COMMENTS OF THE WET COALITION ON EPA'S PROPOSAL TO RATIFY OR WITHDRAW WET TEST METHODS

January 11, 2002

WET COALITION COMMENTS
ON EPA'S PROPOSAL TO RATIFY
OR WITHDRAW WET TEST METHODS

INTRODUCTION

On September 28, 2001, the U.S. Environmental Protection Agency ("EPA") issued a proposed rule entitled "Guidelines Establishing Test Procedures for the Analysis of Pollutants; Whole Effluent Toxicity Test Methods" ("proposed Part 136 rule"). 66 Fed. Reg. 49,794. EPA's proposal seeks to ratify or modify several analytical test procedures previously included in a contested rule promulgated in 40 C.F.R. Part 136. EPA invited public comment, until January 11, 2002, on its proposal. The WET Coalition[1] commends EPA for undertaking this effort. We appreciate the opportunity to submit the following comments and look forward to resolving as many issues as possible prior to publication of the final rule.

The Coalition believes that whole effluent toxicity ("WET") test methods can potentially play a significant role in regulating toxic discharges into the nation's waters. The degree to which WET test methods can successfully contribute toward that goal will depend almost entirely on the availability of: (1) test methods whose performance has been adequately evaluated and found to be acceptable in the appropriate regulatory context; (2) mandatory quality assurance and quality control requirements ("QA/QC") to ensure that the tests perform in practice at least as reliably as demonstrated in the studies conducted to affirm their acceptability; (3) objective rules for interpreting testing results, both in terms of their technical validity and their significance vis-à-vis some regulatory standard; (4) adequate laboratories and laboratory personnel to perform WET testing at a level of proficiency and responsibility commensurate with the regulatory consequences that their test results may trigger; and (5) adequate training to ensure

[1] The WET Coalition consists of the following members: Alliance of Automobile Manufacturers, American Chemistry Council, American Forest & Paper Association, American Petroleum Institute, AMSA, Rubber Manufacturers Association, Utility Water Act Group, VAMWA, WESTCAS, Alcoa, General Electric, Kennecott Utah, and Milliken Company.

that regulators understand WET test procedures, and the significance and limitations of WET test results that may be used in their water quality standards and NPDES programs.

The Coalition is concerned that EPA is forging ahead in the use of WET test methods without having satisfied adequately the prerequisites identified above. In particular, EPA proposes to ratify several test methods in 40 C.F.R. Part 136, and thereby to certify their acceptability for use in a context that carries civil and criminal sanctions, without the necessary scientific and legal basis for its decisions. The WET Coalition is willing to commit its resources to work cooperatively with the Agency in finding practical means of expanding the usefulness and applicability of WET test methods in the regulatory process.

The WET Coalition's interest in a cooperative effort is sincere, as evidenced by the extensive comments and recommendations already submitted to EPA in response to the Draft Interlaboratory Study Design[2] and the Preliminary Interlaboratory Study Report,[3] as well as the letters of July 16, 2001,[4] and September 18, 2001,[5] offering recommendations regarding the "data quality" issues that arose during the Interlaboratory Study.

EPA's review of the WET test methods – for purposes of ratification or withdrawal –

---

[2] Koorse, Steven J. (on behalf of UWAG), Comments on EPA's Proposed Charge to Reviewers: Interlaboratory Study of WET Test Methods (September 15, 1998); Risk Sciences (on behalf of WESTCAS), Comments on EPA's Proposed Charge to Reviewers: Interlaboratory Study of WET Methods (September 14, 1998).

[3] Koorse, Steven J. (on behalf of UWAG and WESTCAS), Comments on EPA's Preliminary Report: Interlaboratory WET Variability Study (December 11, 2000).

[4] Koorse, Steven J. (on behalf of WET Coalition), Letter to Geoffrey H. Grubbs, EPA Office of Water, re: Whole Effluent Toxicity Program (July 16, 2001).

[5] Koorse, Steven J. (on behalf of WET Coalition), Letter to Geoffrey H. Grubbs, EPA Office of Water, re: WET Test Rulemaking (September 18, 2001).

should be no less broad nor intense than that appropriate for original approval of test methods in 40 C.F.R. Part 136. The test methods promulgated in the 1995 Part 136 rule were contested and are not entitled to any presumption of acceptability in the current rulemaking. EPA should judge the WET methods on their own merits based on all of the information now available regarding WET test method performance in the routine regulatory context authorized for Part 136 methods.

The comments below are presented in four parts. The first part discusses why EPA must validate and publish performance characteristics for all analytical methods included in Part 136. The second part discusses concerns over EPA's basis for approving the proposed WET test methods. The third part addresses the Coalition's concerns over the absence of adequate QA/QC protocols in the proposed methods. The fourth part addresses additional concerns, including responses to some of the issues on which EPA specifically sought comment.

I. EPA MUST VALIDATE AND PUBLISH PERFORMANCE CHARACTERISTICS FOR ALL METHODS PROMULGATED IN PART 136.

EPA relies on WET test methods as a cornerstone in the Clean Water Act's integrated program of pollution control. The discussion that follows describes why EPA bears the responsibility to: (1) ensure that all of the test methods to be approved for use in that program are adequately reliable in the context of their intended use, and (2) publish the information permitting authorities need in order to account for analytical variability and other WET test performance limitations in the permitting and/or enforcement process.

A. Due Process Requires Methods Be Validated.

Onerous criminal and civil sanctions[6] may be imposed against persons exceeding NPDES permit limitations established under the Clean Water Act or analogous state law. It is therefore essential that only those analytical methods capable of accurate and reproducible performance be used for measuring compliance with such limitations. EPA's judgment as to which methods qualify for compliance monitoring must be based on an objective and scientifically sound method validation process that includes an assessment of the variability of the test method. If an inappropriate validation process is used, "approved" test methods could lead to inaccurate test results, and thus to constitutional due process violations. Indeed, permittees are entitled to an objective basis for demonstrating compliance, one that will result in enforcement sanctions only where their conduct is actually unlawful.

For example, the Court of Appeals for the District of Columbia, in responding to arguments about the adequacy of test methods available for measuring a lead limit, stated:

[6] *See* CWA § 309, 33 U.S.C. § 1319.

> The possibility of statistical measurement error, which is often unavoidable where regulations set quantitative standards, does not detract from an agency's power to set such standards, it merely deprives the agency of the power to find a violation of the standards, in enforcement proceedings, where the measured departure from them is within the boundaries of probable measurement error.[7]

This case confirms that, under the Clean Water Act, EPA must account for analytical variability either at the time of standard setting and permitting or at the time of enforcement. Indeed, EPA must assure that the irreducible performance limitations inherent in all test methods (including, but not limited to, imprecision) will not act to penalize persons for lawful conduct. EPA cannot provide such assurances absent the collection of adequate performance data and a scientifically sound analysis thereof. This can be accomplished only through the proper conduct of method validation studies.

B. Test Methods Must Be Fully Validated And Their Performance Characteristics Considered In The Regulatory Process.

1. Method Performance Must be Evaluated Fully Before Publication in Part 136.

a) Proper Method Validation Studies Must be Performed

EPA issued a Report to Congress that provides the Agency's recommendations on the use of analytical methods in the regulatory context.[8] In that report, EPA states:

> methods which will be used extensively for regulatory purposes or where significant decisions must be based on the quality of the analytical data normally require more extensive validation and standardization than methods developed to collect preliminary

[7] *Amoco Oil Co. v. EPA*, 501 F.2d 722, 743 (D.C. Cir. 1974) (emphasis in original).

[8] U.S. Environmental Protection Agency, Availability, Adequacy, and Comparability of Testing Procedures for the Analysis of Pollutants Established Under Section 304(h) of the Federal Water Pollution Control Act, Report to Congress, EPA/600/9-87/030 (September 1988) ("Section 518 Report").

baseline data.[9]

In addition, EPA states, "[w]here possible, and in all cases for methods that will have extensive regulatory use, a method should be fully validated and standardized."[10] Surely, the WET test methods at issue in this Part 136 rulemaking — methods for nationwide application — require full validation and standardization.

EPA defines "validate" as follows.

> To verify, using an acceptable scientific process, that a method is based on sound technical principles and has been reduced to practice for routine measurement purposes.[11]

EPA has delineated the "sound technical principles" it considers necessary to "validate" a method.[12] EPA's Section 518 Report describes a three tier validation process, which consists of testing, evaluating, and characterizing the method to the extent necessary to demonstrate that the method achieves a specified performance. In Section II of these comments, the Coalition highlights the particular characteristics of the WET test methods on which the validation process must focus.

More recently, EPA has established an Agency-Wide "Data Quality System" and designated the Office of Environmental Information Quality Staff to administer the program. One of the primary goals of the Agency-Wide Data Quality System is "to ensure that

[9] *Id.* at 3-5.

[10] *Id.* at 3-6 (emphasis added).

[11] *Id.* at viii.

[12] *Id.* at Chapter 3; EPA's draft "Guidelines for Selection and Validation of USEPA's Measurement Methods" also contains guidance for selection and validation of analytical methods. *See* U.S. Environmental Protection Agency, *Guidelines for Selection and Validation of USEPA's Measurement Methods* (August 1987) (Draft) ("EPA Guidelines").

environmental programs and decisions are supported by data of the type and quality needed and expected for their intended use . . . ."[13] A prerequisite to meeting that goal is the availability of reliable test methods.

One of the guidance documents EPA is preparing to finalize discusses the various "Data Quality Indicators" ("DQIs") the Agency believes regulators must consider in evaluating test data to be used for regulatory decisions (e.g., coefficient of variation, standard deviation, relative bias, percent recovery, etc.).[14] DQIs also must be developed in method validation studies so that: (1) EPA can determine whether the test method is sufficiently reliable for regulatory use, (2) regulators have a basis for deciding which WET test method is suitable for the particular regulatory use at issue, and (3) regulators know up front, and have the opportunity to account for, the analytical error and other performance deficiencies the test method is expected to exhibit. Unless the validation studies are conducted properly and their results interpreted properly, the DQIs will mislead, rather than assist, regulators in performing the above tasks.

The practical consequence of failing to validate a method adequately before introducing it for regulatory use is that little confidence can be placed on the data produced by that method. EPA's report on two mercury methods that it previously had deemed acceptable for unconditional application in the regulatory process shows why "validation" is an essential prerequisite to publication of a method in Part 136, and why EPA needs specific criteria to

[13] U.S. Environmental Protection Agency, *EPA Quality Manual for Environmental Programs*, EPA 5360 A1 (May 5, 2000).

[14] U.S. Environmental Protection Agency, *Guidance on Data Quality Indicators* (EPA QA/G-5i) (September 2001) (Peer Review Draft) ("DQI Guidance").

determine acceptable performance.[15] EPA originally approved those two methods on the basis of performance data from a single laboratory. As a result of the markedly poorer performance exhibited by the methods when subsequently evaluated in an interlaboratory study, EPA deemed it necessary to place severe restrictions on the future use of both methods.

EPA previously has expressed serious reservations over the use of inadequately validated test methods and has provided guidance on what it must do whenever such methods are used:

> The user (i.e., the program office) must be cautioned that if one or more of the validation steps is omitted -- because of time or resource limitations -- the accuracy (precision and bias) of the measurement data collected by method users will not have been established as completely as desired, and the data may, therefore, have limited usefulness. In such cases, the agency program should (1) thoroughly evaluate the circumstances and available alternatives, (2) diligently seek compensatory measures such as field evaluation and continual assessment of method performance, (3) understand the potential limitations on the usefulness of the resulting measurement data, (4) insure that the limitations are clearly specified and remain with the data set so that programs do not use the data improperly, and (5) adequately justify and document its decision and rationale for the use of the method under these circumstances for the intended application.[16]

Fundamentally, the Coalition believes that no inadequately validated method should be published in Part 136 due to the potential for improper application notwithstanding the inclusion of the warnings EPA believes to be necessary.

---

[15] Gebhart, J.E., J.D. Messman, and G.F. Wallace, *Interlaboratory Evaluation of SW-846 Methods 7470 and 7471 for the Determination of Mercury in Environmental Samples*, U.S. EPA Environmental Monitoring Systems Laboratory, EPA/600/4-88/011 (April 1988).

[16] EPA Guidelines at 5 (emphasis added).

b) EPA Must Specify the Criteria on Which It Relied in Ratifying a Test Method

In approving test methods for nationwide regulatory use under Part 136, EPA must use an objective set of criteria by which to evaluate reliability of the method (e.g., how much imprecision is too much?). EPA has not identified these criteria. Instead, it approved the WET test methods based solely on a conclusion that their precision is comparable to the precision of approved chemical test methods. However, the document EPA references[17] to support its argument on the precision for chemical methods merely presents the "error-band" figures for those methods. That document does not explain EPA's basis for concluding that those precision figures were acceptable in light of the intended use of the chemical test methods. Nor does the document specify the upper level of imprecision that EPA deems unacceptable, which leaves open the question – what level of imprecision does EPA consider unacceptable for purposes of approving test methods?

In its final WET rule, EPA decided that the "Ames Test" was not sufficiently reliable for inclusion in Part 136.[18] It did not, however, explain its rationale for making that decision. It merely stated that "this test produces many false results, and thus, could potentially confuse or mislead regulators."[19] Yet, that same statement holds true to varying degrees for all of the WET test methods. This failure to identify the benchmark against which the acceptability of test results will be measured is a fatal flaw in EPA's proposal.[20] The public is deprived of the

[17] U.S. Environmental Protection Agency, *Technical Support Document for Water Quality-Based Toxics Control*, EPA 505/2-90/001 (March 1991).

[18] 60 Fed. Reg. 53,529, 53,531 (October 16, 1995).

[19] *Id*.

[20] EPA has not even offered a rationale for its departure from this requirement (e.g., by

opportunity to comment on the standard EPA is using to decide which test methods are reliable and which are not.

Even assuming, for argument sake, that the precision for chemical methods was acceptable, EPA never explains why that level is acceptable for WET test methods in light of the disparate manner in which WET and chemical specific test methods are used in the regulatory process (e.g., EPA has yet to establish a detection/quantification level concept for WET test results akin to that routinely used for chemical test results). Further comments are presented in Section II below regarding the inappropriate comparison of WET and chemical-specific precision estimates.

c) The *Selenastrum* Test, Which is Unreliable, Exemplifies EPA's Arbitrary Part 136 Approval Process

By significant measures of performance, such as precision and completion rate, the *Selenastrum* test is unsuitable for inclusion in Part 136. For example, the CV for the IC_{25} endpoint with EDTA was 34.3%. EPA does not attempt to justify the high CV or discuss the consequences of such imprecision if the method is used in the regulatory process. Nor does EPA attempt to justify the No Observed Effect Concentration ("NOEC") results, which also were extremely variable. Approximately 15 percent of the NOEC results for both the reference toxicant and the receiving water results spanned two or more concentrations above or below the median.[21]

---

explicitly confirming that regulators will be able to adjust NPDES permit limitations to ensure that dischargers are not unjustly penalized for excursions attributable to analytical error).

[21] U.S. Environmental Protection Agency, *Proposed Changes to Whole Effluent Toxicity Method Manuals*, EPA 821-B-01-002 (September 2001) ("Proposed Method Manuals Changes"), p. 31.

Finally, EPA admits that the completion rate for the test using EDTA was only 63.6%. It attempts to justify that excessive failure rate by claiming that "the use of EDTA will improve successful test completion rates for the methods as laboratories consistently culture and test with EDTA."[22] It is unclear how the use of EDTA will improve completion rates when the completion rate was about the same (in fact it was slightly better) when EDTA was not used.

In addition to explaining its rationale for approving *Selenastrum* with EDTA under Part 136, notwithstanding the poor performance discussed above, the Agency must explain its rationale for authorizing the test without EDTA. Test performance, in the absence of EDTA, was poor in every category. Of all the tests initiated, 34% could not be completed. Of all the tests performed on non-toxic water, 33% produced "false" toxicity results. The CV for the tests that were completed was 58.5 %. EPA admits that such variability exceeds levels exhibited by the other chronic test methods.[23]

For the NOEC endpoint, the NOEC values spanned between 2 and 6 concentrations of the median in 60% of the tests for reference toxicants, and between 2 and 4 concentrations in 50% of the tests for effluents.[24] That means, if the actual NOEC were 25%, the results might be anywhere from NOEC = 0% to NOEC = 100%. How could such extreme variability ever be deemed acceptable?

Notwithstanding the great uncertainty associated with the *Selenastrum* test without EDTA, EPA merely "recommends" that the test be run with EDTA. It does not require the use

[22] 66 Fed. Reg. at 49,808.

[23] 66 Fed. Reg. at 49,807.

[24] Proposed Method Manuals Changes at 31.

of EDTA. Moreover, it specifies that testing without EDTA "may be appropriate" where metals are known to be contributing to sample toxicity. It points out that EDTA may bind with some metals, thereby causing tests to underestimate toxicity. The binding effect may be present, but it does not justify EPA's approval of an unreliable test method. EPA has acknowledged that "food" also may "sequester" metals and "confound test results" in Fathead Minnow tests.[25] Yet it does not conclude that it is acceptable to starve the test organisms even though they will perform erratically (if at all) as a consequence. Instead, it provides procedures that "will reduce the probability of reduction of toxicity caused by feeding."

EDTA, to a plant like *Selenastrum*, is essentially no different than food to a minnow. If EPA concludes that EDTA is necessary to achieve acceptable performance with *Selenastrum*, then EDTA should be required in all applications. If EPA has concerns over the binding effects of EDTA on metals, it should state in the test manual that *Selenastrum* is not an appropriate test organism to use if metals may be a source of toxicity.

The proposal to allow testing without EDTA is not rendered acceptable by EPA's statement that testing without EDTA:

> may be conducted if the testing laboratory has demonstrated success in the use of the without EDTA procedure. Demonstrated success should include documentation of meeting appropriate test acceptability criteria and control charts of reference toxicant tests conducted without the addition of EDTA.

EPA does not specify how this provision will be implemented in practice. Do laboratories have to obtain advance approval from their regulatory authority, or does EPA expect

[25] *See, e.g.,* U.S. Environmental Protection Agency, *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Freshwater Organisms*, 3rd Ed., EPA-600-4-91-002 (July 1994) ("Chronic Freshwater Manual"), p. 59.

regulatory authorities to evaluate a laboratory's "demonstrated success" upon receipt of test results? If the former, EPA is establishing a laboratory certification process without any objective standards by which to govern the approval process. If the latter, permittees will be left with no objective basis for knowing which laboratory will be deemed acceptable for purposes of performing its compliance monitoring. In either event, the test method without EDTA is not acceptable for approval under Part 136.

In short, the *Selenastrum* test, with or without EDTA, exhibits extremely poor performance. EPA nevertheless proposed to approve the test under Part 136. Absent an objective standard for making such decisions, EPA's action is arbitrary.

