratories that failed to complete tests due to reasons unrelated to the test methods themselves (i.e., laboratory error) were not included in the test completion rate calculations or statistical analyses." How was a distinction made between "laboratory error" and "test

method error." Based on prequalification criteria, all the laboratories were proficient at conducting these tests and "any" failure could be attributed to "laboratory error" and not the tests themselves. This issue needs to be clarified in the text.

Reviewer Y

Page 91-92

Section 9.1.3. A test either did or did not have a false positive result so a binomial error structure can be assumed. The uncertainty associated with the false positive rate can be calculated based on that assumption. Why wasn't the variance around the rate estimates calculated?

Reviewer X

Page 92, Paragraph 1

Equation: Define "indicating toxicity".

Reviewer Y

Page 92, Section 9.1.4.

The manner in which the censoring was handled results in a bias in the precision estimates for the associated tests. This section is difficult to assess but it seems that, if less than 20% of the observations were censored, a value was substituted for any censored observation, e.g., a data set with >100%, 50%, 100%, 50% would become 100%, 50%, 100%, 50%. This biases estimates of mean and standard deviation. Because the mean and standard deviation are used to estimate CV, the estimated CV is also biased. There are numerous methods ranging from straightforward Winsorization to maximum likelihood methods for producing unbiased estimates from censored data sets.

The simplest approach (Winsorization) can be compared to the substitution method used in the WET Survey report. Assume that 20 LC50 values were obtained: <100, <100, <100, 95, 90, 89, 92, 85, 90, 92, 83, 98, 95, 89, 98, 92, 95, 90, 88, and 92%. If the <100% values are replaced with 100%, and mean and standard deviation estimated for the modified data set, these univariate statistics would characterize the modified set of 20 observations. The estimated mean would be biased and the standard deviation would be too narrow. With this example, the biased mean and narrow standard deviation would be 92.65 and 4.91, respectively. With Winsorization (which carries the assumption of a symmetrical distribution of LC50 values), one avoids this bias by replacing the same number of observations from both tails (the three smallest as well as the three largest) with the nearest (next highest or next lowest) observation value, calculating the mean and standard deviation with this modified data set, and then adjusting the estimates if required. Assuming a normal (and therefore, symmetrical) distribution, the Winsorized mean does not need adjustment but the estimated standard deviation does. Here is an illustration using the above set of numbers.

First, the 20 observations are ranked from smallest to largest: 83, 85, 88, 89, 89, 90, 90, 90, 92, 92, 92, 92, 95, 95, 95, 98, 98, <100, <100, <100%. In this case, the three largest LC50 values are removed and replaced by the next highest value, i.e., <100%, <100% and <100% are replaced by 98%. Then, the three smallest LC50 values are removed and replaced by the next smallest value, i.e., 83%, 85% and 88% are replaced by 89%. The modified data set resulting from these substitutions is 89, 89, 89, 89, 89, 90, 90, 90, 92, 92, 92, 92, 95, 95, 95, 98, 98, 98, 98, and 98%. The mean and standard deviation for this Winsorized data set are 92.90 and 3.61, respectively. Although this Winsorized mean of these 20 observations is unbiased, the standard deviation is biased and needs to be adjusted using the following equation:

where SD = standard deviation estimated for the modified set of observations, i.e., 3.61 in this case; n = the number of observations, i.e., 20 in this case; and v = the number of observations not modified in the data set during Winsorization, i.e., 20-6 = 14 in this case. The Winsorized standard deviation would be 3.61(19/13) or 5.28. So, the "censored" set of twenty observations has an unbiased mean and standard deviation of 92.90 and 5.28, not 92.65 and 4.91. In terms of the CV being used in this report as a measure of interlaboratory precision, the CV from Winsorization is 100(5.28/92.90) or 5.7%, not 100(4.91/92.65) or 4.5%. Details of these methods are provided in sources such as Gilbert (1987) and Newman (1995).

Other methods described in these same sources can also be applied. As an example, maximum likelihood estimation (MLE) can be used and does not require the discarding of observations as done with the Winsorization. Let's use another data set with some observations being <6.25% (left censoring) instead of >100% (right censoring) with a restricted MLE method (Cohen 1959, 1961) to illustrate the approach. The hypothetical data set of 21 observations would be 20, 19, 18, 18, 18, 18, 17, 17, 16, 16, 15, 15, 14, 14, 12, 10, 9, <6.25, <6.25, <6.25, <6.25. Using the substitution method described in the Survey report, the mean and standard deviation would be 13.86 and 4.68, respectively. The resulting CV is equal to 33.7%. To generate maximum likelihood estimates, the mean and standard deviations are first estimated for the observations that are not "<6.25". The results are mean = 15.647 and SD = 3.081 for those 17 observations. Maximum likelihood estimation requires knowledge of the censoring point (e.g., 6.25%), mean and standard deviations for values that are not censored (e.g., 15.647 and 3.081), the total number of observations (n = 21), and number of censored observations (m = 4).

