Enclosure 2

Comments on Interlaboratory Study

I. Issues Related to Conduct of the Interlaboratory Study

A. Representative Laboratories

EPA should state in its report that the laboratories participating in the study are not representative of those that routinely perform tests to assess compliance with NPDES permit limitations. EPA imposed extraordinary prequalification requirements on potential participants. The requirements extended well beyond the record keeping and reporting required by the approved test methods. Those laboratories are atypical of the industry as a whole. Therefore, the prequalification process may have biased the sample in a manner that undermines the Agency's ability to generalize its study conclusions to the population of laboratories at-large. EPA should make this clear in its report.

B. Whole Water v. Ampules

EPA should state in its report that it did not attempt to quantify all sources of variability associated with WET testing. In particular, the variability related to collecting and shipping samples was not characterized because EPA elected to ship, in the case of reagent water, only ampules rather than whole water samples. This is contrary to the proposed research design submitted for peer review prior to initiating the study. Using ampules underestimates the gross variability of the entire WET measurement system. EPA needs to highlight for the peer reviewers this departure from the study protocol and to ask their opinion about its impact on test results and conclusions to be drawn therefrom.

C. Blind Study Benefits

EPA should state in its report that it did not perform the study "in-the-blind." All laboratories knew that they were participating in a formal round-robin research study. Moreover, the labs knew that the primary purpose of the study was to re-validate the WET test methods. As such, they had a financial conflict-of-interest and a desire to demonstrate that WET testing is an accurate and reliable measurement tool. Even if the laboratories made no intentional effort to bias their results, it is likely that the "Hawthorne Effect" undermined the data integrity. The Hawthorne Effect is a well documented phenomena whereby participants in a research project alter their behavior as a result of their awareness that they are being studied. The most likely manifestation is that laboratories will exercise a level of uncommon care not typical or representative of normal testing practices.

Although EPA asserts that the labs did not know which samples they received, it was relatively easy to distinguish blanks from reference toxicants (both in ampules) based on conductivity. In fact, EPA and DynCorp used this very technique to properly identify samples that were cross-labeled accidentally. Since each laboratory was required to measure, record and report sample conductivity, there is no doubt that participants were well aware as to which samples were blanks and which were reference toxicants. Since DynCorp demonstrated the ease with which the sample identities could be unveiled, EPA should acknowledge in its report to the peer reviewers that the experiment was not sufficiently "blind" as required by the study protocol.

D. Laboratory Discretion Circumscribed

EPA should state in its report that it limited the options participating laboratories have to satisfy the mandatory test conditions within the approved Part 136 test protocols. By prohibiting laboratories from exercising their routine lawful discretion regarding test conditions, the Agency underestimated the true level of analytical variability likely to arise in the routine practice of performing toxicity tests.

For example, EPA did not allow laboratories to use dilute mineral water for control water. The Part 136 protocol allows the use of dilute mineral water. Therefore, the level of variability introduced by dilute mineral water when used in practice is unknown. Similarly, EPA counted only offspring from the first three broods in the *Ceriodaphnia dubia* chronic reproduction tests. Therefore, the level of variability introduced by including organisms from the fourth and fifth broods, as per the Part 136 protocol, is unknown.

E. Improper Effluent Sample

EPA should inform the peer reviewers that the study results relating to "effluent" cannot serve as the basis for characterizing test method performance on effluents in general. Aside from the reliance on only a single effluent, the problem lies with the type of effluent EPA selected. As discussed in §4.2.1.2. of the report, EPA deliberately selected a POTW effluent that historically demonstrated low or no acute or chronic toxicity. The Agency had to spike the effluent with KCl before using it in the study. Consequently, the report provides virtually no useful information on how WET tests perform in the multitude of complex effluents to which they are being applied in the regulatory process. The Agency cites "consistency" as its justification for selecting a virtually non-toxic effluent. While consistency may be a relevant factor, it can never trump the study objective itself. Moreover, as EPA acknowledges on page 19 of its report, it could have satisfied the consistency objective by using an effluent that "provided a consistent level of toxicity" (i.e., one that consistently exhibits a high degree of toxicity).

II. Issues Related to Data Validation and Test Acceptance

A. Data Quality Objectives

EPA should inform the peer reviewers that it did not comply with the agency's Order #5360.1 (April, 1984) and 5360.1 A2 (May, 2000). The orders required EPA to establish formal Data Quality Objectives (DQOs) prior to analyzing data from the study. DQOs define the characteristics required to make appropriate regulatory decisions based on the data gathered. EPA must define, in advance, the degree of accuracy and precision necessary to accept or reject WET testing for the purpose of assessing compliance with toxicity limits in NPDES permits. Post-facto conclusions, that the precision of WET testing falls within the range of other analytical methods, do not meet the a priori obligation to establish quantitative DQOs. At a minimum, the Agency must define the threshold for unacceptable accuracy or precision for any given WET test method. Even if EPA intended its Study Plan to serve as its DQO document (see Attachment 1 for mandatory requirements), the Agency deviated from that document frequently and substantially, as discussed in other comments below.