2. All Relevant Performance Characteristics Must Be Published In Part 136.

Test methods cannot be used in the regulatory process absent published information on how they can reasonably be expected to perform.[26] That is because EPA recognizes that all test methods exhibit variability in the way they perform and that method performance must be taken into account in the regulatory process. EPA has not published all appropriate performance information along with its test methods. Indeed, in its Section 518 Report to Congress, EPA stated that "the natural variability in sensitivity [or response of test organisms] . . . must also be accounted for when permit limits, criteria, or standards are set."[27] In that report, EPA also

[26] For example, EPA includes method detection limits and statements of both single operator and interlaboratory precision and bias along with the analytical methods it published at 40 C.F.R. Part 136 for measuring organic pollutants in wastewater. 40 C.F.R. Part 136, App. A (1989) (the "600 series"). The Part 136 regulations set forth the methods used under the CWA. EPA has also published performance characteristics along with the test methods prescribed for use in the RCRA and Superfund programs. *See* U.S. Environmental Protection Agency, *Test Methods for Evaluating Solid Waste, Physical/Chemical Methods, SW-846*, 3rd Ed. (currently being revised, 54 Fed. Reg. 3,212 (January 23, 1989)).

[27] Section 518 Report at 3-11.

recommends that "[p]performance data should be contained in the method for each analyte listed."[28] EPA recently reiterated the importance of understanding the variability:

> the quality attributes of greatest utility from a [data quality objective] planning and statistical design perspective include estimates of the overall (total) study variability and an understanding of the relative contribution of significant components of this total.[29]

In its guidance on DQIs, EPA states that the "DQIs for precision are among the most important quality indicators in an environmental study."[30] Once developed, all DQIs for relevant performance characteristics must be published along with any method intended to be used in the regulatory process. EPA has not published all appropriate performance information along with its test methods. The Coalition's specific concerns regarding performance characteristics are addressed in Section II.

3. EPA Must Perform Interlaboratory Studies for all WET Methods Proposed for Part 136.

Interlaboratory testing is an absolute prerequisite for approval of test methods to be published in 40 C.F.R. Part 136.[31] The measurement data on which compliance determinations and other regulatory decisions are made can be produced in any one of several different laboratories (i.e., industry, government, or commercial laboratories).[32] Thus, it is not enough to determine the reliability of a test by evaluating how it performs in a single laboratory (i.e., an

[28] *Id.* at 3-5.

[29] DQI Guidance at 16.

[30] *Id.* at 17.

[31] EPA already has confirmed that interlaboratory method performance is essential for test methods to be used in the regulatory process. 52 Fed. Reg. 25,699 (July 8, 1987).

[32] 60 Fed. Reg. at 53,532 (col.3).

intralaboratory study). EPA must prove that its tests will perform reliably regardless of which qualified laboratory is involved. Indeed, in its Report to Congress, EPA stated that methods "developed for monitoring purposes must be based on sound scientific principles and be practical for routine use."[33] The only way to determine whether a method is practical for routine use is by evaluating performance data from several laboratories (i.e., *inter*laboratory data). *Intra*laboratory data will tend to underestimate the amount of variability (i.e., error) that a test method will produce in practice. EPA acknowledged this most recently in the WET proposal.[34]

The following EPA statement underscores why WET test methods must be validated for regulatory purposes based on interlaboratory data:

> [p]ossible causes [of between-laboratory variability] may include laboratory differences in concentration series, incorrect or ambiguous calculation or reporting of concentrations . . . laboratory differences in dilution water (e.g., water hardness or pH), laboratory differences in foods and feeding regimes, and laboratory differences in cultures (genotypic and phenotypic differences in sensitivity to various toxicants).[35]

EPA itself has explicitly stated that interlaboratory studies are necessary before a method is acceptable for use in the NPDES program. In response to comments recommending the use of positively charged filters for virus concentration, EPA stated:

> Interlaboratory studies carried out jointly by EPA and ASTM have thus far validated only the use of negatively charged filters for virus concentration. The use of positively charged filters in the

[33] Section 518 Report at 3-1 (emphasis added).

[34] 66 Fed. Reg. at 49,806 (col. 2).

[35] U.S. Environmental Protection Agency, *Understanding and Accounting for Method Variability in WET Applications Under the NPDES Program*, EPA 833-R-00-003 (June 2000) ("WET Variability Guidance"), p. 3-11.

> NPDES Program will not be acceptable until it is established, through collaborative (multi-laboratory) studies, that their performance is equivalent to negatively charged filters, and the method is approved as an alternate method under the procedures established in 40 C.F.R. Part 136.[36]

According to the ASTM standard on determining precision and bias, performance statistics "must be based on data from at least six laboratories that passed all of the outlier tests, [] that is, retained data."[37] Recognizing this, EPA set a data quality objective for the variability study of a minimum of six laboratories.[38] But as discussed below, not all test methods proposed by EPA underwent or successfully completed the interlaboratory testing according to predefined objectives.

4. EPA Did Not Perform an Interlaboratory Study for All WET Methods.

For a few test methods, EPA did not perform the interlaboratory testing required for methods to be used in the regulatory process.

a) *Champia parvula* Reproduction Test

EPA did not perform an interlaboratory study for the *Champia parvula* reproduction test due to insufficient participant laboratory support. EPA justified its inclusion of the method without interlaboratory validation on two grounds: (1) the need for the method because it represents the only approved test method for a marine plant species, and (2) the referee

[36] U.S. Environmental Protection Agency, *Supplementary Information Document, Whole Effluent Toxicity: Guidelines Establishing Test Procedures for the Analysis of Pollutants* (October 2, 1995) (WET SID"), p. 56.

[37] *See* ASTM D2777, *Standard Practice for Determination of Precision and Bias of Applicable Test Methods of Committee D-19 on Water*, § 7.2.3 (1998). Under the National Technology Transfer Act, EPA is required to use the ASTM standard unless it can establish that the standard is impracticable.

[38] 66 Fed. Reg. at 49,804, 49,806.

laboratory data confirmed the estimates of the intralaboratory precision cited at the time of method promulgation.[39]

The justifications provided by EPA for the proposed inclusion of the *Champia parvula* reproduction test in Part 136 are unacceptable. The need for a test method does not and cannot justify circumventing the interlaboratory validation of the method. The necessity does not make the validation any less important. The fact that this is the only method makes it even more important that it be validated so that permittees required to perform the test have confidence in the results.[40]

Nor are the referee data sufficient to validate this method for nationwide use under 40 C.F.R. Part 136. As discussed above, intralaboratory data cannot substitute for interlaboratory data. Moreover, it is arbitrary to use referee data here but reject it elsewhere. For other test methods, "EPA excluded referee laboratory test data from analysis of successful test completion rate, false positive rates, and precision because referee laboratory testing was not conducted on blind test samples."[41] Additionally, EPA stated that it had the opportunity to collect data from one other qualified laboratory, but chose instead not to perform the interlaboratory study at all.[42] That decision also was arbitrary.

---

[39] *Id.* at 49,806.

[40] *See* footnote 16, *supra*.

[41] U.S. Environmental Protection Agency, *Response to Comments: Peer Review of "Preliminary Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods"* (September 2001) ("Response to Peer Review Comments"), p. 13.

[42] 66 Fed. Reg. at 49,806.

Moreover, EPA's willingness to accept the variability for *Champia* in the proposed WET rule is inconsistent with its decision in the WET Variability Guidance. In that guidance document, EPA described the variability for all the WET test methods except *Champia*.[43] Its rationale for excluding *Champia* was "it would be inadvisable to characterize method variability using only 23 tests from only two laboratories."[44] EPA cannot endorse a test method for use in the regulatory process until it can characterize its variability and confirm that the variability is acceptable.

Finally, even if it were acceptable to rely exclusively on the referee laboratory, the test results confirm that the test method is not sufficiently reliable for regulatory use. For example, the referee laboratory conducted tests on an unspiked split sample of receiving water collected on May 23, 2000. The IC_{25} results, which were expected to have been $\geq$ 100%, were 7.53% and 90.4%. In response to a peer review comment that the "level of variability is incredible to say the least,"[45] EPA decided to declare the 7.53% result "inconclusive."[46] Its rationale for that decision was predicated on its conclusion that the percent minimum significant difference ("PMSD") for the test was "above recommended bounds."[47] EPA calculated the PMSD to be 47%, which it claimed to be higher than the upper bound PMSD for other chronic methods (since

[43] EPA's guidance did not describe the variability for all of the WET test method "endpoints."

[44] WET Variability Guidance at 3-8.

[45] Response to Peer Review Comments at 40 (Comment 41.X.2).

[46] U.S. Environmental Protection Agency, *Final Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods, Vol. 2: Appendix,* EPA 821-B-01-005 (September 2001), p. D-25.

[47] *Id*.

the "upper PMSD bounds have not yet been recommended for the *Champia* chronic method").[48]

That decision is inconsistent with EPA's statement in Appendix E that "determinations of test validity were not made based on PMSD bounds."[49] In Appendix E, EPA made reference to its guidance document that recommends invalidating data where the PMSD exceeds the upper bound only "if the test leads to a decision that there is no significant toxicity at the concentration identified in the permit as a limit ("Instream Waste Concentration" ("IWC") or "Receiving Water Concentration")."[50] EPA then states it could not use upper PMSDs bounds for invalidating data in the interlaboratory study, "because IWC concentrations were not established or applicable to an interlaboratory study."[51] Yet, EPA seems to have deviated from that decision in its review of *Champia* data. Thus, the "incredible" degree of variability exhibited by *Champia* remains unexplained, and it confirms that the reliability required for Part 136 test methods is lacking.

b) *Holmesimysis costata* Acute Test

EPA also did not perform an interlaboratory study for the *Holmesimysis costata* acute test. The *Holmesimysis* test proposed by EPA is a **new** method that was not considered in the interlaboratory study. An interlaboratory study was proposed for the originally promulgated *Holmesimysis* test (using *Holmesimysis costata* as an acceptable test species with the *Mysidopsis bahia* acute test procedures) but was not conducted due to insufficient participant laboratory

[48] *Id*.

[49] *Id*. at E-3 (emphasis added).

[50] *Id*.

[51] *Id*.

support.[52] After the referee laboratory attempted to conduct the promulgated test and failed, EPA decided to withdraw it and propose the new method.[53]

EPA justified its inclusion of the new method without interlaboratory validation on two grounds: (1) the method is required only in permits issued in California, and (2) the method development data from California and peer review literature show, given the appropriate test procedures and test conditions, that the test method "can produce reliable and sensitive toxicity results with adequate precision."[54]

Again, the justifications provided by EPA for the proposed inclusion of the *Holmesimysis costata* acute test in Part 136 are unacceptable. The fact that the method is used only in California does not negate the need for validation. Relying on studies other than interlaboratory studies to satisfy the validation requirement is unacceptable as explained above. If EPA wishes to propose a new test method, it first must perform the interlaboratory validation studies necessary to evaluate whether or not the test will be reliable. The method development data and peer review literature cited by EPA are inadequate to demonstrate the reliability of EPA's proposed method for compliance purposes.

EPA cites two peer review journal articles to support the proposed method,[55] one that

---

[52] Two laboratories qualified to perform the promulgated *Holmesimysis* test, but again EPA chose not to conduct any interlaboratory testing. 66 Fed. Reg. at 49,808.

[53] *Id*.

[54] *Id*. at 49,809.

[55] *Id*.

reports acute method development[56] and one that reports 7-day chronic method development.[57] Neither demonstrates that 90% control survival, which is EPA's test acceptability criterion for all acute methods including the proposed *Holmesimysis* method, can be reliably achieved. A number of acute experiments had more than 10% control mortality. The laboratory assumed the mortality was due to salinity greater than 36 parts per thousand, but this theory was not tested. The 7-day chronic experiments include acute 96-hour LC_{50} results, but not 96-hour control survival. It cannot be determined from the paper whether these tests would have met the acute acceptability criterion of 90% control survival – many of the tests show less than 90% control survival at seven days. Interlaboratory validation would be needed to demonstrate that the proposed acute method would result in acceptable control survival.

The acute method development included daily feeding during the 96-hour test, but EPA's proposed *Holmesimysis* method does not. This change would need to be validated as well.

The acute method paper is the source of the multilaboratory precision measurements cited by EPA.[58] These were very limited interlaboratory tests: there were a total of 4 tests and 3 laboratories: the method development laboratory did one paired test with each of the two other laboratories. Since the endpoint determination method was changed after the first paired test due to a significant difference in results between laboratories ($p=0.04$), the interlaboratory comparison of the final method consisted of one paired test. This does not meet EPA's WET

[56] Martin, M., et al. 1989. Experimental evaluation of the mysid *Holmesimysis costata* as a test organism for effluent toxicity testing. *Environ. Toxicol. Chem.* 8:1003-1010.

[57] Hunt, J.W., et al. 1997. Precision and sensitivity of a seven-day growth survival toxicity test using the west coast marine crustacean *Holmesimysis costata*. *Environ. Toxicol. Chem.* 16:824-834.

[58] 66 Fed. Reg. at 49,809.

Variability Study standards, so the EPA-cited references should not be used as a substitute to validate the *Holmesimysis* method.

Reliability of the method for compliance purposes also depends upon organism availability. EPA proposes to add *Holmesimysis costata* to the list of recommended acute test species, which are described in the methods manual as easily cultured in the laboratory and generally available throughout the year.[59] EPA's own experience with the WET Variability Study shows this does not describe *Holmesimysis*: test organisms "are generally obtained from field collected gravid females,"[60] and the reference laboratory "was unable to collect sufficient numbers of gravid females during most of the time frame of the WET Variability Study (September 1999 through April 2000)."[61] An organism that cannot be reliably obtained year-round is not appropriate for compliance testing.

[59] Proposed Method Manuals Changes at 83.

[60] 66 Fed. Reg. at 49,808.

[61] *Id*.

c) *Mysidopsis bahia* Fecundity Test

The *Mysidopsis bahia* fecundity endpoint was completed successfully only 50% of the time in the interlaboratory study. Due to the low successful test completion rate, EPA proposes to modify the protocol to improve performance.[62] EPA, however, has not demonstrated via an interlaboratory study (or otherwise) that the proposed changes will eliminate the unacceptable performance exhibited by the current version of the test method. Until the interlaboratory validation has been conducted on the test method as proposed, neither the method modification nor the current version of the method can be approved under Part 136. More specific comments supporting withdrawal of the fecundity test endpoint from Part 136 are presented below in Section VI.

5. EPA Did Not Perform Adequate Interlaboratory Testing for Certain Test Methods.[63]

For *Selenastrum* chronic and Silverside acute test methods, EPA did not perform adequate interlaboratory testing. While EPA contracted with at least six laboratories to perform these tests, less than six successfully completed the tests for all endpoints. For *Selenastrum*, less than six laboratories completed testing on: the blank sample with EDTA for the NOEC, IC_{25} and IC_{50} endpoints; the blank sample without EDTA for the NOEC endpoint; and the receiving water sample without EDTA for the NOEC and IC_{25} endpoints. For the Silverside acute test method, only five laboratories successfully completed testing for the receiving water LC_{50} endpoint. Given the unacceptable number of participating laboratories, EPA lacks the data necessary to

[62] *Id*.

[63] In light of the data quality problems EPA experienced with the majority of the test results received from participating laboratories in the interlaboratory study (see Section III below), EPA did not perform interlaboratory validation with a sufficient number of laboratories for several test methods in addition to those discussed in this subsection.

determine whether or not those methods are sufficiently reliable to be approved under Part 136.

II. EPA DID NOT VALIDATE THE ESSENTIAL PERFORMANCE CHARACTERISTICS OF THE PROMULGATED TOXICITY TEST METHODS

In order to promulgate a test method under 40 C.F.R. Part 136, EPA must validate the following performance characteristics: accuracy, precision, dynamic range, detection limits, interferences, ruggedness (applicability), reporting, and representativeness/method comparability.[64] Validation is necessary both to determine whether or not a test method is adequately reliable for its intended uses and to provide the performance information needed to properly use the test method and the results it generates. EPA failed to validate the essential performance characteristics for the WET test methods. Further, the available evidence shows that the WET test methods are incapable of meeting minimum acceptable standards for these performance characteristics.

A. EPA Did Not Validate The Accuracy Of The WET Test Methods.

1. EPA Did Not Establish a Data Quality Objective for Minimum Acceptable Accuracy.

EPA Order 5360.1-A2 requires the Agency to establish data quality objectives ("DQOs"). DQOs establish criteria for determining whether the data collected are suitable for their intended purpose. At a minimum, DQOs must identify performance specifications to evaluate data quality.[65] Foremost among these specifications is "accuracy."

Accuracy is a gauge of how close the measured test result is to the true value. Approved test methods for individual pollutants (like copper) have been validated for accuracy, relative to a

[64] Section 518 Report, p. 3-2 to 3-5 and p. 4-49.

[65] U.S. Environmental Protection Agency, *Policy and Program Requirements for the Mandatory Agency-Wide Quality* System, EPA Order 5360.1 A2 (May 5, 2000).

traceable standard, so the user knows approximately how far from the true value (i.e., error) any measurement result is likely to be. For example, if the validation study shows accuracy to be 120% when measuring a reference standard known to contain 10 ug/l of copper, the user will know that a test result of an effluent showing 12 ug/l is just as likely 10 ug/l. However, for a "method-defined parameter"[66] such as WET, there is no means of corroborating the test results. Moreover, EPA's policy of Independent Applicability ostensibly prevents permittees from using bioassessment data documenting the health of the stream to rebut an inference of toxicity from a WET test.

In essence, failing a WET test, according to EPA, creates an irrebutable presumption that the effluent contains an excessive level of "toxicity." Yet, absent a traceable reference standard for "toxicity," there is no means to know how much toxicity actually is in the effluent. Accuracy is essential to avoid decision errors in enforcement actions.[67] EPA recently published guidance documents requiring Agency personnel to define the probability of and tolerance for decision errors.[68]

EPA failed to establish a formal DQO for minimum acceptable accuracy or tolerance thresholds for decision errors related to WET test methods. As such, it is impossible to

[66] EPA sometimes elects to regulate the effects of pollution when the specific chemical cause may be unknown. In such cases, the measured effect becomes the regulated parameter and the method used to measure that effect serves as the operational "definition" of the parameter itself. If the method changes, the pollutant "level" may also change even though the actual concentration of unidentified chemicals causing the effect remains unchanged. Toxicity and Biological Oxygen Demand ("BOD") are examples of method-defined parameters.

[67] U.S. Environmental Protection Agency, *NPDES Permit Writers' Manual*, EPA-833-B-96-003 (December 1996) ("Permit Writers' Manual").

[68] U.S. Environmental Protection Agency, *Guidance for the Data Quality Objectives Process* (EPA QA/G-4), EPA/600/R-96/055 (August 2000) ("DQO Guidance").

demonstrate that the results of any given WET test are appropriate for their intended use under Part 136.

2. EPA Disregarded Accuracy in Developing its WET Test Methods.

The Agency claims that it is impossible to measure the accuracy of biological test methods.[69] Although we disagree with this assertion,[70] even if EPA is correct, the inability to demonstrate the accuracy of a promulgated test method does not nullify the requirement to do so. Accuracy is not optional in the context of test methods utilized for compliance determinations.[71] EPA's own regulations related to mandatory data quality management demand accuracy.[72]

Accuracy is the single most important performance characteristic. If accuracy were not important, there would be no need to perform any analyses; a guess would be as good as a measurement. Indeed, in one of its most recent Part 136 test methods (Mercury Method 1631), EPA states, "the laboratory shall make an initial demonstration of the ability to generate acceptable accuracy and precision within this Method."[73] It states that those "acceptable criteria"

---

[69] Chronic Freshwater Manual, pp. 139, 193 and 225.