The C in the equation is the point of censoring, that is, 6.25% in this case. The mean () and standard deviations (s) on the right side of the equation are those for the observations that are not censored (i.e., 15.647 and 3.081). The ? in these equations is taken from a table (e.g., Cohen (1961) for right censored data or Appendix Table 1 in Newman (1995) for left censored data) using two easily generated values, h and ?. The h is the number of censored observations divided by the total number of observations, i.e., 4/21 or 0.19 in this case. The simple formula below is used to estimate the ?.

So, ? is $3.081^2/(15.647 - 6.25)^2 = 9.4926/88.1158 = 0.1077$. The corresponding ? is 0.2445.

The MLE estimates of the mean and standard deviation for this data set become 13.35 and 5.56.
The calculated CV is 100(5.56/13.35) or 41.6%. This is substantially larger than the 33.7% calculated using the substitution method described in the Survey report.

Beyond the suggested change in calculations for censored data sets, a bias in discussion of each method's precision could be compromised by the exclusion of data sets with more than 20% of values being censored.

A more formal method should be applied to these analyses, e.g., maximum likelihood estimation as described above.

As a minor aside, the term "censored" is not used correctly here. A censored observation is one who's magnitude is known only to be either less than or greater than a specified value. Here, the substitution of the actual threshold value (e.g., 100%) for the censored observation (e.g., >100%) is identified as producing a "censored" observation. That is not correct use of the term. It is a substituted value for a censored observation, not a censored observation. As an example, the statement is made "These values were censored as 12.5%, 25% and 50%, respectively."

Reviewer Y

Page 92, Section 9.1.4, Paragraph 2

Application of an outlier test to the collection of observations from different laboratories should be done very thoughtfully. The goal is to estimate the variation that occurs with the WET methods as applied routinely by testing laboratories. In reality, a laboratory would report results from a WET method if it met all the QC/QA and method conditions. The laboratory would not look at results from 8 to 19 other laboratories analyzing the same sample and do an outlier test if there appeared to be a problem with consistency with these 8 to 19 other numbers. Because of this and the fact that only 15 of 698 observation were rejected as outliers, the application of the outlier test seems atypical and unnecessary. It has the potential for needlessly biasing precision estimates.

Reviewer Z

Page 92, Section 9.1.4, Paragraph 2

The term "censored" seems a little too sophisticated for simply "rounding up" or "rounding down." The reviewer suggests simplifying the language wherever possible.

Reviewer X

Page 92, Paragraph 4

Why use a dated ASTM manual when more recent ones are available. Even though methodology may not have changed, cite the latest version.

Reviewer Z

Pages 92-94

Equations. These equations were not checked due to a lack of time.

Reviewer Z

Page 95, Table 9.1.

This is another good "example" table but it might be more appropriate in an appendix. Included/excluded in the title row needs to be re-formatted.

Reviewer Z

Page 96, Section 9.2, Paragraph 1

This is an excellent summary of the Ceriodaphnia acute test method results but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears and the reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

Reviewer Y
Page 97, Table 9.6.
Caption. Please check significant figures. The significant figures seem to change throughout the report (e.g., see also Page 103, Section 9.3.2). Please check for consistency.

Reviewer Z
Page 98, Table 9.2.
Ceriodaphnia acute on blanks. This is a good table, but as in the other table with a double line on the perimeter, the bottom line is more difficult to read. An extra line should be inserted to provide more space. It also seems as though each table should stand alone and that the reason for the invalid test from Lab #29 should be given here in the table. The reviewer also wonders why the referee laboratory data is not included here whether or not it is included in the calculations the reader would like to know whether inclusion or exclusion was really justified.

Reviewer Z
Page 99, Table 9.3.
Ceriodaphnia acute on ref tox. See comments above on Table 9.2. Also, why are the referee lab data (2 sets) included in this table but not in the summary statistics? Why are there 2 sets of data?