B. Completion Rate

1. EPA miscalculated the true rate of successful test completion. The Agency should correct its report accordingly before submitting it to peer review. The Agency excluded only tests that failed to meet formal Test Acceptance Criteria (TAC). EPA did not exclude the numerous tests that failed to comply with non-discretionary conditions in the Part 136 protocol. Regulators would not consider those test to have been successfully completed. EPA also failed to exclude tests that did not comply with the other mandatory procedures required as part of the official EPA study protocols. EPA justified its decision to impose the additional requirements by asserting that they were analogous to those that would be imposed as a condition of the permit or a part of the contract-for-services with a discharger. In most states, failure to perform tests as specified in the permit would invalidate the results and cause the test to be re-run. It should also be noted that failure to perform tests according to the Part 136 protocol was one of the criteria EPA used to disqualify laboratories during the prequalification review. It is inappropriate to change that standard after the study has begun. EPA obviously takes the mandatory criteria seriously. As pointed out earlier, it insisted that limitations on laboratory discretion were essential to accurately characterize true analytical variability. It is inconsistent to ignore those mandatory criteria here.

   Tables 1A & B (Attachment 2), shows the rate of successful completion based solely on tests that complied with all mandatory protocols and met the minimum test acceptance criteria. EPA must revise the study report to reflect the true rate of successful test completion.

Additional information describing how EPA's data invalidation procedures affected completion rate for the marine species is provided as Attachment 3.

2. EPA failed to apply the Minimum Significant Difference (MSD) test acceptance criteria as recommended in their new guidance manual entitled: Understanding and Accounting for Method Variability in Whole Effluent Toxicity Applications Under the NPDES Program, June, 2000. As a result, the agency over-estimates the rate of successful test completion because several tests should have been deemed "unsuccessful" for failure to meet the MSD criteria.

   If EPA elects to use their new guidance to interpret toxicity test results (as when they assess dose-response relationships to identify and exclude false positives), they cannot selectively apply the guidance in order to artificially enhance the perceived robustness of the methods. The Agency should revise its report to the peer reviewers by excluding all tests that fail to meet the recommended MSD criteria. This is especially true because EPA states that application of the MSD criteria is the single most important recommendation provided in the new guidance manuals.

C. **Outlier Removal**

EPA underestimated the true level of analytical variability by excluding results deemed to be "outliers". The technique used by DynCorp to identify outliers (ASTM h & k statistics) is not included in the standard protocols for whole effluent toxicity testing (40 CFR 136). Nor are the techniques included in EPA's most recent guidance for managing analytical variability in WET testing (see Understanding and Accounting for Method Variability in Whole Effluent Toxicity Applications Under the NPDES Program, June, 2000 and Method Guidance and Recommendations for Whole Effluent Toxicity Testing, July, 2000). Removing outliers prior to evaluating method performance based on an interlaboratory validation study will underestimate the analytical variability that will arise in practice. Dischargers must not be deprived of real world variability data, especially since EPA seems to be unwilling to adjust for it in determining the need for, and in setting, WET effluent limits in NPDES permits.

III. **Issues Related to Data Analysis and Interpretation**

A. **NOAEC**

EPA reports only the LC50 endpoint for acute toxicity tests. Many permits issued by state agencies require dischargers to demonstrate compliance with acute limits or to make other regulatory decisions based on the No-Observed-Adverse-Effect-Concentration

(NOAEC). The agency should calculate and include in its report the rate of false positives for all commonly used endpoints, including the NOAEC.

B. **Variability Is Specific To The Dilution Scheme Used In Study**

1. EPA should state in its report to the peer reviewers that the MSDs and interlaboratory CVs calculated in its study are specific to the conditions and/or the level of toxicity inherent in the test samples used in the study. All of the variability estimates of this study are based on tests conducted with a 0.5 dilution scheme. Such a study design tends to produce "all or nothing" responses in replicates, which increases within and inter-test precision. However, EPA advocates the use of dilutions that focus on the dilution of concern and are closer in magnitude to this dilution than that commonly derived from a 0.5 dilution scheme. Therefore, intra- and inter-test variability, particularly for hypothesis test endpoints, will be a function of whether the dilution of concern is located on the tails of the concentration-response curve or on the transitional slope of this curve. The dilution scheme will have less impact if most of the tested dilutions are on the tails of the curve where response varies less with each concentration and changes in concentration (blanks, very toxic effluents and surface waters, marginally-toxic effluents and surface waters). The dilution scheme will have a significant impact if most of the tested dilutions are located on the transitional slope of the curve because this is where response varies the most with each concentration and changes in concentration (typical reference toxicant test design, some effluents and surface waters).