[70] For example, it is possible to evaluate accuracy as to samples known to be free from toxicity.

[71] *See* Office of Management and Budget, *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies,* 66 Fed. Reg. 49,718 (Sept. 28, 2001).

[72] *See* U.S. Environmental Protection Agency, *EPA Quality Manual for Environmental Programs*, EPA 5360 A1 (May 5, 2000); U.S. Environmental Protection Agency, *EPA Requirements for Quality Management Plans* (EPA QA/R-2), EPA/240/B-01/002 (March 2001); U.S. Environmental Protection Agency, *EPA Requirements for Quality Assurance Project Plans* (EPA QA/R-5), EPA/240/B-01/003 (March 2001).

[73] U.S. Environmental Protection Agency, *Method 1631, Revision B: Mercury in Water by Oxidation, Purge and Trap, and Cold Vapor Atomic Fluorescence Spectrometry*, EPA 821-R-98-002 (May 1999), Protocol 9.1.1. (64 Fed. Reg. 30,417 (June 8, 1999)).

"recognize the variability expected to occur among laboratories."[74] EPA inexplicably does not require such up-front accuracy confirmation for WET tests, even though it states:

> The lack of a standard or common reference toxicant creates a problem for permittees and regulatory authorities attempting to evaluate or compare laboratories. Real or apparent differences occur between laboratories in mean values of EC25, LC50, and NOEC. Some of this difference is random and reflects only the with-in laboratory variance; some may be systematic. Systematic, between-laboratory differences can be inferred only when laboratories use the same test method, use the same reference toxicants and dilution series, use similar dilution waters, and report a sufficient number of tests.[75]

In the real world, laboratories will not be using the same reference toxicants, dilution series, and dilution waters. Moreover, even a single test result could be used to make regulatory decisions. Nonetheless, EPA acknowledges that it cannot determine the differences that might arise depending on the laboratory selected.

The fact that EPA requires dischargers to certify that all reported results are "true, accurate, and complete" affirms the central importance of accuracy in the monitoring system used to evaluate compliance with the Clean Water Act.[76] The Clean Water Act authorizes harsh penalties for those who misrepresent the true quality of effluent discharges.[77]

If EPA is unable to confirm the accuracy of WET test methods, then those methods are no longer suitable for all of their intended purposes. In particular, a discharger will not be able to report its WET test results in its discharge monitoring report ("DMR") and sign a certification

[74] 64 Fed. Reg. 30,417, 30,423 (col. 3) (June 8, 1999).

[75] WET Variability Guidance at 3-11.

[76] 40 C.F.R. § 122.22(d).

[77] CWA § 309(c) and (d) (providing for criminal and civil penalties for false statements).

that its results are "true, accurate, and complete." 40 C.F.R. § 122.22(d). The Agency has issued guidance[78] explaining that, for WET methods, the term "accuracy" merely requires the discharger to write or type on the DMR the exact numerical test result received from the laboratory. The guidance is intended to relieve the discharger of responsibility for certifying that the WET test result shows the actual toxicity level in its effluent. However, if a discharger that has "failed" its WET tests cannot certify that its actual toxicity level exceeds the WET permit limit, it will be placed in the untenable situation of reporting toxicity that may not actually exist, and later having to refute that certification in an enforcement action. DMR challenges generally have been rejected by the courts on the grounds that:

> …[D]ata reported on DMRs may be deemed admissions of liability even where the DMRs are submitted with comments disputing the accuracy of the reports … Moreover, reliance on DMRs to establish liability is consistent with the legislative history and avowed policy of the CWA …
>
> While respondent has presented credible evidence calling into question the reliability of test results from its contract lab, respondent's arguments are ultimately unavailing. Respondent reported the data, certifying it as 'true, accurate and complete' on the DMRs, albeit with 'qualification' or reservation manifested in the comments on the DMRs and cover letters … The legislative history of the CWA as noted above and the required certification on the DMRs, emphasize the need for accurate reporting and simple enforcement, and evidence Congress' and EPA's intent to place heavy reliance on data reported on DMRs in the context of enforcement. Thus, in order to balance such heavy reliance, and not withstanding its 'qualification' of reported data, respondent bears a heavy burden to show laboratory error, in order to prevail under the preponderance of evidence standard of 40 C.F.R. 22.24 … To meet that burden, respondent must show that there were

[78] Sutfin, Charles S., et al., U.S. EPA Office of Water, Memorandum to EPA Regional Water Management and Enforcement Division Directors, *Certification of "Accuracy" of Information Submissions of Test Results Measuring Whole Effluent Toxicity* (March 3, 2000).

errors in the actual tests performed.[79]

No court, however, has dealt with this issue where the permittee argues that the measured result, even though in excess of the permit limit, does not confirm the toxicity of the effluent. In any event, dischargers should not be subjected to the burden of proving a negative (i.e., proving the lack of accuracy). EPA must prove accuracy before its test methods are approved under Part 136. Approving test methods under Part 136, such that they can be used to set and enforce permit limits, is akin to authorizing state troopers to prosecute alleged speeding violations based on readings from radar guns that have not first been calibrated (to assure that a reading of say 60 miles per hour provides a high level of confidence that a motorist is traveling at a speed of at least 56 miles per hour in a 55 mile per hour zone).

If the accuracy of WET tests is unknowable, then the validity of any given result is also unknown. And, regulating without regard for the accuracy of information used to justify those enforcement decisions is the definition of an arbitrary and capricious action.

3. EPA Acknowledges the Poor Level of Accuracy Expected for WET Test Methods.

While EPA has stated that it is impossible to demonstrate the accuracy of any given WET test result, for the NOEC endpoint, the Agency is able to calculate the likely "range" in results. In the procedure manual for each WET test method, EPA states:

> It should be noted here that the dilution factor selected for a test determines the width of the No-Observed-Effect-Concentration and the Lowest Observed Effect Concentration Interval and the inherent maximum precision of the test ...With a dilution factor of 0.5, the NOEC could be considered to have a relative variability of

[79] *In the Matter of: City of Salisbury, Maryland*, EPA Administrative Law Judge Division, Docket No. CWA-III-219 (February 8, 2000).

plus or minus 100%.[80]

Accepting a tolerance range of "plus or minus 100%" is a tacit admission that the WET test methods are incapable of reliably distinguishing between a toxic and a non-toxic sample.

There are sources of inaccuracy beside the fundamental absence of a traceable standard that defines "toxicity" with which to "calibrate" WET tests. For example, EPA states that the rate of probable decision error is set, in advance, when one chooses the level of statistical confidence used to analyze the WET test data.[81] For most toxicity tests, the threshold of statistical significance is set at the 95% confidence level. By design, therefore, approximately 1 in 20 tests will appear to exhibit "toxicity" even when none is present. Other sources of test variability, to be described later in this document, will likely drive the actual rate of false positives much higher.

While a 5% error rate may sound fairly low, it must be evaluated against the very large number of tests that dischargers are likely to run. Many dischargers now perform monthly WET tests. Over the five year permit term, they will run at least 60 toxicity tests. Given the expected error rate and simple algebra, we can calculate that they will observe at least three toxicity test failures during that period. More importantly, only 4.6% of all dischargers with non-toxic effluent will record no WET test failures after five years of monthly analyses using only one species and one chronic test endpoint (e.g., reproduction or growth). Only one-third of all dischargers, with non-toxic effluent, will record no WET test failures after five years of "quarterly" sampling. In either case, the vast majority of all dischargers (66-95%) will appear to

[80] Chronic Freshwater Manual (Section 4.14.6) at 16.

[81] U.S. Environmental Protection Agency, *Guidance for Data Quality Assessment: Practical Methods for Data Analysis* (EPA-QA/G-9), EPA/600/R-96/084 (July 2000).

have toxic discharges regardless of actual effluent quality.

The probability of error increases if the discharger runs more tests on the same species, utilizes additional species, or analyzes multiple endpoints (e.g., survival and reproduction) in a single test. Such inaccuracy renders the test methods inappropriate for their intended purposes under Part 136. In particular, the design of the test precludes the user from ever being able to certify the absence of toxicity.

4. WET Test Methods Produce Significant Errors.

Peer-reviewed studies of EPA's recommended statistical procedures indicate that the actual incidence of false positives (reports of toxicity when none actually exists) is likely to be 3-5 times higher than EPA estimates.[82] This suggests that decision errors occur in 15% to 25% of all tests rather than the 5% predicted by EPA. The experts charged with peer-reviewing the interlaboratory variability study agreed:

> … the actual level of false positives in 'real life' as defined by this study can be expected to be higher. These tests are applied, too often, as decisive when they are far from such.[83]

Non-toxic water can serve as a traceable standard with which to evaluate the rate at which test methods will measure toxicity that is known to be absent (i.e., to evaluate accuracy). Independent studies of actual laboratory performance, using only blind samples of non-toxic

[82] Dhaliwal, B.S., R.J. Dolan, and R.W. Smith. 1995. A proposed method for improving whole effluent toxicity data interpretation in regulatory compliance. *Water Environ. Res.* 67:953-63.

[83] U.S. Environmental Protection Agency, *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 by Versar, Inc. (March 2001) ("Peer Review Report"), pp. 16 and 18.

dilution water, confirm the statistical analyses and peer-reviewer predictions. One such study found that 43% of the non-toxic samples were reported as toxic – an error rate 8-times higher than expected.[84]

The high error rate was confirmed in EPA's own method validation studies.[85] For example, during a formal study of interlaboratory performance, EPA's contractor initiated 38 reference toxicant tests using the chronic method for *Ceriodaphnia dubia*. In those tests, the laboratories spiked "clean" water samples with a concentration of a chemical expected to cause toxic effects to test organisms. Two-thirds of the participating laboratories reported that the sample spiked with a toxin was "non-toxic." EPA asserts that the sample may have been only "marginally toxic" and the individual results depended on the specific test sensitivity at each lab. However, review of the raw data indicates that 11 of the 13 most sensitive valid tests declared the sample to be non-toxic and 8 of the 13 least sensitive valid tests found the sample to be toxic.[86] Statistical re-analysis of EPA's data shows that the probability of passing a test increased as the test became more sensitive. If the sample were truly toxic, however, one would expect the most sensitive tests to detect it first and fail. The best explanation is that the reference samples were not really toxic. Thus, the one-third of all labs that reported toxicity were in error,

---

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

[85] U.S. Environmental Protection Agency, *Final Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods, Vol. 1*, EPA 821-B-01-004 (September 2001) ("WET Study Report").

[86] "Sensitive" tests are better able to identify smaller changes in survival, growth or mortality as "statistically-significant." Such tests are better able to detect toxicity when it is actually present, but are also more likely to misclassify small random variations in survival, growth or reproduction in non-toxic samples as toxicity.

and EPA's results are very similar to the studies described earlier in this sub-section.[87]

Even if one assumes that the samples were toxic, the fact that two-thirds of the labs were unable to detect it affirms the conclusion that the test method is unable to distinguish toxic samples from non-toxic samples and is, therefore, inappropriate for use in the context of gauging compliance with NPDES permit limits.

5. Inaccurate WET Test Methods Result in Unacceptable Impacts.

Inaccurate WET test methods undermine the entire NPDES permitting system. Such tests make it impossible to determine compliance or noncompliance with confidence. Moreover, uncertainty in test results confound the discharger's ability to identify the cause and source of true toxicity when it occurs.

False permit violations may result in inappropriate fines and unjustified public criticism. At a minimum, the need to follow up each test failure (real or not) will significantly increase the cost of testing. Each additional test costs approximately $1,000, and a formal Toxicity Identification Evaluation ("TIE") will cost at least $15-25,000. Several dischargers have spent more than $200,000 chasing "phantom toxicity." This does not include the cost of the plant personnel that participate in, or manage consultants hired to perform, the TIE.

Accuracy also is important from the environmental perspective, because an inaccurate test may fail to identify toxicity when it is truly present. Test methods must be reasonably related to the parameter they are used to regulate. If the accuracy of WET methods cannot be

[87] Risk Sciences, Test Sensitivity for Ceriodaphnia Reproduction Using Reference Toxicants: EPA's Whole Effluent Toxicity Interlaboratory Variability Study (2001).

demonstrated, then the methods must not be included in Part 136.

One cannot defend an inaccurate test solely on the basis that it's better to be safe than sorry. Inaccurate WET tests are no more useful than an over- or under-sensitive automobile airbag and may be just as counter-productive.

6. Test Precision is an Unacceptable Substitute for Accuracy.

When the WET test methods were promulgated, EPA stated that:

> Accuracy of toxicity test results cannot be ascertained, only the precision of toxicity can be estimated.[88]

Since then, EPA has often defended WET test methods by comparing the precision of those procedures to the level of precision in commonly-accepted chemical test methods. However, the degree of precision is irrelevant to the question of accuracy. According to EPA:

> Precision is used to describe the reproducibility of results … Precision refers to the agreement among a group of experimental results and implies nothing about their relationship to the true value.[89]

The fact that a doctor performs surgery with great precision is irrelevant if she amputates the wrong limb. The fact that a pilot executes a perfect landing does little to promote safety if he lands on the taxiway instead of the runway. Precision is no surrogate for accuracy. Even if it were, there is no evidence to indicate that WET test methods exhibit adequate precision. Independent research studies prove quite the opposite:

> The results of this study show that both intra- and interlaboratory

[88] 60 Fed. Reg. at 53,535.

[89] U.S. Environmental Protection Agency, *NPDES Permit Writer's Guide to Data Quality Objectives* (November 1990) ("Permit Writer's DQO Guide"), p. 1-6.

> variance from tests with reference toxicants are often well above the limits that would be considered acceptable by most scientists and permit holders. Combining these sources of variability exacerbates the problem, and the fact that each test point estimate of effect is itself uncertain, due to intra-treatment variance and lack of model fit, has not yet been considered … This study shows that permit toxicity limits could be exceeded because of factors other than effluent toxicity….[90]

B. EPA Did Not Demonstrate Acceptable Precision for the WET Test Methods

1. EPA Did Not Establish a DQO for Minimum Acceptable Precision.

As with accuracy, EPA has a duty to establish thresholds for minimum acceptable precision. The Agency is required to define tolerance limits for and maximum allowable imprecision. Without formal DQOs, there is no objective standard by which to determine whether the level of precision observed in WET testing is appropriate for its intended purpose.

EPA has long suggested that WET test precision may be deemed adequate if it falls within the range of precision recorded for chemical analyses. However, this "bootstrap" argument fails, as EPA never defined tolerance thresholds or DQOs for "unacceptable imprecision" in evaluating the chemical methods either.

2. EPA Did Not Validate Precision for all Endpoints.

What is commonly referred to as "toxicity testing" is actually a large collection of separate test methods using different species, different exposure regimes, different biological endpoints and different statistical endpoints. Each endpoint may be used, independently from any other endpoints, to establish a permit limit for toxicity. Each individual endpoint that may be used to determine compliance must be validated for that purpose. The data presented in

[90] Water Environment Research Foundation, *Whole Effluent Toxicity Testing Methods: Accounting for Variance*, Report #D93002 (1999) ("WERF Variance Report"), pp. 3-24 and 3-25.

EPA's Interlaboratory Study report confirm that precision varies considerably from one endpoint to the next. It therefore is arbitrary to presume, as EPA apparently has, that it is unnecessary to determine the precision each endpoint will exhibit.

EPA failed to validate precision for several endpoints that are routinely included in NPDES permits. For example, EPA failed to evaluate the precision of a No-Observed-Acute-Effect-Concentration ("NOAEC"). EPA evaluated only the LC_{50} and the EC_{25} for mortality endpoints. Many states now use the NOAEC rather than the LC_{50} or EC_{25} to define the threshold for acute toxicity.

Some states (e.g., Virginia) require permittees to use the NOAEC because it is included among the recommended endpoints in EPA's manuals for the acute methods.[91] It is arbitrary to retain such endpoints in Part 136 until adequately validated. It also is arbitrary to subject dischargers to enforceable permit limitations where compliance will be based on a test method whose precision is unknown. States that impose such NOAEC limits will be vulnerable to permit challenges until EPA validates the NOAEC endpoint and includes precision estimates in Part 136. The same is true for any state that intends to use the LC_1 as a compliance threshold in an NPDES permit.

3. EPA Acknowledges the Poor Precision of WET Test Methods.

Data from EPA's own DMR-QA studies demonstrate the lack of precision.[92,93] *See*

[91] U.S. Environmental Protection Agency, *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*, 4th Ed., EPA/600/4-90/027F (August 1993).

[92] *See* Permit Writer's DQO Guide at 3-1 (describing the annual DMRQA program).

[93] One EPA report states "WET methods as tested in DMR-QA studies are twice to four

Table 1. Those data indicate that only about half of the laboratories testing the same sample will provide a nearly identical estimate of the NOEC. One-fourth would report a higher concentration than the true NOEC and the other fourth would report a lower concentration than the true NOEC. Thus, about half of the laboratories reported toxicity estimates that were off by more than a factor of 2.[94]

Table 1: Variation in Reported NOEC Value w/ Reference Toxicants

Ceriodaphnia dubia **Reproduction as NOEC**	**N of Labs**	**Median Value**	**95% Conf. Range**
EPA DMR-QA #12 (1992)	103	20.0	2% – 50%
EPA DMR-QA #13 (1993)	124	25.0	6% – 50%
EPA DMR-QA #14 (1994)	147	25.0	6% – 50%
EPA DMR-QA #15 (1995)	147	25.0	6% – 50%
EPA DMR-QA #16 (1996)	140	25.0	6% – 50%

EPA sponsored the DMR-QA studies to evaluate the ability of dischargers and laboratories to perform standard methods correctly. Results from these tests were deemed "acceptable" if they were within plus or minus one concentration interval in the test dilution series.[95]

Such error bands would be acceptable if they were used also when data are analyzed to determine compliance. However, while the laboratories are allowed to vary by plus or minus 100%, EPA recommends that the variability be ignored when using the data to certify

---

times as variable as chemical analyses." Lazorchak, J.M., P.W. Britton, M.E. Smith, and J.D. Helm. 1997. Summary and Methods Variability Issues of the U.S. EPA Discharge Monitoring Report Quality Assurance Program (DMRQA) Whole Effluent Toxicity Testing (WETT) from 1991-1997. U.S. EPA, Cincinnati, OH.

[94] *See* EPA's DMR-QA Study Results at www.epa.gov/ORD/dbases/PES/index.html.