Reviewer Z
Page 100, Table 9.4.
See comments above on Table 9.2. Also, why are the referee lab data (1 set) included in this table but not in the summary statistics? Why is there only 1 set of data here but 2 in Table 9.3?

Reviewer Z
Page 101, Table 9.5.
See comments above on Table 9.2.

Reviewer Z
Page 101, Table 9.6.
Surprisingly, this table does not have a formatting problem even though the bottom line is in a "bold" font. Perhaps the "bold" font provides extra space around the characters and avoids this problem. The "bold" font also helps set the summary statistics apart from the rest of the table. EPA should consider this format for other tables as well, particularly "bolding" the most important information such as "N", "mean", and "CV."

Reviewer Z
Page 102, Section 9.3, Paragraph 1
See comments above on Section 9.2.

Reviewer Z
Page 102, Paragraph 4
The results presented in this paragraph are startling. "Of the 34 participant laboratories, 24 produced valid results for all samples tested. The 22 invalid tests were concentrated in the remaining 10 laboratories. Of these 10 laboratories, 8 laboratories performed invalid tests on 50% or more of the samples tested. Two laboratories performed invalid tests on all samples tested. This attributed to the relatively low successful test completion rate achieved for the Ceriodaphnia chronic test method in the WET Study." In essence this means that 29.4% of the participant laboratories had significant problems in conducting the tests. Of this 29.4%, 23.5% performed invalid tests on 50% or more of the samples tested, and 5.9% performed invalid tests on all

samples tested. Given that these laboratories were carefully chosen and had to meet certain performance criteria, one would have to assume that they were among the "better" laboratories in the country. If that assumption is correct, then it could also be assumed that "lesser" laboratories would perform more poorly. If that is the case, it appears that there are serious problems with this test method.

Reviewer Y

Page 103, Section 9.3.2, Paragraph 1

[This section] and elsewhere where rates of false positive and test completions are discussed: An estimate of the uncertainty associated with the false positive rate is needed. It could be calculated assuming a binomial error process.

Reviewer Z

Page 103, Section 9.3.2, Paragraph 1

The false positive rate seems reasonable.

Reviewer Z

Page 103, Paragraph 5

The CV values in this section seem high and are comparable to the percentage of labs unable to successfully complete this test. This is another indicator of problems with the test.

Reviewer Z

Pages 105-106, Table 9.7.

This is another example of where inclusion of the results from the referee laboratory might be useful. Again, the reviewer suggests including all the data and then analyzing the data in two different ways, with and without the referee laboratory data. Casual review of this table shows that the results from the referee laboratory are not that much different from the participant laboratories. The data should only be excluded if they are significantly different from the others. Furthermore, additional information could be gained by presenting N, min, max, median for the control mean, control mean (neonates), control CV, and day test termination. This would also provide additional information for comparing the results from the referee laboratory against those of the participant laboratories. The null hypothesis in this case is that there is no difference in referee lab performance in any of these categories, and it really doesn't look like there is. This table also has the same problem as other tables with the double line outline obscuring part of the characters on the last line.

Reviewer Z

Pages 107-108, Table 9.8.

[Same] comments [as] on Table 9.7 except the double underline table outline does not obscure the characters on the last line.

Reviewer Z

Pages 109-110, Table 9.9.

See comments on Table 9.7.

Reviewer Z

Page 111, Table 9.10.

See comments on Table 9.7.

Reviewer Z
Page 112, Table 9.11.
This is another example of a table where "bolding" the "Average" data makes the table easier to read. This approach should be used in other tables as well.

Reviewer Z
Page 113, Section 9.4, Paragraph 1
This is an excellent summary of the fathead acute test method results, but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name. The CV values for the fathead acute test method appear more reasonable than those for Ceriodaphnia.

Reviewer Z
Page 114, Table 9.13.
See comments on Table 9.7.

Reviewer Z
Pages 115-116, Table 9.14.
See comments on Table 9.7.

Reviewer Z
Page 117, Table 9.15.
See comments on Table 9.7.

Reviewer Z
Page 118, Table 9.16.
See comments on Table 9.7.

Reviewer Z
Page 118, Table 9.17.
This is another table where bolding helps the appearance in two ways: 1) the important information stands out and makes the table easier to read, and 2) "bolding" appears to provide more space around the characters and the double underline table format does not interfere with the characters.