   This comment explains why a number of observations made in the referenced study may be interpreted in other ways. For example, the MSD distribution percentiles reported from this study are similar in many cases to those reported previously by EPA (2000 EPA variability guidance). One might conclude that this data merely corroborates EPA's conclusions regarding averages, percentiles and distributions for MSDs specific to each endpoint. This conclusion is not surprising, however, since the data used by EPA to establish these MSD statistics were based primarily, if not entirely, on tests using a 0.5 dilution scheme. This test design provides numerous "all or nothing" responses, which are characterized by low intra- and inter-treatment and inter-test variability. Low variability among and between treatments, within and between tests, results in low MSDs and inter-test CVs.

   This point further explains why, for the most part, inter-test CVs calculated in the inter-lab study do not vary significantly from that previously reported in EPA's variability guidance. Using a 0.5 dilution scheme will influence inter-test variability by minimizing variation in the test endpoint from test to test, particularly for hypothesis test endpoints. Therefore the inter-lab study may support conclusions previously made by EPA for inter-test CVs, but this is

specific to tests using a 0.5 dilution scheme.

It is most curious why the average MSDs and inter-test CVs calculated for chronic test sub-lethal endpoints do not differ significantly between the non-toxic (blank) and the toxic (reference toxicant, effluents, surface waters) samples. Examples include:

Cyprinodon variegatus Growth

	MSD	CV
Effluent	11	8.1
Receiving Water	9.4	4.7
Reference Toxicant	14.3	6.0
Blank	13	7.9

Mysidopsis bahia Growth

	MSD	CV
Effluent	25	23
Receiving Water	22	17
Reference Toxicant	24	22
Blank	24	20

Blank and reference toxicant MSDs and CVs may not have differed because some of the reference toxicant experiments failed to measure toxicity (apparent test design flaw). However, one would expect the MSDs and CVs for tests based on dilutions showing little variability in response within and between concentrations (non-toxic/marginally-toxic and very toxic samples) to be lower than those based on dilutions showing significant variability within and between concentrations (dilutions residing on the slope of the concentration-response curve). One reason for average MSDs varying little with the level of toxicity inherent to a sample is that the 0.5 dilution scheme used artificially reduces intra- and inter-treatment variability. Variability which is characteristic to responses along the concentration-response curve will be selectively omitted if one does not test many dilutions along the slope of that curve (result of low dilution schemes). This is how the dilution scheme can affect MSDs and conclusions and guidance based on these MSDs.

This same effect would be expected for inter-test CVs since they are also dependent on whether dilutions located on the slope of the concentration-response curve are tested. A good example of this effect can be seen in the LC50s for the acute and chronic sheepshead minnow tests. Regardless of sample or test type, 25% of all tests reported gave identical LC50s plus or minus 0.01% effluent. This can not be due solely to performance because this occurred for different samples and test types. This effect is due primarily to the dilution factor chosen

for the study, which results in high precision even when different samples and tests are used.

Finally, it is not clear why the saltwater tests did not provide any false positive indications of toxicity when testing blank water. One would expect at least a small percentage of blank samples to show some indications of toxicity based on random chance alone. The dilution scheme chosen in the study may have influenced the false positive rate by only exposing one high concentration (100%) of the non-toxic sample to organisms. The next highest concentration would be 50% using the 0.5 dilution scheme. The probability that the dilution water, which presumably would be similar or identical to that used to culture/hold organisms, would mitigate any potential negative effects originating with the blank water would be very high in the 50% dilution. More concentrations approaching the 100% dilution should have been tested using a higher dilution factor. These concentrations would contain much more blank water and would increase the probability of measuring impact due to blank water. Again, the dilution factor chosen in this study may have artificially influenced its results.

In conclusion, it must be stated clearly in the report that the MSDs and inter-test CVs calculated from this study are specific to the conditions and/or the levels of toxicity inherent to the samples tested. Based on these comments, one would conclude that the MSDs and inter-test CV estimates of this study are lower than those expected with dilution schemes greater than 0.5 and when the dilution of concern is located on the transitional slope of the concentration-response curve. This study will underestimate intra- and inter-test variability expected under other test design and circumstances.