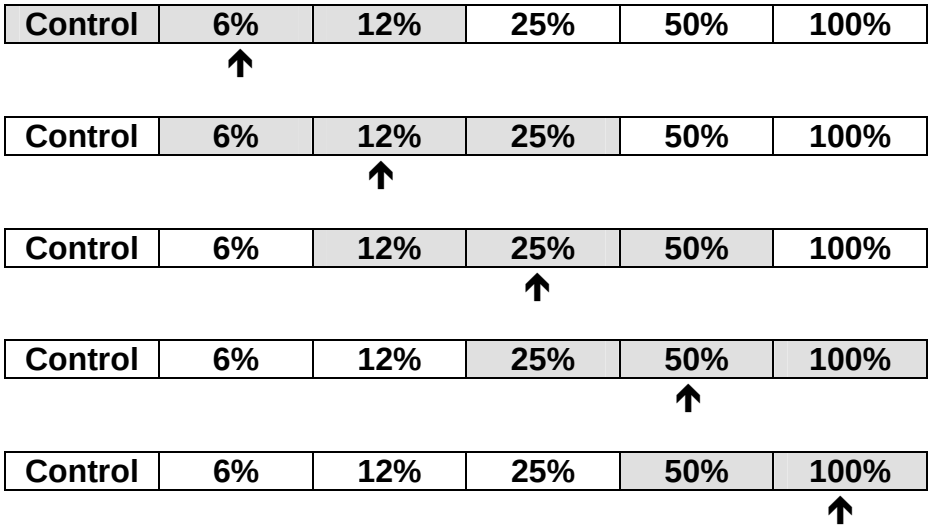
[95] Chronic Freshwater Manual (Section 9.3.1.1) at 38.

compliance with an NPDES permit limit.[96] If EPA's desire to use WET test methods is strong enough such that it is willing to deem acceptable the associated imprecision, it must establish formally a procedure to ensure that permittees are not penalized for the error in reasonable potential and compliance determinations.

Figure 1 illustrates the standard (0.5) dilution series used in most WET tests. Control organisms are exposed to non-toxic dilution water. Organisms exposed to undiluted effluent are in the "100% treatment group." The arrow indicates the "true" threshold of toxicity. The shaded boxes represent the range of results that would be deemed within the range of acceptable precision error.

[96] *See* WET Variability Guidance.

Figure 1: Imprecision Allowed in EPA's DMR-QA Studies



The concept is easier to understand if viewed, by analogy, as a gas gauge. With a half-full tank, the gauge would indicate somewhere between full and one-fourth full. With a full tank the gauge would read as much as half empty.

Thus, a sample that is not-toxic when diluted to one-quarter strength (NOEC = 25%) may, when tested, indicate the presence of toxicity anywhere in the continuum from 12% effluent to 50% effluent (e.g., "plus or minus 100%"). A sample that is not toxic at all (NOEC = 100%) may, nevertheless, indicate that it is toxic until diluted to one-half strength.

EPA recently published new guidance defining the expected precision, expressed as the average coefficient-of-variation ("CV"), for each major test method.[97] EPA demonstrates the relatively poor precision of the WET test methods.[98]

[97] WET Variability Guidance at A-5.

[98] *Id.* at pp. 3-4 and 3-5; *see also* WET Study Report at xiv.

Each of the freshwater test species exhibited a CV of about 0.26 for the sublethal endpoint.[99] If a radar gun exhibited a similar CV, and 100 state patrol officers each measured a vehicle traveling at exactly 55 mph, one-third of the officers would observe a speed in excess of 61 mph, one-fourth of the officers would record a speed greater than 64 mph, 5% of the officers would believe the car was traveling nearly 78 mph, and one officer would conclude the car was speeding more than 88 mph.[100]

Furthermore, EPA's estimates of the CV are only the "median" value observed across many labs. The actual CV varies by a factor of 2-3x between laboratories. For example, EPA reports that 50% of the labs they studied recorded a CV between 0.17 and 0.45 for one measure of *Ceriodaphnia* reproduction; 25% of the labs had CV values greater than 0.45.[101] This means the actual level of imprecision can sometimes be far worse than described by the radar gun analogy above.

EPA relied on the <u>average</u> CV to suggest that the precision of WET methods was acceptable. It is not enough for a method to perform adequately on average. It also must do so routinely. This is particularly true because dischargers generally are not permitted to average the results of multiple WET tests. Each WET test must pass independently. It is impossible to record consistent compliance using a test that is allowed to be so inconsistent and imprecise.

---

[99] Fathead minnow growth, Ceriodaphnia dubia reproduction, *Selenastrum capricornutum* (algae) cell density.

[100] Max measured speed = mean speed * (z * std dev); where std dev = CV * mean and z is defined by a probability value in a normal distribution (usually found in a table in the appendix to any introductory textbook on statistics).

[101] WET Variability Guidance at 3-4. The Agency has admitted that a CV of 0.50 is "quite high." DQO Guidance at 6-10.

Experts charged with peer-reviewing the results of EPA's interlaboratory study agreed:

> … there is much more variability occurring in a regulatory sense than is 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 interval of the median is misleading and not useful.[102]

By comparison, assume that a driver with a true blood alcohol content ("BAC") of 0.05 was asked to submit to a sobriety test. If the breathalyzer had a CV of 0.4, the reported result would vary between 0.01 and 0.09 (w/ 95% confidence). In this case, the driver faces an unacceptable risk of being falsely charged with DUI despite have a true BAC well below the legal limit of 0.08%.

Moreover, as EPA has stated, "[t]he assumption that WET precision will vary among toxicants is also logical."[103] The variability differences for the different matrices used in EPA's Interlaboratory Study confirm that position. EPA has not explained why the matrix type influences variability and under what circumstances the variability estimates derived from its study may not be representative of the myriad of effluent types on which its Part 136 test methods will be used.

Finally, it is arbitrary to suggest that a CV of say 0.3 is acceptable for WET testing just because EPA has accepted it for some chemical methods (as discussed in the next subsection). Any common appliance with a CV as high as 0.3 would be deemed defective and quickly

[102] Peer Review Report at 16.

[103] WET Variability Guidance at 4-2.

replaced. Household thermostats with a CV of 0.3 only hold temperature somewhere between 29º and 111º when set to 70º. ATM machines with a CV of 0.3 would dispense somewhere between $42 and $158 when $100 was requested. A grocery store scale or gas pump that exhibited a CV of 0.3 would be prohibited by the Bureau of Weights and measures in every state. Surely, precision is more important when measuring toxicity than when weighing lettuce.

That position was reached in a recent WERF report, which stated:

> [I]n conclusion, to ensure a fair regulatory process, intra- and interlaboratory variation in WET test results must be explicitly incorporated into the evaluation of compliance with WET limits. It is recommended that … protocols that show poor performance (low comparability and reproducibility) be dropped from the decision-making process until good performance can be assured or until additional test acceptability criteria to reduce variability are adopted.[104]

4. Comparisons to Chemical Test Precision are Irrelevant.

EPA routinely claims that the WET test precision level is acceptable because it is comparable to the level of precision observed in chemical test methods. This is true only if the best WET tests are compared to chemical methods operating at concentrations where imprecision is highest.[105] Regardless, the impact of that analytical variability on regulatory decision-making is more significant for WET tests than chemical tests. Very few chemical pollutants are

[104] WERF Report, p. ES-2 and ES-3.

[105] EPA cites to the TSD for support. The TSD says the coefficient of variation ("CV") for manganese is 129%. But when you trace that CV back to its origin (a 1983 EPA methods manual), you find that it applies to Method 243.1. The CV was calculated with interlaboratory study data for tests performed at a concentration of 11 ug/l. In that very same method description, however, EPA specifies that for measurements below 25 ug/l, Method 243.2, rather than Methods 243.1, should be used. Thus, the variability to which EPA now cites, with approval, would be expected only if the test were run below the concentration at which EPA deems the test method to apply. In short, EPA's statement is erroneous and misleading.

regulated to the same zero-tolerance standard as WET. According to EPA:

> Analytical precision varies over the range of a procedure and is worst near the detection limit.[106]

Furthermore, unlike WET testing, it is possible to independently corroborate the accuracy of a chemical analysis. Imprecision is more tolerable where the impact on accuracy is well-defined. For example, a chemical test also may have a high coefficient-of-variation, but because we can calibrate the results against traceable standards (i.e., known concentrations), we can easily calculate whether the resulting imprecision makes it impossible to determine (with confidence) whether the measured value is higher or lower than the permit limit. For example, if the permit limit is 10 ppb and the measurement is 20 ppb, when the precision is known to be $\pm$ 5 ppb, and the bias is 103%, the excursion is a clear violation. However, where accuracy is unknown and unknowable, as is the case with WET methods, greater precision is essential to support informed decisions because toxicity is a method-defined parameter that cannot be corroborated by any other test procedure.

5. WET Test Imprecision Results in Unacceptable Impacts.

The implications of inadequate precision are profound. When dischargers split identical effluent samples between two laboratories, they are likely to receive different results, perhaps vastly different. Sometimes the difference is merely one of degree (high toxicity vs. low toxicity). More often, the difference is one of conclusion (toxic vs. not-toxic). Courts have ruled that such disparate results in split samples is evidence of analytical error and inaccuracy.[107]

---

[106] Permit Writer's DQO Guide at 1-6.

[107] *Public Interest Research v. Elf Atochem*, 817 F. Supp. 1164, 1180-81 (D.N.J. 1993) (holding laboratory error is a partial defense to liability under the CWA).

Imprecision is particularly vexing when a large number of WET tests are run simultaneously during a TIE. Inconsistent results confound interpretation of test data and frequently delay a successful conclusion to the TIE. As a result, the discharger is unable to verify whether toxicity is actually present and, if it is, to identify the most likely cause. Finally, test imprecision makes it more difficult to confirm that toxicity has abated. According to EPA:

> Accurate, reliable stressor identification procedures are necessary for EPA and the states/tribes to accurately identify the cause(s) of water quality standards violations … Accurate stressor identification can be very critical in NPDES permitting cases, both for fairness and success in stressor control … A high degree of accuracy and reliability in the stressor identification process is necessary.…[108]

If a method is intended to reliably assess compliance with NPDES permit limits, test results must be tied to actual effluent quality. EPA warns that the health of stock culture organisms may significantly affect the sensitivity and outcome of any given WET test.[109] Variations in test sensitivity can cause the compliance threshold to become unstable.

In one test, a 23% reduction in reproduction or growth may be insufficient to cause test failure. In a subsequent test using the same test species but a different batch of organisms, an 18% reduction in reproduction or growth may be deemed statistically-significant. Thus, “better" effluent quality may lead to a permit violation, while poorer quality is regarded as compliant. The inconsistency in outcomes is due solely to differences in test sensitivity, not to changes in effluent quality.

---

[108] U.S. Environmental Protection Agency, *Stressor Identification Guidance Document*, EPA 822-B-00-025 (December 2000), pp. ES-2 and 1-7.

[109] 66 Fed. Reg. at 49,797 (col. 3).

In an enforcement context, compliance cannot be based on anything other than the actual effluent quality compared to one consistent standard of performance. It is up to the permit writer to determine that standard of performance. And, the standard should remain consistent from month to month and between individual tests, just as it does for any chemical pollutant.

Whole effluent toxicity is a method-defined parameter designed to regulate the magnitude of "effect" on the environment. As such, it is inappropriate to allow the level of "effect" that constitutes a permit failure to vary from test to test as it does in the proposed methods. The level of imprecision inherent to the WET test methods renders them unsuitable to the purpose EPA intends (e.g., determining compliance based on the results of a single toxicity test result).[110]

For any chemical test method, EPA would reject results where adequate accuracy and precision could not be demonstrated.

> If either the precision or accuracy test is failed, the test must be repeated until the laboratory is able to meet the precision and accuracy requirements.[111]

The application and use of WET test results must be strictly limited for the same reason.

> In the absence of a record supporting the trustworthiness of agency decision-making tools as they were applied, we cannot uphold those tools' application … No matter how sophisticated or involved the methods employed by EPA in reaching its decisions, in order to uphold those decisions under the Clean Air Act we must

[110] U.S. Environmental Protection Agency, *Whole Effluent Toxicity (WET) Control Policy*, EPA 833-B-94-002 (July 1994).

[111] U.S. Environmental Protection Agency, *Guidance on Evaluation, Resolution, and Documentation of Analytical Problems Associated with Compliance Monitoring*, EPA 821-B-93-001 (June 1993), p. 11.

be able to see that the Agency's actions were not arbitrary.[112]

C. EPA Did Not Define the Dynamic Range or Establish a Detection Limit for Each WET Test Method.

1. EPA Must Define Tolerance Limits to Minimize Decision Error When Test Methods are Used.

Detection limits and dynamic range are closely related concepts:

> In general, methods can only be used to measure analytes over a specified concentration range, defined as the linear dynamic range. The linear dynamic range is limited at the lower level by the detection limit ….[113]

It is necessary to identify the dynamic range and detection limits in order to establish boundaries around the tolerable level of decision error when using any given test method. EPA recommends use of the DQOs process to define those tolerance limits:

> When data are being used in decision-making by selecting between two alternative conditions (e.g. compliance/non-compliance with a standard), the Agency's recommended systematic planning tool is called the DQO Process … DQOs define the performance criteria that limit the probabilities of making decision errors by considering the purpose of collecting the data; defining the appropriate type of data needed; and specifying tolerable probabilities of making decision errors.[114]

A critical part of the DQO process is to choose analytical methods that are appropriate for the intended decision. In particular, EPA concurs that the method must be capable of measuring the relevant parameter within the specified error tolerance:

> In Step 5 of the DQO process … select the measurement and

---

[112] *Ohio v. United States Environmental Protection Agency*, 784 F.2d 224, 230-31 (6th Cir. 1986).

[113] Section 518 Report at 3-4.

[114] DQO Guidance, pp. 0-4-0-7.

> analysis methods capable of performing over the expected rate of values and … determine the detection limit for each potential measurement method … [115]

EPA failed to define an objective performance criteria for acceptable dynamic range related to WET test methods. Nor did the Agency establish a detection limit for these methods. Therefore, it cannot be shown or assumed that the promulgated methods are suitable for their intended purpose.

2. The Level of Natural Biological Variability Inherent to WET Test Methods Necessitates That the Dynamic Range and Detection Limits be Clearly Defined.

Like all living things, growth, reproduction, and lifespan vary among individual organisms in any given test species. The purpose of toxicity testing is to distinguish differences caused by pollutants from those that occur naturally. The sensitivity and utility of the test depends on controlling background variability during the statistical analysis, thereby eliminating confounding factors.

In order to avoid decision errors, the method must be capable of detecting toxicity when it is actually present, and the method must not indicate toxicity when it is actually absent.[116] EPA defines the point at which a method is capable of minimally acceptable precision as the "detection limit." Historically, when establishing detection limits for chemical analysis methods, EPA required at least 99% confidence (only 1% risk of error):

> … Detection Limit refers to the minimum concentration of an

[115] *Id*. at 5-2 and 5-4.

[116] U.S. Environmental Protection Agency, *Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing (40 CFR Part 136)*, EPA 821-B-00-004 (July 2000) ("WET Testing Guidance").

> analyte that can be measured and reported with a 99% confidence that the analyte concentration is greater than zero.[117]

Data from EPA's WET interlaboratory validation studies indicate that the level of natural variability is often very high for some test endpoints (see Table 2). Note that the variability is greatest for sub-lethal metrics such as *Ceriodaphnia* reproduction or Fathead minnow growth. For example, while the average water flea may produce approximately 23 offspring in a one week period, any given test organism may produce as many as 49 or as few as zero.

Table 2: Range of Natural Variability When Exposed to Non-Toxic Water[118]

Test Species For Chronic Method	Number of Measurements	Percent Mortality	Median Weight or Reproduction	99% Range of Performance
Fathead minnow	920	36%	0.47 mg/fish	0.21 – 0.89 mg/fish
Ceriodaphnia dubia	2960+	6%	23 offspring per female	0 – 49 offspring per female

If zero reproduction is within the normal dynamic range of the test organism even when exposed only to non-toxic dilution water, it is very difficult to determine when reductions in reproduction are the result of pollutants rather than natural causes.[119] Similar data from other species and biological endpoints evaluated in EPA's WET interlaboratory study demonstrate the same natural variability in measured endpoints and the same need to define the dynamic range and establish appropriate detection limits.

---

[117] 40 C.F.R. § 136.2(d).

[118] WET Study Report (raw data provided electronically by EPA and re-analyzed by Risk Sciences).

[119] *See* Risk Sciences, Memorandum to Jim Pletl re: Power Analysis (December 26, 2001).

If the performance (median weight or reproduction) of test organisms varies by plus or minus 100% of median, as EPA's data indicate, then it is essential to know when the observed difference is greater than one expects to occur naturally. It is necessary to define the dynamic range and detection level of each method to account for the natural background variability of each test species and endpoint; otherwise, small changes in survival, growth, or reproduction may be mistaken for evidence of toxicity, and the risk of decision error becomes unacceptable.

3. EPA Erred by Failing to Establish a Detection Limit for WET Methods.

All of EPA's recent test methods in Part 136 include a minimum detection limit ("MDL") established as the detection limit laboratories must demonstrate their ability to achieve before being eligible to perform analyses for regulatory application. EPA provides no such detection level concept in any of the WET methods. EPA apparently believes that the concept of a detection limit applies only to analytical methods that rely on mechanical instrumentation to measure pollutant concentrations. It claims it is impossible to establish a detection limit for biological organisms. Even if that were true (which is not the case[120]), it would not relieve the Agency from the responsibility to provide the same level of protection against decision errors in WET testing that detection limits ensure for chemical methods.

If EPA is correct in stating that the variability of WET tests is similar to that observed in chemical analyses, then the risk of decision errors is also likely to be similar. It is essential to minimize the risk of error by properly accounting for intrinsic analytical uncertainty when reviewing data from a toxicity test.

[120] A paper offering recommendations for calculating detection limits for WET test methods is attached. *See* Risk Sciences, Developing A Detection Level for Whole Effluent Toxicity (WET) Testing (2002).

> Specifying a statistical confidence threshold for calculating test results (e.g. 95% or 99%) does not take the place of establishing a detection limit. The statistical confidence level (aka "alpha') reduces the risk of inferential error within any single test, but it does not address the issue of variability between tests even on identical samples. This is contrary to EPA's own guidance:
>
> Between laboratory differences in test sensitivity are important and need to be addressed.[121]

EPA provided no process or mechanism to manage the uncertainty caused by inter-test variability. Independent research studies have investigated the impact that such variability has on decision errors:

> … incorporating uncertainty into the decision-making process for WET limits would significantly affect interpretation of WET test results.[122]

WET test methods, as promulgated, are incomplete because EPA failed to define an acceptable dynamic range and establish an appropriate detection level. The methods do not meet the standard-of-performance established for chemical test procedures despite EPA's admission that the level of variability is approximately equivalent.

[121] WET Variability Guidance at G-7.

[122] WERF Variance Report at ES-2-ES-3.

D. EPA Did Not Establish Procedures to Correct for Sources of Test Interference.

1. Extraneous Factors, Other Than Chemical Pollutants, Interfere With Toxicity Tests.

EPA's test procedure manuals warn that small changes in pH, temperature, hardness, culture health, and diet are all known to effect test sensitivity.[123] Standardized methods are designed to control for such factors. However, EPA's own studies demonstrate that significant variations in reported toxicity levels continue to result within and between labs on identical split samples.[124,125]

Similar results were reported by the Water Environment Research Foundation ("WERF"). WERF's research found that more than half of the apparent variation in WET test results is due to factors unrelated to actual effluent quality (such as choice of lab, the particular lab technician, the health of the test cultures, and even the season of the year).[126]

Among the most common sources of WET test interference are: ionic imbalance, pH drift, and pathogen contamination.