Reviewer Z
Page 119, Paragraph 1
This is an excellent summary of the fathead chronic test method results but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

Reviewer Z
Pages 122-123, Table 9.18.
See comments on Table 9.7. This is another example of how "blind rejection" of referee laboratory data has obscured the full story on the results. The footnote indicates that the referee laboratory data were excluded from summary statistics. In this particular case, however, the referee laboratory conducted an invalid test.

Were the results excluded because it was the referee lab or because of the invalid test, or both? "Two tests conducted by the referee laboratory were invalid due to control survival of 65%. These two tests were initiated on the same day, so poor health of organisms used for testing on that day is a likely cause." This also raises the issue this reviewer has already alluded to in Section 9.1.2., How do we know that this unsuccessful test completion was caused by "laboratory error" or "test method error." Compelling evidence has not been provided for either explanation. Furthermore, this "failure," regardless of the cause, would seem to argue that advance knowledge of sample contents had nothing to do with successful test completion or any other parameter since the referee laboratory could not even complete the test. This would seem to be another piece of evidence suggesting that the referee laboratories performed no better or worse than the other laboratories on variability, successful test completion, or false positives. It is not enough to just say that these data were excluded because there is a problem, one has to demonstrate that there is a problem by using the data. Either way, the data should be presented and analyzed. It would be extremely disturbing to learn that there really was no difference in the performance of the referee laboratories and that the data were excluded primarily for legal rather than scientific reasons.

Reviewer Z
Pages 125-126, Table 9.19.
See comments on Table 9.7.

Reviewer Z
Pages 126-127, Table 9.20.
See comments on Table 9.7.

Reviewer Z
Page 128, Table 9.21.
See comments on Table 9.7.

Reviewer Z
Page 129, Table 9.22.
These short summary tables are all enhanced by "bolding" the "Average" information on the last line of each. It also seems to avoid the problem with encroachment by the double underline table outline. Many other tables which are much longer would benefit even further by "bolding" some of the summary information at the end, particularly on tables which cover more than one page or on tables where it may be more difficult to identify the beginning, middle, and end.

Reviewer Z
Page 130, Paragraph 1
This is an excellent summary of the Selenastrum chronic test method results, but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

Reviewer Z
Page 130, Section 9.6.1, Paragraph 2
The successful test completion rate of about 60% is very poor and raises serious questions about the appropriateness of the Selenastrum test methods.

Reviewer Z
Page 131, Section 9.6.3, Paragraph 3
Between laboratory variability ranging from 24.1 to 87.5% for the total variability component is very poor and raises serious questions about the appropriateness of the Selenastrum test methods.

Reviewer Z
Page 133, Table 9.24
See comments on Table 9.7. Another invalid test by a referee laboratory.

Reviewer Z
Page 134, Table 9.25
See comments on Table 9.7. Another invalid test by a referee laboratory.

Reviewer Z
Page 135, Table 9.26
See comments on Table 9.7.

Reviewer Z
Page 136, Table 9.27
See comments on Table 9.7.

Reviewer Z
Page 137, Table 9.28
See comments on Table 9.7.

Reviewer Z
Page 138, Table 9.29
See comments on Table 9.7.

Reviewer Z
Page 139, Table 9.30
See comments on Table 9.7.

Reviewer Z
Page 140, Table 9.31
See comments on Table 9.7.

Reviewer Z
Page 141, Table 9.32
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z
Page 141, Table 9.33
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z

Page 143, Section 9.7, Paragraph 1

This is an excellent summary of the mysid chronic test method results, but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full scientific name. There is a real potential for ambiguity because there are two mysid tests. While it might be expected to remember that the Mysidopsis test is the only chronic mysid test and the Holmesimysis test is the only acute mysid test, this information should be included so there is no doubt which is which.

Reviewer Z

Page 143, Section 9.7.1, Paragraph 2

Measuring a successful fecundity endpoint in only 50% of the tests is extremely poor and raises serious questions about the mysid chronic test method.

Reviewer X

Page 143, Paragraph 4, Lines 3-4

I know this is not the way to do it in the manual, but I believe this laboratory did it "right". How for goodness sake can you calculate growth based on missing organisms?

Reviewer Z

Page 144, Section 9.7.3, Paragraph 1

An IC25 CV for total variance of 41.3% is extremely poor and raises serious questions about the mysid chronic test method.

Reviewer Z

Page 145, Table 9.35

See comments on Table 9.7. Another invalid test by a referee laboratory. Another way to compare the performance of the referee laboratories relative to the participant laboratories would be to calculate the variability, successful test completion and false positives for each test and see if the performance of the referee laboratories was significantly different from the participant laboratories.