C. **False Positives Underestimation**

1. EPA underestimated the true rate of false positives by reporting results only after test data had been analyzed to confirm the presence of a valid dose-response relationship. Given the selected alpha-level of .05, a Type-I error should only be observed once in every 20 tests even before a dose-response is verified. The initial number of false positives was frequently much higher than expected thereby indicating the possibility of systematic bias within the experimental design. EPA should explain in its report to the peer reviewers why the observed number of Type-I errors failed to conform to the limit imposed by the .05 critical alpha level. Table 2 shows the number of Type-I errors before and after considering dose-response characteristics and compares the level to the number of false positives expected in the given sample size. See Table 2 in Attachment 4.

2. EPA underestimated the true rate of false positives by misinterpreting results from the reference toxicant tests. The Agency acknowledged that many laboratories failed to observe toxicity in the chronic *Ceriodaphnia* tests on reference toxicant

samples. The agency asserts, incorrectly, that the failure was due to "differences in test sensitivity between laboratories." In fact, 9 of the 11 most sensitive tests (based on percent minimum significant difference) indicated that the reference toxicant sample was not toxic. Conversely, 9 of the 11 least sensitive tests showed the sample was toxic. On average, tests that indicated toxicity were 50% less sensitive than tests that indicated no toxicity. The difference in test sensitivity was statistically-significant (p=.05). If the reference toxicant sample was actually toxic, then the most sensitive tests would be the most likely to confirm the presence of toxicity. Because that did not occur in EPA's study, and because two-thirds of the laboratories (including the referee lab) reported no statistically-significant difference in *Ceriodaphnia* reproduction, the only logical conclusion is that the sample was not toxic. Therefore, the laboratories observing test failures were, in fact, reporting false positives. Based on data from the non-toxic reference toxicant tests, the true rate of Type-I error exceeds 33% for the chronic *Ceriodaphnia* reproduction method. EPA should provide this information in the report to peer reviewers.

3. EPA failed to adjust its interpretation of all other sample results (blanks, effluent and receiving water) based on unacceptable performance reported for reference toxicant samples. Assuming that the reference toxicant samples were actually toxic, then two-thirds of the laboratories failed to record the presence of that toxicity. In routine practice, if a lab fails a reference toxicant check sample, all tests running concurrently with that check sample must be deemed invalid and rerun. EPA failed to invalidate concurrent samples despite admitted deficiencies in the reference toxicant results. It is likely that EPA's study design underestimated the true rate of false positives because the reference toxicant results indicate the test organisms were uncharacteristically insensitive during the round-robin study. Therefore, they would be less prone to the spontaneous reductions in reproduction caused by the intrinsic biology of test organisms, not water chemistry, that typify false positives.

> *"In toxicity tests, the detection limit is determined by the 'sensitivity' of the test organisms. The sensitivity of organisms to pollutants is an intrinsic quality, which may vary greatly between speciess, but also varies somewhat among organisms within the same species, and is affected by the condition or 'health' of the organisms. Because the sensitivity of the test organisms cannot be 'calibrated' before each toxicity test, the tests must include standards to ensure data integrity. The final rule promulgated today includes the use of standard 'reference' toxicants to maintain that integrity."*

USEPA, <u>Supplemental Information Document for Whole Effluent Toxicity: Guidelines Establishing Test Procedures for the Analysis of Pollutants</u>. October 2, 1985 p. 25 [SID was referenced in original promulgation 10/16/95]

D. **False Negatives**

The extraordinary number of Type-II errors (false negatives) reported for the chronic *Ceriodaphnia* reproduction test indicates a severe defect in study design. From 1992 to 1996, bioassay laboratories analyzed more than 1,300 reference toxicant samples as part of EPA's DMR-QA/QC program. During that entire period, only three Type-II errors were reported. Yet, there were 33 false negatives observed in EPA's Interlaboratory WET Variability Study. The rate of Type-II errors in the most recent study is nearly 300-times higher than historical experience in the DMR QA/QC studies would cause us to expect. Obviously, the observed difference is highly statistically-significant ($p<.0001$). EPA should explain this discrepancy in the report to peer reviewers.

IV. Issues Related to Findings and Conclusions

A. Comparison With Chemical Tests

1. EPA asserts that analytical precision for WET test methods is within the range of variability commonly observed for chemical analyses. The Agency fails to note that the adverse regulatory implications of analytical variability are significantly reduced by using detection and quantification levels to define the limits of the test's valid dynamic range. No such safety factor is applied to bioassay testing and, therefore, the WET test methods are inherently more vulnerable to decision errors. The Agency must assess and inform the peer reviewers of the actual impact of test variability on making a correct compliance determination given the zero-tolerance standard for toxicity and the known coefficient-of-variation for each standard method.