Some of the designated test species are particularly susceptible to the ionic chemistry of natural waters. Indicator species cultured in one kind of water may not fare well when exposed to water with a fundamentally different structure. For example, the freshwater of many southern states is naturally low in hardness. The lack of calcium may suppress normal levels of

---

[123] *See* Chronic Freshwater Manual.

[124] *See* WET Variability Guidance.

[125] *See* WET Study Report.

[126] WERF Variance Report at 3-3 - 3-4.

reproduction in *Ceriodaphnia dubia*. In some cases, the water is so low in ions that it causes mortality.[127] In such instances, it is not pollution causing toxicity – it is the lack of chemicals essential to aquatic life.

In western states, standardized test species may be incapable of tolerating the naturally high mineral content of western stream water. *Ceriodaphnia* have a relatively narrow tolerance range for hardness and conductivity.[128] Even differences in the balance between major anions and cations can interfere with the normal reproductive cycle of sensitive invertebrate species.[129]

EPA is aware that natural ionic chemistry can interfere with WET test results. In fact, the Agency warns laboratories not to use some well waters as a dilution source because of this problem.[130] The extremely low levels of natural minerals also may cause rain water to fail a toxicity test.

These problems suggest that WET test procedures may not be suitable for evaluating all analytes. EPA failed to describe such limits on the applicability of the method as required by the Agency's own guidance.[131] Interference problems may undermine the utility of WET methods as a tool for testing groundwater, storm water, and streams with relatively high or low conductivity. EPA warned users of these effects, but failed to specify limits on the application of

---

[127] Goodfellow, W.L., et al. 2000. Major ion toxicity in effluents: a review with permitting recommendations. *Environ. Toxicol. Chem.* 19:175-82.

[128] *See* Chronic Freshwater Manual.

[129] Goodfellow, et al. (2000).

[130] *See* Chronic Freshwater Manual.

[131] U.S. Environmental Protection Agency, *Guidelines and Format for Methods to be Proposed at 40 CFR Part 136 or Part 141,* (July 1996) (Draft), p. 15.

the method due to known interference problems.

pH shock is another source of potential interference with WET tests. Test organisms are often cultured in water with a pH near 8.0. Effluent frequently is discharged at a pH between 6.5 and 7.0. Although such discharges comply with NPDES permit limits, they may fail a WET test. Organisms transferred from culture water to effluent experience an instantaneous pH shock. Such shocks may exceed the general tolerance of the test organism. Worse, biochemical interactions between the test organisms and the sample water tend to cause pH to drift up over time (sometimes more than 1 standard unit in 24 hours). When the sample water is renewed each day, the test organisms again are shocked by instantaneous change in pH when they are transferred from the old effluent sample to the new effluent sample. During the week-long chronic test, the process is repeated seven times, causing further shock to test organisms. The end result may cause mortality or inhibit (or delay) normal reproduction cycles.[132] However, the effect is entirely an artifact of the test design and bears no relationship to real world conditions. EPA's proposed revisions to the methods discuss controlling pH after a test begins but do not describe how to adjust or account for the phenomena described above.

Interference caused by ionic chemistry and pH-shock usually affect only the invertebrate test species. Pathogen interference is more common among WET tests performed on fish species.[133] This phenomenon is most commonly observed in once-through cooling facilities.

---

[132] Cruze, R. 1993. Effects of pH variation on chronic toxicity test reliability. Riverside Regional Water Quality Control Plant, Riverside, CA.

[133] Downey, P.J., et al. 2000. Sporadic mortality in chronic toxicity tests using *pimephales promelas* (Rafinesque): cases of characterization and control. *Environ. Toxicol. Chem.* 19: 248-255; Kszos, L.A., A.J. Stewart, and J.R. Sumner. 1997. Evidence that variability in ambient fathead minnow short-term chronic tests is due to pathogenic infection. *Environ. Toxicol. Chem.* 16:351-356; Grothe, D.R. and D.E. Johnson. 1996. Bacterial interference in

Pathogens found in natural stream waters often cause mortality (unrelated to toxicity) to standard test organisms.[134] Side-by-side toxicity tests often indicate that the effects are worse in upstream water prior to diversion than in discharged effluent. Mortality caused by pathogen interference adversely impacts the survival endpoint. It also may impact adversely the biomass endpoint because the mean weight of the test organisms is calculated by dividing the total weight of a replicate by the number of original organisms. Here again, EPA has acknowledged the problem but has failed to modify the mandatory test protocols to account for such interference when it occurs.[135] While EPA mentions that "dual controls" may be used to characterize pathogenic toxicity, the Agency does not describe how to report such WET results when assessing compliance status for an NPDES permit limit.[136] EPA also is proposing a method modification to control pathogen interference that has not been sufficiently validated and has not been proven to control interference.

In each instance described above, the WET test is measuring a genuine stressor. However, pollution is not the source or cause of the "pseudo-toxic" effects observed. If interference remains uncontrolled, the methods are incapable of distinguishing true toxicity from artifactual anomalies and are unsuitable for their intended purpose.

---

whole-effluent toxicity tests. *Environ. Toxicol. Chem.* 15:761-764.

[134] Chronic Freshwater Manual at 128.

[135] *See* Proposed Method Manuals Changes.

[136] WET Testing Guidance at 2-3.

2. Some Method Requirements Interfere With the Validity of the Data Analysis.

Some procedures may unintentionally bias the results. For example, EPA established mandatory Test Acceptance Criteria ("TAC") in an attempt to address test sensitivity. If Fathead minnow control organisms exhibit less than 80% survival or weigh less than 0.25 mg/fish at the end of a chronic test, the test is deemed invalid and must be repeated. A chronic *Ceriodaphnia* test must be repeated if control mortality is greater than 20% or reproduction is less than 15 offspring per female.

The unintended side effect of these procedures is that they truncate one tail of the data distribution for control organisms. The resulting distribution is no longer normally-distributed . More important, it violates the underlying assumption that the population of test organisms is identical for the control group and the effluent exposed groups. In addition, the standard deviation for control organisms is likely to be artificially reduced. The combined effect of these "adjustments" creates a systematic bias in the statistical analyses used to evaluate WET test data.[137] Over time, this bias significantly inflates the expected number of false toxicity reports.

A similar bias occurs when laboratories report toxicity based on EPA's Linear Interpolation Method (a.k.a. "the IC_{25} procedure"). The IC_{25} method requires data to be intentionally manipulated by "smoothing it" to meet certain statistical assumptions prior to analysis.[138] Unfortunately, EPA's recommended procedure has the effect of altering the data in a way that only tends to increase apparent control performance.

[137] *See* Swygert, Bruce, South Carolina DEHC, Letter to Glenn Stoner, Milliken & Company, re: Third Revised Draft NPDES Permit No. SC0003191 (October 11, 2001).

[138] Chronic Freshwater Manual, Appendix M.

> The ICp method can overestimate the toxicity of a sample because of smoothing of nonmonotonic data . . . The inability of the ICp to deal with hometic effects led to a 35% difference between the observed effect and the modeled effect. Because the confidence intervals for the nonparametric model are derived from bootstrapping procedures rather than from standard regression statistics, the confidence interval coverage may not be representative of the true variance either.[139]

The smoothing process never causes the mean biological endpoint in control data to be reduced. This introduces another systematic bias to the statistical analysis and tends to increase the number of false positives observed. Because the performance of control organisms is used to evaluate whether or not the performance of organisms exposed to effluent is statistically significant, exaggerating the performance of control organisms will suggest incorrectly that the effluent is more toxic than it actually is. Even EPA's contractor refers to such results as "smoothing error" when analyzing data from the interlaboratory study.[140]

Errors in the IC_{25} results are further compounded when EPA's software is used to calculate confidence limits around the estimated test endpoint. Estimated values greater than the highest tested concentration are transformed to equal the highest tested concentration during the bootstrapping process. When all of the bootstrap estimates are averaged to calculate the confidence limit, the value must be lower than the highest measured value because all values higher than that were eliminated.

The IC_{25} bias is particularly problematic for those that must demonstrate compliance with WET limits in undiluted effluent (e.g., $IC_{25} \geq 100$). Since it is impossible to test a concentration

[139] Markle, P., et al. 2000. Effects of several variables on whole effluent toxicity test performance and interpretation. *Environ. Toxicol. Chem.* 19:123-132, p. 130.

[140] WET Study Report (Section 9.3.2) at 72.

greater than 100% effluent, IC_{25} values (generated in EPA's program by a statistical procedure called bootstrapping) that are equal to or greater than that will be censored, and confidence limit estimation often fails because of excess censoring. When the IC_{25} program fails to estimate a confidence interval, it should rightly be interpreted as providing no assurance that the IC_{25} is less than 100%. Nevertheless, it often will be interpreted quite the opposite. Problems in confidence interval estimation by the IC_{25} software have led EPA to warn against using it:

> EPA recommends that confidence intervals for the ICp method not be reported or used in WET testing until the ICp software has been thoroughly reviewed by experts and possibly modified.[141]

Retraction of the confidence interval estimation leaves the IC_{25} method without an approved measure of estimate reliability. Without a tool to quantify uncertainty in the ICp, one cannot distinguish between effects caused by effluent toxicity and effects caused by random behavior of organism responses. Thus, permit holders cannot certify toxicity as required by permits.

Another source of unintentional bias occurs in the way in which some tests are terminated. For example, EPA states that the chronic *Ceriodaphnia dubia* test is expected to run from six to eight days and should be terminated when 60% of the control organisms have produced at least three broods (provided that the average is at least 15 offspring per female).

Ceriodaphnia normally take seven days to produce three broods under non-toxic conditions. On occasion, they may take a bit longer. When this occurs, the test method allows them an extra day to perform. On occasion, they finish a bit early, and the test is terminated accordingly.

[141] WET Testing Guidance at 3-2.

The problem is that the control organisms are given considerable flexibility to achieve the three-brood, 15 offspring-average. Effluent-exposed organisms are not. If control organisms finish early (e.g., six days), effluent-exposed organisms are not allowed to go the normal seven days, let alone the eight days that the species sometimes requires. The test is designed to evaluate whether reproduction is reduced, not delayed. And, even if the test were measuring delayed reproduction as an indicator of toxicity, it must first demonstrate that the observed delay is "statistically-significant."

It is not uncommon for control organisms to release the third brood overnight between the sixth and seventh day. When the technician counts offspring each morning, she will conclude that the stopping criteria has been met and terminate the test. The effluent-exposed organisms may have produced a third brood in as little as three or four hours after the control group, but they will not be given the chance to do so. Any apparent difference in performance then may be mislabeled as evidence of toxicity. In fact, EPA's method manuals instruct technicians to terminate the test and finish counting in less than two hours lest more organisms be born and confound the results.

In short, if control organisms need more time to perform, they are entitled to as much as 48 additional hours to reproduce. If effluent-exposed organisms need only an hour or two longer, they are out of luck. It is important to note that it is not unusual for effluent-exposed organisms to need slightly more time to reproduce. Moving from one culture medium to another is a shock to the system that requires a short period of reacclimation. It is not unlike the jet-lag effect humans experience when traveling across time zones. Eventually, the body adjusts, but for a while, everything is slightly off-balance.

EPA deems acceptable control organisms that delay normal reproduction by one day. Why should a similar delay be interpreted as toxicity in an effluent-exposed group? EPA's mandatory procedures consistently bias the test toward "discovering" a biometric difference even where none really exists. Unless the Agency can demonstrate that a 24-hour delay in an otherwise normal reproduction cycle is a significant threat to the environment, such temporal anomalies interfere with the data analysis and should not be misconstrued as "toxicity."

E. EPA Failed to Validate the Ruggedness of WET Test Methods

1. EPA Must Demonstrate That New Test Methods are Adequately Robust.

Standard methods are used routinely to demonstrate compliance with NPDES permit limits. As such, it is essential that the method can be performed correctly by most laboratories.

> As with the chemical method characteristics, applicability of the biological method is a key criterion for assessing the method's adequacy … For the test method to be applicable, particularly for widespread NPDES biomonitoring, biological test methods must be adaptable to a wide variety of labs. The availability of labs that can realistically perform the test methods with reproducible results should be a key criterion in determining applicability of method. The key criterion for determining applicability is the ease with which the test can be performed on a routine basis.[142]

In order to ascertain whether a method is adequately robust, EPA must define a DQO to evaluate that performance characteristic.

> A method developed for regulatory use should represent state-of-the-art technology that has been demonstrated to be practical for routine use …Once the data quality objectives of a particular need are defined, then it can be determined if a method is adequate for its intended purpose … A fully validated and standardized method is a method that has been ruggedized by a systematic process and

[142] Section 518 Report at 3-12.

is applicable for its intended use.[143]

EPA failed to establish a specific performance criterion to evaluate whether the WET test methods are sufficiently robust. In the absence of such a criterion, the Agency cannot determine whether the WET test methods can be performed, correctly, on a routine basis.

> Once DQOs have been developed and a design for the data collection activity expected to achieve these objectives has been selected, DQOs are used to define the quality assurance (QA) and quality control (QC) requirements specifically tailored to the data collection program being initiated ...Without first developing DQOs, a QA program can only be used to document the quality of obtained data, rather than to ensure that the data quality obtained will be sufficient to support a permitting decision.[144]

2. <u>The Majority of Laboratories are Unable to Complete the Specified Procedures Required in the WET Test Methods.</u>

In late 1999 and early 2000, EPA performed a large-scale interlaboratory validation study of WET test methods. In the final report for the study, EPA claimed that a very large percentage of all tests initiated were "completed successfully."[145] That statement is incorrect. EPA's definition of a successful completion was not based on whether the test was performed in accordance with the method. Instead, it stated that:

> [a] valid test was defined as a test that met the required test acceptability criteria for the method as stated in the WET method manuals. Tests that deviated from specified test conditions were identified with data qualifier flags but were not excluded as invalid tests.[146]

EPA considered test results to be valid even when produced by a laboratory that deviated

---

[143] *Id.* at 3-4 - 3-5.

[144] Permit Writer's DQO Guide at 1-4.

[145] *See* WET Study Report, Chapter 9.

[146] *Id.* at 65 (Section 9.1.1).

from the mandatory requirements in the test method protocols (other than TACs). Strict compliance with the methods, as promulgated, was a mandatory prerequisite for data acceptance, as confirmed by:

> Participant laboratories were required to analyze each blind test sample according to the promulgated WET test method manuals and specific instructions in participant laboratory standard operating procedures (SOPs) developed for the study.[147]

Table 4 clearly demonstrates that the vast majority of all laboratories participating in EPA's validation study were unable to perform the method without at least one significant deviation from the mandatory conditions in the method. To be specific, Table 4 does not include deviations from the requirements that EPA imposed in its DQOs over and above the mandatory provisions in the test methods per se.

[147] WET Study Report at xiii.

Table 4: Percent of Initiated Tests that were "Completed Successfully"

Freshwater Species	Test Protocol	Valid Tests (%)	
		Excluding Method Deviations	***Including Method Deviations***
Ceriodaphnia dubia	Chronic	82%	34%
Ceriodaphnia dubia	Acute	95%	50%
Fathead minnow	Chronic	98%	59%
Fathead minnow	Acute	100%	32%

Data from all of the other methods tested show a similar pattern; the true rate of successful test completion was less than half the number EPA claimed when deviations from method requirements are considered.

It is important to note that the actual rate at which WET method deviations occur is likely to be higher than observed during EPA's interlaboratory validation study. EPA selected "extraordinarily qualified" laboratories to participate in its Interlaboratory Study by imposing strict pre-qualification requirements (i.e., WET testing experience, proficiency, capacity, and quality control) on all candidates. The participating laboratories also were given written Standard Operating Procedures to follow, briefed at special training sessions prior to the study, and contractually required to comply with specified procedures. Therefore, the participating laboratories are atypical from (i.e., better than) the laboratory population as a whole. In other words, results from the study represent a "best case" scenario. As discussed above, even the "best" laboratories were unable to routinely perform and complete the WET tests. As underscored during the peer review process, EPA has provided no evidence that real-world performance, away from the Agency's rigorous scrutiny, is likely to be better:

> 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.[148]

If the best laboratories cannot properly perform and complete the methods, EPA clearly has not established that the methods are sufficiently rugged for use in the regulatory context. The consequences to the permittee of a laboratory failing to properly perform and complete a WET test are too great.

If a permittee failed to perform a WET test exactly as required by the mandatory provisions, as did the participating laboratories, the permittee would assume it is obligated to repeat the test. If there was insufficient time remaining to repeat the test during the monitoring period, the permittee would be exposed to liability for failure to comply with the permit provisions. Moreover, EPA apparently takes the position that permittees whose laboratories do not strictly follow the mandatory test protocols cannot certify the test results on their DMRs.

To the permittee, it makes no difference whether or not regulators are authorized to exercise discretion regarding which deviant test results they will accept or reject.[149] The permittee has no assurances whatsoever regarding how the regulator might exercise its discretion regarding any particular deviant data point. Moreover, the test protocol provisions that give permitting authorities the discretion to accept otherwise unacceptable data contain no objective criteria by which the permittee can judge the likelihood that a deviant data point will or will not

---

[148] Peer Review Report at 52.

[149] *See* Section III.F. *supra* for examples of test conditions that are mandatory (use terms like "must" or "shall"), but that the test manuals authorize the regulator to treat as discretionary after receiving the data from a permittee.

be accepted. And even if a state regulator accepts a deviation, there is no assurance that EPA will concur; absent objective criteria applicable to all, EPA might view such deviations as permit violations.

Thus, from the permittee's perspective, the test protocols (unless they are preceded by the term "may" or "should") are mandatory. If an NPDES permit contains a WET limitation, the permittee has no choice other than to strictly follow the WET test protocol. The fact that the government may later "accept" a data point from a test that deviated from the protocols is no different from the government exercising prosecutorial discretion in regard to an excursion of a permit limit. In both cases, the permittee has no basis for assuming up front that its action was lawful. While relief afterwards is possible, it is neither guaranteed nor predictable.

Likewise, any regulator that chooses to perform its own WET testing for compliance purposes will need to consider the test protocols absolutely mandatory, even where the test method provides the regulator discretion to accept data points that originated from an improperly performed test. Indeed, it is highly doubtful that a court in an enforcement action would accept government evidence (or accord it very much weight) consisting of data from improper test procedures. The court is unlikely to find persuasive an argument that the very same regulator adducing the evidence has made a *post hoc* determination that its otherwise flawed data can be accepted, absent any criteria in the test protocols for doing so.

Accordingly, EPA simply cannot ignore the finding derived from its Study that a very large percentage of even the best laboratories will be unable to perform and complete the WET tests according to the mandatory requirements. That problem obviously is not the product of sloppy laboratories seeking to cut costs. The Study demonstrates that the problem is real and

unavoidable, as evidenced by the repeated statements admonishing the participating laboratories to adhere strictly to the test protocols.