Reviewer Z

Page 146, Table 9.36

See comments on Table 9.7.

Reviewer Z

Page 147, Table 9.37

See comments on Table 9.7.

Reviewer Z

Page 148, Table 9.38

See comments on Table 9.7.

Reviewer Z
Page 149, Table 9.39
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z
Page 150, Section 9.8, Paragraph 1
This is an excellent summary of the sheepshead acute test method results, but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

NOTE: THERE NEEDS TO BE A SUMMARY TABLE OF CV VALUES THAT COMPARES TESTS ON A SINGLE PAGE.

Reviewer Z
Page 151, Table 9.41
See comments on Table 9.7.

Reviewer Z
Page 151, Table 9.42
See comments on Table 9.7.

Reviewer Z
Page 152, Table 9.43
See comments on Table 9.7.

Reviewer Z
Page 152, Table 9.44
See comments on Table 9.7.

Reviewer Z
Page 153, Table 9.45
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z
Page 153, Section 9.9, Paragraph 1
This is an excellent summary of the sheepshead chronic test method results, but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

Reviewer Z
Page 155, Table 9.46
See comments on Table 9.7.

Reviewer Z
Page 156, Table 9.47
See comments on Table 9.7.

Reviewer Z
Page 157, Table 9.48
See comments on Table 9.7.

Reviewer Z
Page 158, Table 9.49
See comments on Table 9.7.

Reviewer Z
Page 159, Table 9.50
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z
Page 160, Section 9.10, Paragraph 1
This is an excellent summary of the silverside acute test method results but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name.

Reviewer X
Page 160, Paragraph 3, Last sentence
Was this done within holding times? Unclear.

Reviewer Z
Page 161, Table 9.52
See comments on Table 9.7.

Reviewer Z
Page 162, Table 9.53
See comments on Table 9.7.

Reviewer Z
Page 163, Table 9.54
See comments on Table 9.7.

Reviewer Z
Page 164, Table 9.55
See comments on Table 9.7.

Reviewer Z
Page 164, Table 9.56
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z
Page 165, Section 9.11, Paragraph 1
This is an excellent summary of the silverside chronic test method results but it seems to come far too late in this report. Perhaps some of this information could be included in the Executive Summary. This is the first place in the text where the most important information appears. The reviewer believes that this information should be provided sooner and in a different format for summary purposes. Since this is a summary heading, it might be helpful to include the full common name and the full scientific name. A CV value of 40% seems higher than should be acceptable for this type of test.

Reviewer X
Page 166, Paragraph 3, Last line
I assume these data where pH>9.0 were excluded? If not they should have been.

Reviewer Z
Page 167, Table 9.57
See comments on Table 9.7.

Reviewer Z
Page 168, Table 9.58
See comments on Table 9.7.

Reviewer Z
Page 169, Table 9.59
See comments on Table 9.7.

Reviewer Z
Page 170, Table 9.60
See comments on Table 9.7.

Reviewer Z
Page 171, Table 9.61
This is another excellent summary table where the important "Average" information is bolded. This makes the important information easier to find and read within the table and avoids the problem of the double underline table outline obscuring the bottom line of data in each cell.

Reviewer Z

Page 172, Section 9.12, Paragraph 1

This is the most important section of the entire report, and it is unfortunate that it appears at the very end of the report! Perhaps some or all of these summary tables should appear in the Executive Summary. The reader should not have to wait until the very end to find the most important information. It also leaves the reader "hanging" because there is no overall summary and conclusions to evaluate the significance of the results; i.e., answering the "so-what" question. The summary table that appears in the Executive Summary should be in the Results Summary as well.

Reviewer X

Page 172, Table 9.63

Table needs to have a column detailing which tests would have "passed in a regulatory context". See other comments in this review relative to the differences between the regulatory context ("real life") and this variability study.

Reviewer X

Page 173, Paragraph 1, Lines 8-10

This is too positive, failure to use EDTA really messes up the *Selenastrum* test, period.

Reviewer Y

Page 174, Table 9.65.

Because of the manner in which outliers may have been excluded and censoring was handled, it is difficult to know if this table quantifies the actual precision for normal application of the different WET methods.

Reviewer X

Page 175, Tables 9.65 and 9.66

[These tables] or similar table needs to show min and max range, which as noted above is much more important in a regulatory context than CVs.