2. The fact that WET testing precision is similar to that of chemical testing is irrelevant if the two methods are implemented differently. Unless EPA intends to develop and apply detection and quantification reporting thresholds for toxicity testing, then the Agency must establish a specific level of accuracy and precision that shall be deemed "unacceptable" when using results to assess compliance with WET NPDES limits. This is particularly important for methods that suffer from an excessive number of incomplete tests or have extraordinarily high coefficients-of-variation (e.g. Minnow growth, *Ceriodaphnia* reproduction, *Selenastrum* cell density, *mysid* fecundity). EPA should clarify this issue in the report to peer reviewers.

3. EPA claims that the accuracy and precision of WET test results is directly related to the abilities and experience of the bioassay lab performing the test. They recommend careful review of laboratory practices before contracting for services.

However, EPA failed to identify the specific performance criteria that it used to distinguish competent laboratories from incompetent ones.

B. Confidence Limits

In its report, EPA should provide a table that describe the confidence limits (or error bands) for the true toxicity value at different coefficients-of-variation. The table would be used to determine whether WET test results were sufficiently definitive to support certification on the Discharge Monitoring Report (DMR). Table 3 in Attachment 5 provides an example of such a tool.

C. PMSD Criteria

The Percent Minimum Significant Difference (PMSD) criteria recommended by EPA do not appear to significantly reduce the number of Type-I or Type-II errors encountered in whole effluent toxicity testing. The Agency was unable to corroborate the claims made in other guidance documents that proper application of the PMSD criteria would substantially improve interlaboratory test precision. Although the coefficient-of-variation may decline, the absolute number of false positives and false negatives remains relatively unchanged. EPA should note this in its report to peer reviewers.

D. Precision Estimates

Section 9.14 of the Report states that "precision estimates were not calculated for data sets that required greater than 20% of data to be censored in this manner." EPA should explain in its report to peer reviewers how precision should be estimated in those circumstances, or why it believes such estimates are not significant.

E. Test Method Withdrawal

EPA should specifically identify for the peer reviews those test methods it does not consider to be sufficiently reliable to support the intended regulatory use. As stated earlier in the document, EPA needs to define its criteria (DQO) for making those determinations.

ATTACHMENT 1

USEPA, Preliminary Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods. Volume 2: Appendix A WET Study Plan (p.4-5).

"Section 2 Six data quality objectives (DQOs) have been identified as necessary to ensure that data produced will meet the study objectives described above. They are..."

"Section 2.1 All data in the study must be generated in accordance with the analytical and quality assurance/quality control (QA/QC) procedures defined in this study plan and the following documents... (list of EPA's method manuals and clarification memos)... The test requirements in Sections 4.4.3 and 4.4.4 of this study plan and the specific instructions provided by EPA will define the allowable flexibility in WET methods included in this study." [p. 4 & 5]

"Section 2.2 Test parameters must meet the range of chemical and physical conditions (such as temperature, hardness, ammonia, conductivity, pH, salinity, etc.) outlined in the appropriate methods manual and as detailed in Section 4.4.3 and 4.4.4 of this study plan."

"Section 4.4.3.1 Physical and chemical properties of the test samples must be in the ranges specified in this study plan, the SOW, specific instructions, and the methods manuals." (p 19)

"Section 4.4.3.5 The specified dilution and control waters (listed in Tables 6 - 17 for each test method) must be used and prepared according to instructions in Section 7 of the methods manuals." (p. 20)

"Section 4.4.3.7 All tests must be conducted using the number of replicates and the number of test containers per concentration as specified in Section 4.4.4. (p. 20)

"Section 4.4.3.8 Test chambers used within a test must be the same type, size, shape, and material. The material must be allowed by the methods manuals for the method used." (p. 20)

"Section 4.4.4 Method-Specific Requirements. The summary of test conditions for the twelve WET methods to be evaluated in the WET Study are provided in Tables 6 - 17. These tables are extracted from the summary test condition tables in the methods manuals and modified to fit the scope of this study. Items that are in bold italic in these tables represent conditions standardized for the purposes of this study where methods manuals provide a range." (p. 21)

Preliminary Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods. Volume 2: Appendix B Participant Laboratory Standard Operating Procedures. EPA 821-R-00-028B (October, 2000).

[Note: all examples taken from SOP for Chronic Ceriodaphnia dubia Survival and Reproduction testing. However, similar "boilerplate" language was used in all the SOPs distributed to all participating laboratories by DynCorp on behalf of EPA.]