At a minimum, EPA must present, for each test method and each endpoint, the percentage of laboratories that encountered completion problems. If EPA considers acceptable a problem of that magnitude, given the manner in which the WET tests will be used, and the undeniable consequences that the problem will impose on dischargers, it must say so explicitly and explain its rationale. If the Agency concludes that the WET tests can be promulgated in Part 136 notwithstanding the problem, it must include measures in the protocol prescribing what permit writers must include to protect permittees from liability for the test completion problems their laboratories will encounter, notwithstanding their best efforts.

EPA's decision to overlook deviations from the QA/QC requirements in its test protocols is inconsistent with the statement it made in the Settlement Agreement arising from the challenge to the 1995 rule:

> EPA acknowledges that the test methods manuals . . . distinguish between requirements (by use of the compulsory terms "must" and "shall") and recommendations and guidance (by use of the discretionary terms "should" and "may") so as to indicate the instances when the analyst has flexibility to optimize successful test completion and **when standardization is necessary to assure the predictability of the methods to provide reliable results.**[150]

EPA's statement underscores that QA/QC "requirements" are just that - mandatory. Even if it were acceptable for laboratories to deviate from them "to optimize successful test completion," that clearly was not the motivation for participating laboratories to deviate from the

[150] Settlement Agreement, *Edison Electric Institute, et al. v. EPA*, No. 96-1062 and consolidated cases (D.C. Cir.) (July 24, 1998) ("Settlement Agreement"), p. 2 (emphasis added).

requirements. The record contains no information whatsoever explaining why so many laboratories deviated so frequently from the mandatory QA/QC provisions. The only valid conclusion to be drawn from the Interlaboratory Study, therefore, is that most laboratories are unable to perform the WET test methods in accordance with the procedures "**necessary to assure the predictability of the methods to provide reliable results."** Given the extensive period in which the WET test methods have been used, and the extensive experience of the participating laboratories, in particular, it will not be sufficient for EPA to justify the problem by claiming that performance will improve as laboratories become more familiar with the test.

3. Results From EPA's Interlaboratory Validation Study Demonstrate the WET Methods are not Sufficiently Robust to be Included in Part 136.

Although EPA claims that WET methods perform similarly to more traditional chemical analyses, the Agency does not provide any data to support that conclusion with respect to the rate at which test technicians deviated from mandatory methods and procedures. Even if the comparison is legitimate, it does nothing to demonstrate that the low rate of test completion is "acceptable." That's because EPA failed to establish a criterion to distinguish acceptable from unacceptable performance.

More specifically, the DQO for the interlaboratory WET validation study require a minimum of nine laboratories to evaluate the performance of the test protocols.[151]

When results from invalid tests are excluded, there is insufficient data to meet the formal DQO or to demonstrate the methods are suitable for their intended use:

> The primary objectives of the WET Study are to (1) generate data

[151] WET Study Report (Vol. 2) at A-3.

> to characterize the interlaboratory variability of the 12 WET methods targeted in the study, (2) obtain data on the rate at which participating laboratories successfully completed WET tests initiated, and (3) generate data on the rate at which WET tests indicate 'toxicity' is present when measuring non-toxic samples… Six data quality objectives (DQOs) have been identified as necessary to ensure that data produced will meet the study objectives described above. These are: (1) All data produced 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 [method manuals].[152]

EPA previously recognized that more studies were necessary to demonstrate the ruggedness of WET test methods. Indeed, as noted earlier, that was one of the primary objectives of the interlaboratory WET variability study.

> The minimum requirements for a demonstration of adequacy is that the methods have been subjected to ruggedness testing, and that single and multilaboratory precision of the methods have been established … Further standardization for the chronic toxicity test methods for freshwater and marine organisms is necessary to reduce the costs of performing the experiments and to ensure consistent data. Ruggedness, single laboratory, and multilaboratory precision studies are still needed for several test methods.[153]

It is not sufficient that EPA merely complete the studies; the results also must demonstrate that the methods are, in fact, robust. EPA's official (and anonymous) peer reviewers concluded that the WET methods failed to make that demonstration:

> … 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 CV values are too high.[154]

If the procedural error were the fault of the laboratories participating in the study, then

[152] WET Study Report (Vol. 2) at A-6 - A-7.

[153] Section 518 Report at 4-49 and 4-52.

[154] Peer Review Report at 19.

the results indicate that there are very few laboratories capable of performing the methods correctly. Since WET testing is mandated in approximately 6,500 permits[155] and, since strict conformance with the methods is also mandatory for NPDES permit testing,[156] EPA's data clearly demonstrates a lack of adequate testing capacity at qualified laboratories.

If the laboratories are assumed to be highly competent, as EPA did when the laboratories were pre-qualified for participation in the study, then the only reasonable conclusion is that the methods themselves are not adequately robust for routine use. In either case, EPA failed to demonstrate that the WET test methods met the performance criteria for ruggedness and applicability.

Additional issues and concerns related to using invalid data from EPA's WET interlaboratory variability study are set forth in Section V to follow.

4. The *Ceriodaphnia* Test Exemplifies the Completion Problem.

EPA acknowledges that one-third of laboratories participating in the chronic *Ceriodaphnia* test series were unable to complete the test "successfully." That is, the laboratories were unable to meet the minimum Test Acceptance Criteria for control survival and reproduction, and thus the tests were deemed invalid. EPA implies the results were anomalous and concentrated among eight labs with particularly poor performance due to "poor culture health." EPA offers no evidence to substantiate the claim that stock cultures were in poor health at the laboratories in question.

---

[155] Shukla, R., et al. 2000. Bioequivalence approach for whole effluent toxicity testing. *Environ. Toxicol. Chem.* 19:169-74.

[156] *See* Section IV-B and Section V.

An alternative explanation is that the remaining two-thirds of the laboratories were able to successfully complete their tests because they performed the tests differently than those that had completion problems. Perhaps either of the two groups of laboratories deviated from the mandatory procedures in the chronic test protocol.[157] EPA warns that failure to conform to the required methods also increases test variability.

> A number of problems with WET tests are caused by misapplication of the tests, misinterpretation of the data, lack of competence of the laboratories conducting WET testing, poor condition/health of test organisms, and lack of training of laboratory personnel, regulators and permittees … The critical steps in minimizing WET test method variability are … conducting tests properly to generate the biological endpoints.[158]

Even if poor culture health was the primary barrier to successful test completion, that leaves open two very substantial issues. First, EPA does not explain why the "poor culture health" it predicts to have existed in one third of the participating laboratories from around the country is not likely to be present in at least that large a proportion of the laboratories that would be using *Ceriodaphnia* if it is ratified in Part 136. It is not sufficient to state, without any technical support, that "completion rates for this method improve when testing laboratories are allowed flexibility in the timing of sample collection and can avoid initiating tests during periods of marginal to poor culture health."[159] Laboratories performing WET testing on behalf of permittees will be expected to analyze those results when the samples arrive; they will have no more "flexibility" in timing than the laboratories that participated in the Interlaboratory Study.

---

[157] Citation back to Table of deviations in Section II-E (Ruggedness).

[158] WET Variability Guidance at 4-1 and 5-1.

[159] 66 Fed. Reg. at 49,806 (col. 1).

Moreover, EPA assumes that the laboratories can make a reliable determination when their brood cultures are impaired and when they are recovered. It offers no evidence to support that assumption, or criteria that laboratories might use to assess the health of their brood cultures.

The second reason that the Agency's "poor culture health" explanation raises concerns is that EPA fails to explain how the 8 laboratories that failed half of the tests they initiated were able to meet the Test Acceptance Criteria in the remaining 50% of tests. The results demonstrate that even when EPA believes a laboratory has impaired brood stock, the permittee only a 50:50 chance of identifying that deficiency in a single test.

Independent research studies, using EPA databases, indicate that poor test completion is only marginally related to the choice of laboratory. Only 2% of the variance in chronic *Ceriodaphnia* reproduction results, when evaluating identical reference toxicant samples, is correlated with the specific laboratory performing the test.[160] The date the test was performed had 15-times more influence on the final results. This suggests successful test completion has a significant random component. If so, wastewater samples would also appear to vary in estimated toxicity due to random influences on control performance rather than changes in actual effluent quality.

Culture "crashes" are actually quite common. Most laboratories experience at least one or two of these events per year. Although there appears to be a semi-regular pattern to the culture crashes, it is not necessarily seasonal or predictable. EPA's suggestion that the phenomena is isolated to only a few laboratories is unfounded. The Agency's failure to investigate completion rates over longer time periods obscures the true level of difficulty

[160] *See* WERF Variance Report.

experienced by all laboratories at one point or another.

In short, the Interlaboratory Study confirms that laboratories using the *Ceriodaphnia* test are likely to experience very significant completion problems. The Agency does not offer a suitable explanation for that result. Moreover, this is not the first time completion problems have arisen with *Ceriodaphnia*. The 1994 Chronic Manual (at page 193) itself states that laboratories in an earlier studied failed to complete 48 of the 91 tests that had been initiated (i.e., over 50 percent), (adjusted to reflect the 1994 protocols). The completion problem, therefore, is not anomalous, as EPA suggests. It is real, and it must be addressed before EPA can conclude that *Ceriodaphnia* is a sufficiently rugged test to be ratified in Part 136. At a minimum, if EPA believes it can support a ratification decision, the Agency must include explicit language in the test protocol itself stating that the Part 136 method is not intended to be used to support mandatory NPDES monitoring requirements, unless the permit specifies that dischargers will not be legally responsible for any deviations from monitoring requirements attributable to failure to complete the test (i.e., failure to achieve the TAC when all mandatory test protocols are followed).

F. EPA Failed to Establish Clear and Correct Reporting Requirements for WET Methods.

1. EPA Failed to Follow Agency Guidance Recommending That Analytical Variability Must be Accounted for When Reporting Test Results.

> The precision of toxicity measurements is similar to that of finely tuned instruments operating at detection limits. Thus users of biological methods must account for the inherent variability in response. Typically for toxicity test methods, this means using replicate exposures at each concentration and running parallel tests with each sample or batch of test organisms using a standardized toxicant so that the 'health' or sensitivity of the test organisms can be independently measured. It also means that the natural variability in sensitivity will have to be accounted for. More

> importantly, this variability must also be accounted for when permit limits, criteria or standards are set.[161]

Although EPA has resisted adjusting water quality criteria or permit limits for analytical variability, they have allowed such corrections when using data to evaluate compliance with chemical-specific limits.

> Although analytical precision can and does affect variability, it can be quantified and taken into consideration when reporting data for NPDES permits. Usually, the methods used for water and wastewater analysis have precision and accuracy factors reported.[162]

As noted earlier, the most common method of accounting for analytical variability in a traditional chemical test is to report results differently depending on whether data fall above or below the detection or quantitation level. However, since no such levels have been established for WET methods, EPA failed to include the necessary procedure to adjust WET test results for analytical variability before reporting or certifying compliance on a DMR.

EPA came close to providing such a procedure in the Technical Support Document for Water Quality-based Toxics Control:

> The allowable frequency for criteria excursions should refer to true excursions of the criteria, not to spurious excursions caused by analytical variability or error. In evaluating data on chemical concentrations or toxicity units, it is desirable to subtract the analytical error log variance from the observed log variance in order to arrive at the true log variance contributing to criteria excursions.[163]

Unfortunately, EPA failed to provide a specific formula for calculating Error Log

[161] Section 518 Report at 3-11.

[162] Permit Writer's DQO Guide at 1-7.

[163] TSD Responsiveness Summary at 11.

Variance or Observed Log Variance. Hence, this particular adjustment procedure has no practical utility and is not used by state permitting authorities.

The need to account for analytical variability is particularly important when assessing historical effluent quality data to determine whether there is "reasonable potential" to violate a water quality standard. The results of such assessments often determine whether a permit limit is imposed.

It is well understood that any test that relies on statistical analysis, as WET methods do, will incur a predictable number of false failures over the course of a large number of samples (see discussion in Section II.A.3. above). EPA failed to warn WET method users that this will occur and failed to provide an equation for calculating the number of expected errors in a given number of tests.

EPA's current guidance instructs states to conclude that reasonable potential exists if there is even a single exceedence during the previous five-year monitoring period.[164] However, EPA fails to appropriately take into account the cumulative number of false failures that inevitably will occur. Consequently, the vast majority of all dischargers performing routine WET testing will appear to have reasonable potential to exceed water quality standards regardless of their actual effluent quality (assuming a state has promulgated a legitimate standard for WET).

A similar reporting problem arises when dischargers elect to analyze identical split aliquots of a single effluent sample. If the two labs disagree as to whether the sample is toxic or

[164] *See* TSD, Chapter 3.

not, and assuming both tests are valid, permittees are uncertain as to how to certify compliance status on the DMR.

All states require both test results to be reported. And most states require dischargers to declare a permit violation based on the failed test. Such an approach lacks legal or scientific merit. Both the 'presumption of innocence' and the 'null hypothesis' cause us to assume there is no toxicity until there is strong evidence to the contrary. Inconsistent results between identical split samples is not sufficient evidence that toxicity is present or that a violation occurred. It is not unreasonable to require additional testing in such circumstances, but it is unreasonable to require permittees to certify non-compliance on the basis of such inconclusive data. This is especially true given that the courts generally have prohibited dischargers from utilizing a defense based on analytical variability after test results are certified.

This reporting problem is particularly troublesome, given EPA's recent finding that the vast majority of laboratories make errors in the way they analyze and report WET test data. Indeed, it turns out that:

> Laboratory data submitted during EPA's interlaboratory variability study of whole effluent toxicity test methods were reviewed to evaluate the accuracy with which laboratories routinely analyzed and reported WET test data. It was found that 74% of the 35 laboratories reporting results for the Ceriodaphnia dubia chronic test and 82% of the 28 laboratories reporting results for the fathead minnow chronic test made one or more errors in the calculation or reporting of 54% of the 126 Ceriodaphnia dubia chronic tests results and 63% of the 105 fathead minnow chronic test results … Only 17% and 19% of errors resulted in a difference greater than 10% in reported results for the Ceriodaphnia dubia and fathead minnow chronic methods, respectively. These results indicate a need for increased laboratory quality control and personnel training, increased emphasis on state and regional laboratory certification programs, and increased client attention to laboratory

selection and result reporting.[165]

2. EPA Failed to Publish Complete Guidelines for Many of the Procedures Used to Account for Variability or Control for Test Interferences.

In various supplemental guidance documents, EPA has recommended: (a) using dual controls, (b) reporting confidence ranges, (c) confirming dose-response relationships, (d) rejecting outliers, and (e) tracking long-term trends in laboratory performance. In every instance, EPA provides little or no guidance on how to implement such recommendations.

a) Dual Controls.

EPA suggests using dual controls to account for ionic interference in receiving waters.[166] Effluent results would be compared, simultaneously, to control organism performance in standard laboratory dilution water and in upstream receiving waters. However, EPA does not explain what conclusion should be drawn or how to report compliance on a DMR when the results of these two comparisons are inconsistent.

b) Reporting confidence ranges

When the WET methods were promulgated in 1995, EPA stated a preference for using point estimates rather than NOECs to evaluate toxicity.

> In the NPDES permitting program, the recommended statistical procedure is the point estimate because confidence intervals can be placed around the point estimate.[167]

Unfortunately, EPA's recommended approach for calculating point estimates often fails

[165] Brent, R., et al. Accuracy of Laboratory Reporting in EPA's WET Interlaboratory Variability Study. Abstract for 2001 SETAC Conference in Baltimore, MD.

[166] *See* WET Testing Guidance, Section 6-5.

[167] 60 Fed. Reg. at 53,539.

to generate appropriate confidence intervals. This problem most frequently occurs when the intervals would tend to show that the point estimate for toxicity cannot be distinguished from the non-toxic condition with the necessary statistical confidence. EPA encourages states to continue using the point estimate procedures, rather than refraining from using those procedures until the problems can be corrected. That recommendation lacks scientific foundation and undermines the validity of all results reported using the handicapped procedure.

c) Confirming dose-response relationships

EPA recently published new guidance explaining how to evaluate WET data to determine whether a concentration-response relationship (a.k.a. "dose-response relationship) exists.[168] This guidance is consistent with EPA's previous instructions on the subject:

> A predictable dose response curve is one of the mandatory requirements for a valid toxicity test. We would never accept analytical results from an instrument producing an abnormal standard curve. The predictable dose response curve, that is increasing toxicity with increasing concentration, is the analogue of the analytical standard curve and is of equal importance in toxicity testing.[169]
>
> The agency [EPA] is concerned that single concentration, pass/fail, toxicity tests do not provide sufficient concentration-response information on effluent toxicity to determine compliance. It is the Agency's policy that all effluent toxicity tests include a minimum of five effluent concentrations and a control.[170]
>
> The dose response curve is the basis for the validity of the toxicity test. The control serves as the starting point from which the dose response is evaluated. If a dose response is not obtained, the

[168] *See* WET Testing Guidance, Chapter 4.

[169] Mount, Donald, National Effluent Toxicity Assessment Center, U.S. EPA Environmental Research Laboratory - Duluth, NETACommunique re: Number of Test Concentrations Needed (January 1990).

[170] WET SID at 28.

> toxicity cannot be inferred.[171]

EPA proposes to require that:

> . . . the concentration-response relationship generated for each multi-concentration test must be reviewed to ensure the calculated test results are interpreted appropriately.[172]

"Reviewing" the concentration-response is critical, but it is not enough. It is necessary to confirm the presence or absence of a valid concentration-response relationship before concluding a sample is toxic. It is surprising and arbitrary that EPA neglected to make this a mandatory element in the method. This is especially true given that EPA relied on dose-response analysis to reject several tests that otherwise would have been falsely labeled toxic during the interlaboratory WET variability study.[173]

In addition, although EPA provides some illustrations as to how to interpret dose-response relationships, the Agency failed to develop or publish an objective statistical test to confirm the presence or absence of a valid dose-response relationship.

> The use of NOECs … assumes either (1) a continuous does-response relationship, or (2) a non-continuous (threshold) model of the dose-response relationship … The data should be plotted, both as a preliminary step to help detect problems and unsuspected trends or patterns in the responses, and as an aid in interpretation of the results.[174]

Although EPA recommends graphing the data, it does not specify how to define the

---

[171] Norberg-King, Teresa J., U.S. EPA Environmental Research Laboratory - Duluth, Memorandum to Rob Pederson, EPA Region X, *Review of the Toxicity Results from West Boise and Landers Street POTWs* (June 5, 1989).

[172] Proposed Method Manuals Changes at 64.

[173] WET Study Report (Section 8.2.3) at 62.

[174] Chronic Freshwater Manual (Section 9.1.1.2) at 44.