3.4 Miscellaneous Comments

Reviewer X

I am confused as to why *Selenastrum* tests were done both with and without EDTA. It is a "no-brainer" that lack of EDTA will result in poorer test results, this has been demonstrated many times. Doing testing with and without EDTA seems a waste of effort and resources.

Since interlaboratory testing could not be done for the *Champia* chronic and *Holmesimysis* acute test methods, referee preliminary testing results for these two tests are not useful.

There is a lot of repetition in the document (e.g., page 6 text). However, whether this is "bad" or not is uncertain. Given that this report will be read by both a technical and a non-technical audience, the repetition may be useful.

Reviewer Y

No additional comments.

Reviewer Z

No additional comments.

3.5 Additional References Recommended For Inclusion in The Document

I am somewhat disappointed in this document simply because its Discussion of the Results is so limited. I had expected/hoped that there would be extensive Discussion putting the findings into context with regulatory usage and the overall usage of WET tests. However, since this was not identified as a requirement of the Study Plan, I cannot fault the authors. I do, however, note the advisability and need for such detailed Discussion at some point and make some points in this regard below.

First, single species toxicity tests (e.g., WET tests) are valuable first tier assessments. Results should then be used as guidance for additional studies such as exposure characterizations to provide insight on causality (e.g., TIEs), or biological assessments to provide data for detecting ecological impairment. As noted by Hall and Giddings (2000) and Chapman (2000), WET tests are the beginning, not the end of evaluations.

Second, variability is a complex issue which needs to be adequately addressed at some point relative to this study. Warren-Hicks and Parkhurst (1992) noted the following based on three interlaboratory studies: (1) variations in percent survival were lowest at the two extremes of test concentrations where sample toxicity was either high or low; (2) variations in percent survival were greatest at concentrations of intermediate toxicity, which is where NPDES toxicity limits are often set; and (3) comparisons of coefficients of variation show that percent survival at a given concentration is more variable than LC50 values. However, the variability of IC25 or LC50 values is not relevant in cases where NPDES permit limits are expressed in terms of percent effect at a particular effluent concentration. The inherent variability of single-species measurement endpoints such as percent effect on survival, growth, and reproduction at a single effluent concentration can result in both false positive and false negative predictions of ecological risk.

Third, false positives are more usefully considered in terms of when toxicity tests provide results indicating impacts when in fact no impacts occur (de Vlaming and Norberg-King, 1999; de Vlaming et al., 2000). This study did not evaluate those false positives, but rather the false positives that occur when a sample that is not toxic indicates toxicity. The fact that *Ceriodaphnia* tests sometimes indicate toxicity in non-toxic water samples has been reported by Moore et al. (2000); such previous findings should be included in some future Discussion of the results of this particular interlaboratory variability study.

Chapman, P. M. 2000. Whole effluent toxicity testing - Usefulness, level of protection, and risk assessment. Environ. Toxicol. Chem. 19: 3-13.

Cohen, Jr., A.C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. Technometrics 1:217-237.

Cohen, Jr., A.C. 1961. Tables for maximum likelihood estimates: singly truncated and singly censored samples. Technometrics 3:535-541.

de Vlaming, V. and Norberg-King, T. J. 1999. A Review of Single Species Toxicity Tests: Are the Tests Reliable Predictors of Aquatic Ecosystem Responses? Office of Research and Development, Duluth, MN. EPA 600/R/97/11. Technical Report.

de Vlaming, V., Connor, V., DiGiorgio, C., Bailey, H. B., Deanovic, L. A., and Hinton, D. E. 2000. Application of whole effluent toxicity test procedures to ambient water quality assessment. Environ. Toxicol. Chem. 19: 42-62.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold Co, New York. Hall, L. W. Jr. and Giddings, J. M. 2000. The need for multiple lines of evidence for predicting site-specific ecological effects. Human Ecol. Risk Assess. 6: 697-710.

Moore, T. F., Canton, S. P., and Grimes, M. 2000. Investigating the incidence of Type I errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia*. Environ. Toxicol. Chem. 19: 118-122.

Newman, M.C. 1995. Quantitative Methods in Aquatic Ecotoxicology. CRC/Lewis LLC, Boca Raton.

Warren-Hicks, W. and Parkhurst, B.R. 1992. Performance characteristics of effluent toxicity tests: Variability and its implications for regulatory policy. Environ. Toxicol. Chem. 11: 793-804.

APPENDIX A

REVIEWER COMMENTS

REVIEWER X

REVIEWER Y

REVIEWER Z