"This standard operating procedure document (SOP) is a supplement to the participant laboratory Statement of Work (SOW). This SOP details specific information in the SOW regarding the conduct of the *Ceriodaphnia dubia* survival and reproduction test method in the WET Interlaboratory Variability Study. All modifications to the SOP must be approved by DynCorp prior to implementation." [p. 1]

Section 3.2.1 General Testing Requirements: EPA acknowledges that the promulgated WET methods distinguish between requirements (indicated by the compulsory terms 'must' and 'shall') and recommendations and guidance (indicated by discretionary terms 'should' and 'may'). The latter terms indicate that the analyst has flexibility to optimize successful test completion and when standardization is necessary to assure the predictability of methods to provide reliable results. Additionally, the method manuals allow variations of the methods which are typically fixed in the permit; therefore, for purposes of this study, a set of variables will be defined by EPA (for example, dilution water, salinity, and acute test duration). Any deviation from defined test procedures and/or conditions, such as the necessity to change reagents, equipment, test conditions, or other specified test parameters must be reported to SCC, recorded at the time of modification, noted in telephone logs of communications, documented in a memorandum, and approved by EPA.

"Section 3.2.1.2 Physical and chemical properties of the test samples must be in the ranges specified in this SOP, the SOW, specific instructions, and the methods manuals." [p. 4]

"Section 3.2.1.5 The specified dilution and control waters must be used and prepared according to instructions in Section 7 of the methods manuals." [p. 5]

"Section 3.2.1.7 All tests must be conducted using the number of replicates and number of test containers per concentration as specified in Section 3.2.2." [p. 5]

"Section 3.2.1.8 Test chambers used within a test must be the same type, size, shape and material. The material must be allowed by the methods manuals for the method used." [p. 5]

"Table 2 provides a summary of test conditions that shall be followed for the conduct of all *Ceriodaphnia* survival and reproduction tests performed in the WET Interlaboratory Study. This table is extracted from the summary test condition table in the method manual and modified to fit the scope of this study. Items that are bold italic in this table represent conditions standardized for the purposes of this study where method manuals provide a range..." [p. 6 & 7]

ATTACHMENT 2

TABLE 1A: Method Robustness for Blank Samples

WET Test Species	Protocol	# of Tests Initiated	# of Valid Tests	% of Tests Completed Successfully
Ceriodaphnia dubia	Acute	34	27	79%
Ceriodaphnia dubia	Chronic	35	14	40%
Fathead minnow	Acute	28	16	57%
Fathead minnow	Chronic	25	9	36%
Selenastrum w/ EDTA	Chronic	9	0	0%
Selenastrum w/o EDTA	Chronic	14	2	14%
Silverside minnow	Acute	7	1	14%
Silverside minnow	Chronic	7	2	29%
Sheepshead minnow	Acute	8	2	25%
Sheepshead minnow	Chronic	8	1	13%
Mysid shrimp	Chronic	9	2	22%

TABLE 1B: Method Robustness for Reference Toxicant Samples

WET Test Species	Protocol	# of Tests Initiated	# of Valid Tests	% of Tests Completed Successfully
Ceriodaphnia dubia	Acute	32	15	47%
Ceriodaphnia dubia	Chronic	49	29	59%
Fathead minnow	Acute	39	21	54%
Fathead minnow	Chronic	38	18	47%
Selenastrum w/ EDTA	Chronic	14	0	0%
Selenastrum w/o EDTA	Chronic	14	5	36%
Silverside minnow	Acute	13	5	38%
Silverside minnow	Chronic	14	3	21%
Sheepshead minnow	Acute	8	2	25%
Sheepshead minnow	Chronic	8	1	13%
Mysid shrimp	Chronic	14	4	29%

Note: only tests that complied with all required protocols and met all Test Acceptance Criteria are deemed "valid." Only valid tests may be used to evaluate the "robustness" of the method for inclusion in 40 CFR 136.

ATTACHMENT 3

Following is a discussion of how EPA's data invalidation procedures affected the outcome of its completion rate and variability analyses for the marine methods.

V. *Menidia beryllina*

A. Acute

Success Rate

There was a total of 40 tests conducted including the reference lab tests. EPA concluded that all but 2 tests were conducted successfully. However, EPA did not invalidate the following number of tests for failing to meet the corresponding test requirements or current guidance:

9	DO limits
4	reference toxicant control chart limits
3	salinity test limits
2	sample temperatures exceeded 4^{o}C upon receipt
5	samples > 36 hours old at first use
3	samples > 72 hours at any use
3	minimum # organisms/replicate
4	temperature test limits
5	test incomplete
11	renewals not conducted properly

Regulatory agencies generally reject tests for failing to meet these requirements. Invalidation of these tests leaves 13 tests as valid, which is a 33% test completion rate in contrast to that reported by EPA. This is extremely poor and does not compare well with chemical-specific methods.