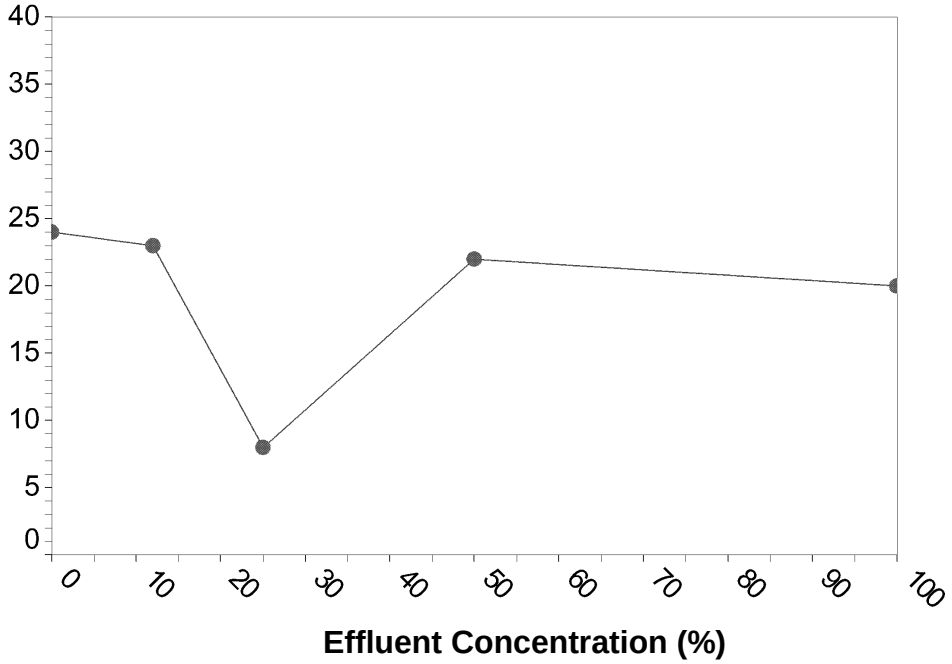
difference between a valid dose-response and an invalid or inconclusive concentration response. The inherent subjectivity associated with visually analyzing graphed results makes the technique unsuitable for its intended purpose as a tool to determine compliance.

A similar issue arises with respect to the definition of an NOEC vs. the Lowest-Observed-Effect-Concentration (LOEC). Occasionally, toxicity test data exhibit an anomalous pattern like that shown in Figure 3.

Some state regulators would conclude that the NOEC occurred at the 25% effluent concentration. Others would note that there was no statistically-significant effect at the 100% concentration and label it the NOEC. Indeed, a WERF study demonstrated that there frequently is disagreement between state regulators over interpreting WET test results even when examining identical data sets.[175]

[175] Water Environment Research Foundation, *WET Testing Program: Evaluation of Practices and Implementation,* Report #D83001 (1998).

Figure 3: Anomalies in Dose-Response Analysis



The problem is that EPA's guidance offers inconsistent definitions of an NOEC:

> The NOEC [no observed effect concentration] is the highest concentration of toxicant, in terms of percent effluent, to which the test organisms are exposed that causes no observable adverse effect.[176]

vs.

> If in the calculation of a NOEC, two tested concentrations cause statistically adverse effects, but an intermediate concentration did not cause statistically-significant effects, the test should be repeated or the lowest concentration must be used. For example: 6.25, 12.5, 25, 50 and 100% effluent concentrations are tested. The 12.5 and 50% concentrations are statistically-significant, but 25% is not statistically significant. If the test is not repeated, then the NOEC is 6.25%.[177]

[176] TSD at 4.

[177] U.S. Environmental Protection Agency, *EPA Region IX Whole Effluent Toxicity Training Course Manual* (1999), pg. 3.

Rigorous dose-response analysis would likely eliminate such anomalies.[178] However, EPA's failure to require such analysis and provide clear decision rules perpetuates regulatory inconsistency and renders the methods too confusing and inadequate for their intended purpose.

d) Rejecting outliers

While reviewing data collected during the WET interlaboratory validation study, EPA's contractor rejected results deemed to be "outliers."[179] This is a common practice when validating chemical methods. However, it assumes that the true concentration of the tested analyte is known. Without such information, it is not possible to discern which toxicity results are representative and which are outliers.

The practice of rejecting outliers simply because the WET test results were markedly different from other laboratory results is inconsistent with any permit requirement to report the results of all valid tests of effluent quality when evaluating compliance. In addition, discarding such results causes EPA to significantly underestimate variability and overestimate precision of toxicity test methods.

Even if one assumes that it is legitimate to reject outlier data, the procedures used by EPA's contractor cannot be applied in an equivalent manner when reporting individual WET tests on a DMR. The ability to identify outliers depends on analyzing a large number of identical samples. This is a practical impossibility for permittees. Consequently, while outliers are known to occur, EPA failed to include a reasonable procedure in the method to identify and reject such outliers prior to reporting compliance.

---

[178] *Id*. at Chapter-Statistics, p. 1.

[179] WET Study Report (Table 9.1) at 60.

If EPA had not used its subjective dose-response judgments and special outlier criteria, its estimate of false WET test failures would likely have been considerably higher than the Agency actually reported.[180]

e) Tracking long-term trends

Finally, EPA recommends that laboratories track test performance over time.[181] Specifically, the Agency suggests that laboratories record the average response and variability for laboratory dilution water (non-toxic culture water) and for reference toxicant conditions. These data can be used to assess the general health of test organisms.

> The health of the test organisms or biological system being monitored is a unique attribute without a complimentary attribute for analytical methods. The health of a toxicity test organism has a profound effect on the quality of the data, and must therefore be considered a key criterion for assessing adequacy … The health of test organisms and biological systems cannot be 'calibrated' before the experiment in the same way as analytical instrumentation ... There are no knobs to turn to adjust for these factors to achieve consistent performance during a test method. For these reasons, the biological procedure must include biological standards (e.g., standard reference toxicants) in order to ensure data integrity.[182]

Unfortunately, while EPA recommends that such data be collected, it does not require it. And the Agency fails to establish any national QA/QC criteria by which to evaluate laboratory performance. Even the labs that voluntarily track such information measure their performance only against their own historical results. There is no independent national standard by which to compare the performance of one lab against another.

---

[180] *See* WET Study Report.

[181] WET Variability Guidance at 7-3.

[182] Section 518 Report at 3-11.

Ironically, then, the more variable a given laboratory's performance over time, the less likely that any single reference toxicant test will fall outside its 'normal' range. The same holds true for variability in stock cultures used to supply test organisms. Thus, even if WET test methods may exhibit variability similar to chemical test methods in an interlaboratory study, the difference in standardized QA/QC procedures causes the variability to go largely unchecked when toxicity testing is performed in practice.

Incomplete test requirements undermine the utility of WET methods. Even relatively simple issues, such as whether to count or ignore the offspring produced in a fourth brood during the *Ceriodaphnia* test or whether to include the weight of dead minnows in the chronic growth test, create great uncertainty among laboratories and increase the imprecision of split-sample tests. EPA has recommended many procedures to reduce inter-test variability but failed to require any of them.[183] If test precision can only be maintained in a tolerable range by adjusting the standard procedures as Kentucky, Tennessee, Washington, North Carolina, South Carolina, Wisconsin, and Texas have done, then the methods are incomplete until such refinements are made mandatory in Part 136.

In fact, the whole issue of what elements are mandatory and what elements are discretionary also must be clarified. Dischargers are required to certify that they performed the WET test in accordance with the required methods. There is considerable uncertainty as to whether a test remains valid even when it fails to comply with one or more of the experimental conditions set forth in the procedure manuals. Moreover, allowing test conditions to change from lab to lab allows factors other than effluent quality to effect the outcome of a test. EPA

[183] *See* WET Variability Guidance, Appendices E and F.

must revise the methods to indicate clearly where discretion is allowed to deviate from specified procedures and where no deviations will be tolerated. Until then, the "method" may be what any given lab says it is on any given day. This cannot be the basis for adopting a standard procedure into Part 136.

EPA's general conclusion that WET testing is suitable for use in NPDES permitting is misleading. The more relevant issue is which specific tests, which specific endpoints, which specific statistics, and which specific laboratory conditions using which specific QA/QC criteria make the methods suitable for which specific permitting purpose (general water quality characterization, process control monitoring, reasonable potential determination, toxicity identification evaluations, or compliance and enforcement)?

Each and every WET test endpoint must be shown to be suitable for each and every purpose for which it is authorized to be used. EPA has not yet met that prerequisite.

G. EPA Did Not Validate the Applicability and Comparability of WET Test Methods.

The need for EPA to confirm that WET test methods can reliably measure instream toxicity potential was underscored by an EPA Administrative Law Judge in an opinion stating:

> [t]here must be a reasonable basis to believe the permittee discharge could be or become acutely toxic. In addition, the proposed tests must be reasonably related to determining whether the discharge could lead to real world effects. The CWA objective to prohibit the discharge of 'toxic pollutants in toxic amounts' concerns toxicity in the receiving waters of the United States, not the laboratory tank.[184]

[184] *In the Matter of Metropolitan Dade County, Miami-Dade Water and Sewer Authority* (NPDES Permit No. FL0224805), EPA Administrative Law Judge Division, 1996 EPA ALJ LEXIS 80 (October 3, 1996) at 20.

As discussed below, EPA has not yet confirmed that its proposed Part 136 methods are capable of predicting "real world effects."

1. EPA Must Demonstrate the Representativeness and Comparability of Part 136 Methods.

EPA has specified how WET test methods are supposed to be used in the regulatory process, but it has never established that the WET methods it proposes under Part 136 are suitable to support those uses. For example, EPA has stated that WET effluent limitations must be imposed where WET testing shows that an effluent has the reasonable potential to cause or contribute to an excursion of a narrative criterion within a state's water quality standards.[185] So EPA concludes that WET methods are capable of identifying "excursions" of narrative water quality criteria. Indeed, in its test manuals, EPA states:

> [t]he objective of aquatic toxicity tests with effluents or pure compounds is to estimate the "safe" or "no effect" concentration of these substances, which is defined as the concentration which will permit normal propagation of fish and other aquatic life in the receiving waters.[186]

That quote confirms EPA's position that its WET methods can be used to identify excursions of the particular narrative criterion that many states express in terms of "propagation" of aquatic life. EPA states that WET methods also can be used "to interpret their narrative requirements of 'no toxics in toxic amounts.'"[187] Some states define toxic amounts as any amount of pollutant that significantly impairs the richness, abundance, or community structure of

[185] 40 C.F.R. § 122.44(d)(1)(v) (2001).

[186] Chronic Freshwater Manual at 3. WET tests are used to "identify effluents and receiving waters containing toxic materials in chronically toxic concentrations… the data are used for NPDES permits development and to determine compliance with permit toxicity limits."

[187] 63 Fed. Reg. 36, 742, 36,768 (July 7, 1998).

aquatic organisms.

Moreover, EPA has clarified in the proposed rule that:

> EPA's promulgation of WET test procedures are not water quality criteria recommendations under section 304(a). When States develop and implement water quality standards, including narrative water quality criteria, States should translate those criteria into measurable expressions of toxicity. The test methods themselves are not per se translators of the narrative criterion: "no toxics in toxic amounts." The test methods are merely the measurement tools according to which such criteria may be translated.[188]

In summary, EPA is clear about how WET test methods are supposed to be used in the regulatory process – "tools" for translating narrative criteria into numerical WET values (e.g., toxicity units or "TUs") from which WET effluent limitations can be derived (and for making reasonable potential determinations). If EPA wishes the particular WET methods it proposes to ratify in Part 136 to be used for that purpose, it must confirm that those WET test methods (each of them and their endpoints) can support that intended use. It is not enough just to evaluate whether or not the WET methods can reliably measure whether or not a numeric WET effluent limitation has been exceeded. That level of validation may be sufficient for chemical-specific test methods. Unlike WET methods, chemical-specific methods are not expected to be used as "tools" to translate the water quality criteria; they are expected only to be used to gauge compliance with the effluent limitations derived from numeric water quality standards.

As EPA previously told Congress:

> Where possible, and in all cases for methods that will have extensive regulatory use, a method should be fully validated and standardized. This increased level of validation verifies that the

[188] 66 Fed. Reg. at 49,796 (emphasis added).

method is suitable for its intended purpose.[189]

As discussed below, EPA has not established that its proposed WET methods are suitable for their intended uses.

2. EPA Did Not Validate all WET Test Endpoints in Common Use.

EPA claims that WET tests provide reliable evidence of instream conditions.[190] The Agency bases its claim on a number of studies that sought to confirm a correlation between WET test results and instream biology. However, these studies only examined a few of the biological endpoints used in toxicity testing.

Several WET test endpoints have never been validated at all. These include nearly all of the marine test endpoints and the *Selenastrum capricornutum* cell density endpoint. EPA has substantially altered many of the other test methods since the time their field validation studies were done.[191] The Agency has made no attempt to re-validate the correlation between WET test results and actual instream conditions using the newly modified methods. This is particularly problematic where the new methods show toxicity where none previously existed.[192]

> Protocol changes and options contained within U.S. EPA whole effluent toxicity tests represent variables that have the potential to affect bioassay performance and interpretation of results.[193]

---

[189] Section 518 Report at 3-6.

[190] 60 Fed. Reg. at 53,529.

[191] For example, changing the age requirement in the fathead minnow test to use larval rather than juvenile fish.

[192] A more detailed discussion of the specific method modifications is presented in Section IV.

[193] Markle, et al. (2000), p. 123.

Many of the statistical endpoints now in common use have not been shown to be accurate predictors of actual instream conditions. In particular, most "instream effects" studies cited by EPA defined toxicity based on the LC_{50} rather than the NOAEC. However, the NOAEC now is being used frequently to determine compliance with NPDES permit limits.

Similarly, EPA now recommends that point estimates (such as the IC_{25}) be used to assess effluent toxicity. However, the relationship of the IC_{25} endpoint to instream conditions was not evaluated in any of the studies the Agency cites to demonstrate validity of the method. In fact, the Agency has not yet shown that a 25% reduction in growth or reproduction, as measured in a laboratory-based toxicity test, has any significant long-term effect on aquatic populations in the field. To do so would require EPA to account for, among other things, the frequency, duration, and magnitude of the instream exposure, which will vary considerably from one waterbody to the next (and even within a single waterbody). EPA acknowledged in the Settlement Agreement that its Part 136 rule "does not specify means to adjust [WET test methods] for the frequency, duration, or magnitude of instream exposure conditions. . . ."[194] EPA states that regulators need to make those "adjustments" in setting water quality standards and making permitting decisions.[195] But regulators will be unable to do so unless EPA confirms in the Part 136 process that WET methods can be "adjusted" in a manner that assures their reliability in measuring actual instream impacts.

Rather than demonstrate the validity of the method for its intended purpose, as required by 40 C.F.R. Part 136, EPA inappropriately deferred its responsibility. Further, the Agency

[194] Settlement Agreement at 2.

[195] *Id*.

neglected to acknowledge this critical limitation and prerequisite requirements in the method itself.

> The interpretation and application of [toxicity] test results are part of the implementation policy and are not addressed in this rulemaking ... It is not always obvious that an effect level that is determined to be statistically significant is also biologically significant. The implied question, concerning the biological significance of (threshold) statistically significant occurrences of adverse biological effects observed in toxicity tests, is an implementation question, and is not addressed in this rulemaking.[196]

On a related note, EPA continues to claim that the results from a single WET test are suitable for assessing effluent compliance (and water quality standards compliance).[197] However, the Agency has never demonstrated the validity of this assertion particularly when the failures occur infrequently and inconsistently in an otherwise dominant pattern of passing tests. Nor has the Agency justified the validity of single tests when two identical split samples report inconsistent results regarding toxicity.

3. EPA's Field Validation Studies do not Demonstrate Comparability of WET Methods.

EPA cites a number of studies to support its claim that WET tests provide a reliable indicator of instream conditions.[198] In the time since EPA initially relied on these studies, the research has been criticized extensively in peer-reviewed scientific literature.[199] [200] [201] EPA has

---

[196] WET SID at 28 and 33.

[197] *See* U.S. Environmental Protection Agency, *Whole Effluent Toxicity (WET) Control Policy,* EPA 833-B-94-002 (July 1994).

[198] *See* TSD.

[199] Parkhurst, B.R., W. Warren-Hicks, and L.E. Noel. 1992. Performance characteristics of effluent toxicity tests: Summarization and evaluation of data. *Environ. Toxicol. and Chem.*

not responded to these criticisms, nor has the Agency submitted its own studies to formal scientific peer review as now required by Agency procedures.

Among the most salient criticisms is the fact that EPA's studies did not attempt to validate all of the endpoints commonly used in WET testing. For example, if inhibition of reproduction or growth, by itself, can be defined as an excursion of a narrative criterion, then these "sublethal" endpoints must be evaluated independently of other endpoints. In EPA's studies, the sublethal impacts were frequently confounded by (and caused by) acute mortality effects. Thus, while sublethal effects may be reliable predictors of instream impairment when accompanied by severe reductions in survival, there is no evidence to indicate that the sublethal effects are predictive when there is no increase in mortality.[202]

Another significant criticism of EPA's studies is that the research sites were not randomly selected. For example:

> Because study design and analysis were site specific in most of these studies, and because site selection was nonrandom in most instances, past studies were unable to establish predictive relationships that could be applied elsewhere.

In fact, most of the sites were chosen specifically because they were known to be impaired and the water quality was known to be of concern. Control sites were also selected

---

11:771-791.

[200] Marcus, M.D., and L.L. McDonald. 1992. Evaluating the statistical bases for relating receiving water impacts to effluent and ambient toxicities. *Environ. Toxicol. and Chem.* 11: 1389-1402.

[201] Chapman, P.M. 1995. Extrapolating laboratory toxicity results to the field. *Environ. Toxicol. and Chem.* 14: 927-930.