Invalidation of tests based on these criteria also results in a much smaller data set for analysis. Only 1 blank test, 2 receiving water tests, 5 effluent tests and 5 reference tests are available for analysis. However, it is clearly inappropriate to characterize test performance using such a small sample size.

B. Chronic

1. **Success Rate**

There were a total of 44 tests conducted including the reference lab tests. EPA concluded that all tests were conducted successfully. However, EPA did not

invalidate the following number of tests for failing to meet the corresponding test requirements or current guidance:

2	DO limits
12	reference toxicant control chart limits
5	salinity test limits
12	sample temperatures exceeded 4^{o}C upon receipt
4	samples > 36 hours old at first use
1	minimum # organisms/replicate
11	temperature test limits
11	test incomplete
20	feeding regime
4	pH <6 or >9

Regulatory agencies generally reject tests for failing to meet these requirements. Invalidation of these tests leaves 8 tests as valid, which is a 18% test completion rate in contrast to that reported by EPA. This is extremely poor and does not compare well with chemical-specific methods.

Invalidation of tests based on these criteria also results in a much smaller data set for analysis. Only 2 blank tests, 2 receiving water tests, 1 effluent tests and 3 reference tests are available for analysis. However, it is clearly inappropriate to characterize test performance using such a small sample size.

2. **Other Comments**

The inter-lab study failed to calculate MSD percentiles and control CVs for both survival and growth endpoints. This information is critical to appending and updating EPA's recently released variability guidance. EPA must conduct this analysis for survival (not conducted in the new guidance document) and update its calculations for growth. Although EPA's guidance does not address inter-lab data for MSDs, there is no reason that this data can not be used to calculate MSD limits for WET methods.

VI. *Mysidopsis bahia*

A. Chronic

1. Success Rate

There were a total of 48 tests conducted including the reference lab tests. EPA concluded that all but 2 tests were conducted successfully. However, EPA did not invalidate the following number of tests for failing to meet the corresponding test requirements or current guidance:

4	DO limits
14	reference toxicant control chart limits
17	salinity test limits
6	sample temperatures exceeded 4^{o}C upon receipt
2	samples > 36 hours old at first use
19	minimum # organisms/replicate
2	temperature test limits
4	information incomplete
16	vessel size
4	aeration not applied appropriately

Regulatory agencies generally reject tests for failing to meet these requirements. Invalidation of these tests leaves 9 tests as valid, which is a 19% test completion rate in contrast to that reported by EPA. This is extremely poor and does not compare well with chemical-specific methods.

Invalidation of tests based on these criteria also results in a much smaller data set for analysis. Only 2 blank test, 2 receiving water test, 1 effluent tests and 4 reference tests are available for analysis. However, it is clearly inappropriate to characterize test performance using such a small sample size.

2. **Other Comments**

The inter-lab study failed to calculate MSD percentiles and control CVs for all three endpoints. This information is critical to appending and updating EPA's recently released variability guidance. EPA must conduct this analysis for survival and fecundity (not conducted in the new guidance document) and update its calculations for growth. Although EPA's guidance does not address inter-lab data for MSDs, there is no reason that this data can not be used to calculate MSD limits for WET methods.

Fifty percent of the tests conducted and validated by EPA failed to provide the minimum proportion of females with eggs in the controls required to calculate the fecundity endpoint. This component of the test completion rate for this test was not presented or discussed by EPA in its report. This completion rate is extremely poor compared to most WET and chemical-specific methods and should be sufficient reason to eliminate this endpoint from the test. The test endpoint should not be retained if sufficient data are not available to document its performance and reliability. The number of tests remaining after thorough validation of the data set provides almost no information on this endpoint. In addition to the endpoint's poor completion rate, this endpoint is undoubtedly the most costly of all of the WET methods because of the time and expertise required to properly identify females with eggs, females without eggs, males and juveniles. Finally, there is no evidence in this study or others that this

endpoint consistently provides more sensitivity than the other two endpoints for any given sample. Research should be conducted to choose another test with a reproductive endpoint that will have an acceptable completion rate, can be determined with reasonable ease, and offers information in addition to that already collected from other endpoints.

VII. *Cyprinodon variegatus*

A. **Acute**

Success Rate

There were a total of 32 tests conducted including the reference lab tests. EPA concluded that all tests were conducted successfully. However, EPA did not invalidate the following number of tests for failing to meet the corresponding test requirements or current guidance:

6	DO limits
13	salinity test limits
3	temperature test limits
5	information incomplete
6	renewals not conducted properly
1	aeration not applied properly
2	vessel size
8	test solution volume

Regulatory agencies generally reject tests for failing to meet these requirements. Invalidation of these tests leaves 8 tests as valid, which is a 25% test completion rate in contrast to that reported by EPA. This is extremely poor and does not compare well with chemical-specific methods.

Invalidation of tests based on these criteria also results in a much smaller data set for analysis. Only 2 blank tests, 2 receiving water tests, 2 effluent tests and 2 reference tests are available for analysis. However, it is clearly inappropriate to characterize test performance using such a small sample size.

B. **Chronic**

1. **Success Rate**

There were a total of 32 tests conducted including the reference lab tests. EPA concluded that all tests were conducted successfully. However, EPA did not

invalidate the following number of tests for failing to meet the corresponding test requirements or current guidance:

6	DO
4	reference toxicant control chart limits
10	salinity test limits
2	sample temperatures exceeded 4^{o}C upon receipt
1	samples > 36 hours old at first use
1	minimum # organisms/replicate
4	temperature test limits
2	aeration not applied properly
3	vessel size
20	feeding regime

Regulatory agencies generally reject tests for failing to meet these requirements. Invalidation of these tests leaves 4 tests as valid, which is a 13% test completion rate in contrast to that reported by EPA. This is extremely poor and does not compare well with chemical-specific methods.

Invalidation of tests based on these criteria also results in a much smaller data set for analysis. Only 1 blank tests, 1 receiving water tests, 1 effluent tests and 1 reference tests are available for analysis. However, it is clearly inappropriate to characterize test performance using such a small sample size.

2. **Other Comments**

The inter-lab study failed to calculate MSD percentiles and control CVs for both survival and growth endpoints. This information is critical to appending and updating EPA's recently released variability guidance. EPA must conduct this analysis for survival (not conducted in the new guidance document) and update its calculations for growth. Although EPA's guidance does not address inter-lab data for MSDs, there is no reason that this data can not be used to calculate MSD limits for WET methods. This degree of uncertainty must be compared only to that of chemical-specific tests in their respective range of quantitation before a defensible conclusion regarding acceptability of method variability can be made.

ATTACHMENT 4

TABLE 2: TYPE-I ERRORS IN METHOD BLANK TESTS

WET Test Species	Test Method Protocol	# of Tests Accepted By EPA	# of Type-1 Errors Expected to Occur in Given Sample Size (alpha = .05)	# of Type-1 Errors Before Analyzing Dose-Response	Percent Type-1 Errors Before Analyzing Dose-Response	# of Invalid Dose-Response Relationships	# of False Positives After Analyzing Dose-Response	% False Positives After Analyzing Dose-Response
Ceriodaphnia dubia	Acute	33	1.6	0	0.0%	0	0	0.0%
Ceriodaphnia dubia	Chronic	27	1.3	3	11%	2	1	3.7%
Fathead minnow	Acute	27	1.3	0	0.0%	0	0	0.0%
Fathead minnow	Chronic	24	1.2	2	8.3%	1	1	4.1%
Selenastrum w/ EDTA	Chronic	5	<1	0	0.0%	0	0	0.0%
Selenastrum w/o EDTA	Chronic	6	<1	2	33%	1	1	16%
Silverside minnow	Acute	6	<1	0	0.0%	0	0	0.0%
Silverside minnow	Chronic	7	<1	0	0.0%	0	0	0.0%
Sheepshead minnow	Acute	7	<1	0	0.0%	0	0	0.0%
Sheepshead minnow	Chronic	7	<1	0	0.0%	0	0	0.0%
Mysid shrimp	Chronic	7	<1	0	0.0%	0	0	0.0%

Note: analysis based on number of "valid" tests and invalid dose-response relationships as defined and reported by EPA.

ATTACHMENT 5

TABLE 3: EXPECTED TEST VARIABILITY FOR IDENTICAL SPLIT SAMPLES

True IC-25	95% Confidence Range at Given C.V.				
	10% CV	**20% CV**	**30% CV**	**40% CV**	**50% CV**
10%	8-12%	6-14%	4-16%	2-18%	ND-20%
20%	16-24%	12-28%	8-32%	4-36%	ND-40%
30%	24-36%	18-48%	12-48%	6-54%	ND-60%
40%	32-48%	24-56%	16-64%	8-72%	ND-80%
50%	40-60%	30-70%	20-80%	10-90%	ND->100%
60%	48-72%	36-84%	24-96%	12->100%	ND->100%
70%	56-84%	42-98%	28->100%	14->100%	ND->100%
80%	64-96%	48->100%	32->100%	16->100%	ND->100%
90%	72->100%	54->100%	36->100%	18->100%	ND->100%
100%	80->100%	60->100%	40->100%	20->100%	ND->100%