[202] Diamond, J. and C. Daley. 2000. What is the relationship between whole effluent toxicity and instream biological condition? *Environ. Toxicol. Chem.* 19:158-168.

non-randomly to assure that only the "best" locations served as a reference baseline. In some cases, locations downstream of a discharge were selected as control sites because they showed the highest level of species richness and abundance. This selection bias caused EPA to overestimate the correlation between toxicity test results and instream condition. It also caused the Agency to underestimate the occurrence of false positives.[203] In short,

> The U.S. EPA studies have been criticized for selecting sites with high instream toxicity and known biological impact. Further, none of these studies demonstrated predictive accuracy.[204]

EPA acknowledges that toxicity testing produces false positives regarding stream impact. According to EPA's Technical Support Document for Water Quality-based Toxics Control ("TSD") the number of false positives ranged from 7% to 9.4%. EPA defines a false positive as an occasion when toxicity testing predicts adverse instream impacts but no such impacts were observed. Although it appears that false positives were only 9.4% of the total sites examined (in the pie chart below), two-thirds of all unimpaired sites (i.e., 9.4% divided by the sum of the 9.4% that were unimpaired but falsely deemed by WET tests to be impaired and the 4.4% that were unimpaired and deemed by WET tests to be unimpaired) were incorrectly predicted to be impaired by failed toxicity tests. The errors were a small percentage of all sites examined but a large percentage of unimpaired sites. In other words, even if the TSD information were to support that WET tests can confirm sites that are in fact impaired (setting aside the concerns mentioned earlier), the TSD information shows that WET tests routinely will find unimpaired

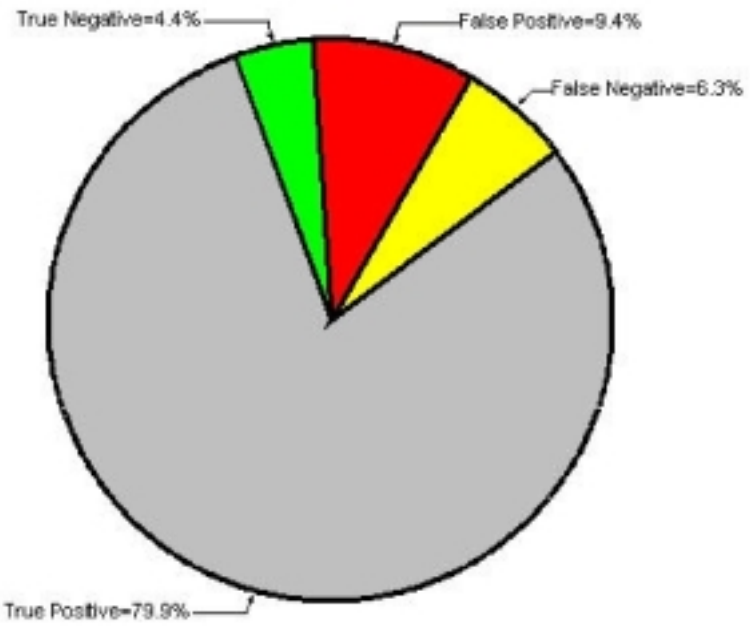
---

[203] Parkhurst, B.R. 1996. Predicting Receiving System Impacts from Effluent Toxicity, in Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts. D. R. Groethe, et al. (eds.). SETAC Press, Pensacola, FL, pp. 309-321.

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

sites to be impaired.

Figure : Summary of WET Decision Errors as it Appeared in EPA's TSD (1991)[205]



EPA underestimated the true rate at which the WET methods incorrectly predict instream impairment. Most of the sites examined in EPA's Complex Effluent Toxicity Testing Program ("CETTP") studies (85%), in fact were impacted and, hence, could only produce a true positive, never a false positive, result. Ceriodaphnia tests also indicated toxicity at more than half of the 53 sites in which no instream impacts were found. This tendency is overlooked if the frequency of presumed false positives is expressed as a percentage of all sites examined rather than as a percentage of unimpacted sites. Reanalysis of EPA's data, adjusting for the unbalanced sampling protocol used by the Agency, shows that the true rate of false positives is much higher:[206]

[205] Hall, Jr., L.W. and J.M. Giddings. 2000. The need for multiple lines of evidence for predicting site-specific ecological effects. *Hum. Ecol. Risk Assess.* 6:679-710.

[206] Novartis Crop Protection. *An Ecological Risk Assessment of Diazinon in the*

Table 4. Frequency of positive *Ceriodaphnia* ambient toxicity tests associated with sites at which no instream impact was observed. U. S. EPA Complex Effluent Toxicity Testing Project

Location	Number of sites examined	Non-impacted sites *Ceriodaphnia* tests at non-impacted sites	Positive *Ceriodaphnia* tests at non-impacted sites	% false positives	Reference
North Carolina	43	12	3	25	Eagleson et al., 1990
Elkhorn Creek, KY	160	22	15	68	Birge et al., 1989; Dickson et al., 1992
Trinity River, TX	72	10	6	60	Dickson et al., 1992
CETTP[1] (8 studies)	80	9	6	67	Dickson et al., 1992
TOTAL	355	53	30	57	

Moreover, all of the instream studies EPA relied on to demonstrate the validity of WET testing, in fact, were performed on ambient water samples drawn directly from the stream. These were not whole effluent samples. By definition, such stream samples took into full account the actual dilution available. By contrast, most WET tests performed to demonstrate compliance with NPDES permit limits make very conservative (7Q10) dilution assumptions. EPA acknowledges that incorrect dilution assumptions undermine the validity of the test, but has not made more accurate assumptions a mandatory part of the method.

> Therefore, a discharger's chance of being charged incorrectly with causing instream toxicity is low <u>if and only if</u> dilution in the receiving water is considered.[207] (emphasis in original)

EPA's field validation studies also failed to account for other significant factors, such as

---

Sacramento and San Joaquin River Basins (November 1997).

[207] TSD at 8.

habitat or natural local water chemistry, that are known to affect species richness and abundance. Nor was any attempt made to demonstrate that the surrogate test species were, in fact, representative of resident species in each waterbody.

> Whole effluent toxicity test species are generally not the same as resident species that the results of WET testing are aimed at protecting … Differences exist between sensitivities and tolerances of WET species. Such differences are not unexpected … however, these differences can become profound when regulatory use of WET test results involves a bright line that does not adequately account for species differences.[208]

The most important consideration is this: even if EPA's field validation studies demonstrate that there is a predictive relationship between acute toxicity endpoints when measured using the LC-50, that does not validate the existence of a relationship between chronic sublethal endpoints and instream condition. EPA cannot generalize the validity of some WET tests to all WET tests, as underscored by the following statement:

> General validation consists of testing, evaluating and characterizing the method to the extent necessary to demonstrate that the method achieves a specified performance. This process establishes quantitative measures of performance under typical conditions of use … A method that is generally validated cannot unequivocally be assumed to be valid for every specific use.[209]

[208] Chapman (2000).

[209] Section 518 Report at 3-6.

4. Independent Scientific Studies are Unable to Demonstrate a Correlation Between WET Test Results and Actual Biological Conditions in the Receiving Waters.

Recognizing the inherent deficiencies in EPA's field validation studies, other researchers have attempted to determine whether WET test results reliably predict instream conditions using superior data and study designs. These studies do not confirm EPA's conclusions:

> Although WET tests are useful in predicting aquatic individual responses, they are not meant to directly measure natural population or community responses … Consequently, a more direct evaluation of ecosystem health, using bioassessment techniques may be needed to properly evaluate aquatic systems affected by wastewater discharges.[210]

> There is nearly a 50% probability that toxicity exhibited in WET tests may not be reflected instream, even for those effluents exhibiting a relatively high failure rate (>90%) … A surprising result of this study was the lack of relationship between Ceriodaphnia acute or chronic WET endpoints and instream biological results.[211]

It is not surprising that WET tests are unreliable predictors of actual instream conditions given the degree of inaccuracy and imprecision for the methods, as noted elsewhere in these comments, and below:

> Comparisons to field conditions indicate that WET tests are not reliable predictors of effects in the receiving environment. Whole effluent toxicity tests are only the first stage in a risk assessment and as such identify hazard, not risk …WET tests have a degree of innate variability that will persist despite the most diligent attempts to eliminate it … Variability within about a factor of two is generally agreed to be the maximum acceptable range. In other words, if two laboratories conduct the same tests on the same sample, a difference within a factor or two may not be

[210] La Point, T.W. and W.T. Waller. 2000. Field assessments in conjunction with whole effluent toxicity testing. *Environ. Toxicol. Chem.* 19:14-24.

[211] Diamond and Daley (2000).

> unreasonable. This means that regulatory bright lines are generally not appropriate for judging WET test results … single exceedence of a WET test standard should not be considered to represent a significant potential for environmental damage. Uncertainty exists for WET tests … in some cases uncertainty can be high, particularly when single WET test results are extrapolated to receiving environments.[212]

5. Special Cases Must be Independently Validated, But Were Not.

Because WET test procedures have been designated "standard methods," they are routinely used by default even where it may not be appropriate to do so. EPA has cautioned that some methods may not be appropriate in all cases, but it did not delineate such instances within the method as required by 40 C.F.R. Part 136. For example, the Agency has said:

> EPA has not mandated which test methods NPDES permitting authorities must use under different exposure conditions ….[213]

> The existing whole effluent toxicity testing requirements do not specify whether applicants should test for acute or chronic toxicity ... Permit applicants should consult with the permitting authority to determine applicable testing requirements. Permitting authorities retain discretion to require testing for either acute or chronic toxicity.[214]

There are a large number of special circumstances where the standard WET method may not be appropriate, including: effluent-dependent streams, storm water exposures, streams with naturally high or low conductivity or other ionic imbalance, and estuarine systems. Discharges to effluent-dependent streams may indicate toxicity in a WET test, but on balance, support a greater richness and abundance of species than would arise naturally in the absence of

[212] Chapman (2000) at 3 and 8.

[213] Davies, Tudor T. and Michael B. Cook, U.S. EPA Office of Water, Memorandum to EPA Regional Water Management and Environmental Services Division Directors, *Clarifications Regarding Whole Effluent Toxicity Test Methods Recently Published at 40 CFR Part 136 and Guidance on Implementation of Whole Effluent Toxicity in Permits* (July 21, 1997).

[214] 64 Fed. Reg. 42,434, 42,449 (August 4, 1999).

augmented flows.

Storm waters are frequently deficient in several key ions, thereby causing WET tests to fail for reasons unrelated to chemical contamination. And, even when storm water runoff is exposed to pollutants, the exposure regime mandated by the WET test method bears no resemblance to the actual exposure conditions in the stream (e.g., 48 hours in an acute test versus a couple of hours in a rain event). The test tends to overestimate toxicity in such cases.

Many areas of the country have ambient stream chemistry that is naturally low in certain ions. Streams in other areas may be naturally high in conductivity, hardness, or alkalinity. In situations where a facility uses the stream for its production water, these factors may cause WET tests to fail for reasons unrelated to actual effluent quality.

> … test organisms may be sensitive to noncontaminant effects. For instance, increased hardness is a feature of some effluents, which can have an adverse effect on daphnids irrespective of contaminant concentrations. Variations in salinity and total dissolved solids can significantly affect WET test organisms.[215]

Estuarine systems present a unique challenge because they are too saline to use freshwater test organisms and, yet, not salty enough to use standard marine organisms (without manipulating the sample salinity). In such cases, the ionic imbalances alone may be sufficient to cause a sample to appear toxic.

In summary, EPA has not demonstrated the representativeness, comparability, or validity of WET test methods. Therefore, the methods are unsuitable for their intended purpose.

EPA's concept of independent application, in which WET test

[215] Chapman (2000) at 4.

> results can be considered in isolation from information on the receiving environment, is an example of misuse. The concept of independent application can only be justified if each component is predictive of the others. This is clearly not the case; WET tests are not always reliable predictors of receiving environment conditions … Risk assessments involve joint applicability, not independent application … Alone, WET tests cannot fulfill their stated purpose ("to identify, characterize, and eliminate toxic effects of discharges on aquatic resources").[216]

The Pellston Conference, which EPA cites to support the claim that WET testing is suitable for NPDES permitting, actually concluded that there must be limitations on the use of WET methods:

> The SETAC review concluded that … WET methods were not considered sufficient by themselves to make environmental hazard assessments. Other chemical-specific and bioassessment methods were still needed to understand the variety of stress mechanisms that affect any given receiving stream.[217]

III. THE PROPOSED METHODS DO NOT CONTAIN ADEQUATE QA/QC REQUIREMENTS

EPA's Section 518 Report at 6-1 states:

> While a fully validated and standardized method is desirable, it alone is not sufficient for the generation of valid data. A quality assurance and quality control (QA/QC) program including QA/QC procedures incorporated into methods are essential to detect and correct problems in the measurement process and to assure that the results generated are of a known and acceptable quality.

The Coalition emphatically concurs with EPA's statement; however, the proposed biological test methods contain inadequate QA/QC requirements.

EPA's failure to incorporate adequate QA/QC procedures into the proposed biological

[216] Chapman (2000).

[217] Ausley, L. 2000. Reflection on whole effluent toxicity: the Pellston Workshops. *Environ. Toxicol. and Chem*. 19:1-2.

tests is inconsistent not only with its statement to Congress, but also with previous decisions it has made under 40 C.F.R. Part 136. For example, EPA incorporated mandatory QA/QC protocols into the methods it promulgated for measurement of various organic compounds. In the preamble to that rule, EPA stated the objective of QA/QC:

> QA/QC seeks to assure that analyses of the same substances taken by different analysts at different times and places are of the same quality and are comparable within known statistical confidence limits.[218]

Nowhere could that objective be more important than in the context of compliance monitoring. By analogy, maintaining compliance with an NPDES effluent limitation using a test method to which inadequate QA/QC controls apply is akin to complying with speeding limits while on an automobile trip through successive towns, each of which uses radar guns that measure speed (i.e., are calibrated) differently.

In the context of NPDES compliance monitoring, where permittees are potentially subject to civil and criminal sanctions, it is essential that all measurements relied on by the regulatory authority reflect reliable and consistent laboratory practices. The proposed biological test methods do not contain the QA/QC protocols necessary to enable regulatory authorities, permittees, or labs to have a sufficient level of confidence that this is the case. Unless the test protocols are modified, the WET test methods will not be sufficiently reliable to be used for their intended regulatory purposes.

EPA's DQI Guidance states:

> In its broadest sense, data quality is a measure of the degree of

[218] 49 Fed. Reg. 43,234, 43,237 (col. 3) (Oct. 26, 1984).

> acceptability or utility of data for a particular purpose. To simplify the way data quality is examined, and to facilitate communication about data quality issues, certain data quality attributes can be defined and measured. The principal quality attributes important to environmental studies are precision, bias, representativeness, comparability, completeness, and sensitivity.[219]

EPA clearly believes that these attributes must be considered in the context of quality assessment. This section of the WET Coalition's comments will address, in part, several of the principal quality attributes and the deficiencies of the current WET methods in this regard.

A. EPA's Proposed Methods Provide No Basis For Ensuring Comparability of Test Results Within And Between Laboratories.

Inclusion of WET test procedures in 40 C.F.R. Part 136 and NPDES permits requires that they provide comparable results within and between labs. Comparability is a principal quality attribute important to environmental studies.[220] Users of data define data quality based on comparability and its DQIs and measurement quality objectives ("MQOs"). EPA's NPDES program, which includes the use of WET test data in the reasonable potential ("RP") and limit derivation processes, assumes the comparability of data. Yet DQIs and MQOs have not formally been established for the WET methods. This is a significant concern, because reasonable potential and compliance determinations typically rely on data from one or more analysts in a single laboratory and often on data from more than one laboratory. For purpose of this discussion, we will refer to data from more than one analyst or more than one laboratory as "pooled" data.

1. EPA Acknowledges The Importance of Comparability

[219] DQI Guidance at 5.

[220] *See* DQI Guidance.

According to EPA, establishment of MQOs or DQIs is necessary to address the impact of specific error sources on total study design. Error sources refer to:

> any factors that build uncertainty into a measured value. Generally speaking, these error sources are the result of natural variability in the sampled media and inherent imprecision in the measurement process.[221]

EPA states in its guidance that total error is a function of error due to measurements within sampling units and between sampling units. WET testing includes three different sampling units: the effluent tested, the organisms tested, and the laboratories doing the work. Each population of these units is sampled in every WET test, and each sample has error associated with it. As noted above, identification of error sources is important because it drives study design. For example, replication within a treatment should increase if within sample unit error exceeds between sample unit error. The value of the current WET test designs as proposed is problematic since study design determines the test result (WET is a method defined parameter) and the error sources that should be considered when developing study design have not been addressed in WET test design (the methods). Sources of error that should be addressed and are specific to the test methods include organism sensitivity and representativeness as well as lab expertise. Characterization of error sources is important to comparability, but it also applies to DQIs and MQOs for other data quality attributes.

EPA's guidance on DQIs states several times that comparability cannot be ignored when assessing data quality:

> Before pooling data, the comparability of data sets generated at different times or different organizations must be evaluated to

[221] DQI Guidance at 9.

> establish whether two data sets can be considered equivalent in regard to the measurement of a specific variable or groups of variables.
>
> Data sets that are representative of two different populations are generally not comparable with respect to pooling.
>
> Comparability is a very important qualitative data indicator for analytical assessment, and is critical when considering the combination of data sets with the same analytes.
>
> Only comparable data sets can be readily combined.[222]

2. EPA Does Not Provide The Means For Ensuring Comparability

One of the DQIs that EPA includes in its comparability guidance is the need for similar detection levels between measurements. The DQI Guidance states:

> Combining data sets having different detection or quantitation levels leads to difficulties in analytical interpretations.[223]

Nowhere in the 40 C.F.R. Part 136 methods is the problem of combining data sets with different sensitivities (detection levels) more significant than with the WET methods. Review of any method included in EPA's interlaboratory WET study illustrates the degree of variability in detection capabilities that exists within and between labs. For example, if one assumes the PMSD to be a reflection of detection capability in a WET hypothesis test, the 5^{th} and 95^{th} percentiles for the *C. dubia* reproduction PMSD distribution range from 10% to 43%, respectively.[224] Comparability of WET data used in the reasonable potential and limit derivation processes cannot be ascertained since capability to detect toxicity is not a DQI listed in the method, nor is there an MQO provided in the method to assess this aspect of comparability.

[222] DQI Guidance at 7-8, 33.

[223] DQI Guidance at 35.

[224] *See* WET Variability Guidance.

EPA attempts to address this issue using PMSD percentiles. However, this approach only modifies the interpretation of the data and ignores the actual comparability and quality of the data, because the upper PMSD percentiles used by EPA are not required to be met for the data to be considered valid. The PMSD is a measure of within test variability. As the PMSD varies between tests, so does the reliability and certainty of conclusions regarding each test. EPA ignores this fact and interprets all data with a PMSD greater than the 90th percentile of equal certainty (according to the proposed language of this rulemaking); this is not defensible. Data sets such as these are as dissimilar in their reliability as any two data sets with differing PMSDs below the 90th percentile. This has been and continues to be a primary criticism of the use of hypothesis tests in the WET program. Additionally, this aspect of comparability cannot be assessed since a detection limit has not been provided in the proposed WET methods.

Another comparability concern within the context of WET tests is that a number of parameters in the proposed methods can vary between tests and influence test results while still meeting method requirements. These parameters include dilution water quality, pH, food quality, and organism age (acute tests mostly). EPA states in its draft DQI Guidance that data conclusions without consideration of comparability may be based on:

> an artifact of methodological differences among the studies rather than differences in experimentally-controlled conditions.[225]

The parameters that vary within the limits of the methods may affect not only comparability of data, but also reasonable potential, permit compliance, or toxicity identification evaluation processes. Failure of the promulgated test method protocols to quantify the significance of methodological differences among labs prevents the use of such data in these

[225] DQI Guidance at 38.

processes. EPA does provide some recommendations on ways to assess the comparability of WET data statistically in its WET Variability Guidance, but these tools and approaches are not found in the methods as proposed.

3. EPA's Reference Toxicant Testing Procedure Does Not Ensure Comparability.

EPA clearly has failed to provide a basis for one laboratory to compare its test results with that of other laboratories. Instead, EPA has laboratories compare their performance against their own past performance. This entails laboratories (1) periodically performing biological tests on chemicals (i.e., reference toxicants) of known concentrations, and (2) developing a performance chart showing both the mean performance of that laboratory and the range of performance that the laboratory is expected to have 95% of the time. If the test results fall within the 95 percent confidence interval, the data are considered acceptable.

There are three fundamental flaws in EPA's only DQO for WET test comparability. First, the collection of long-term reference toxicant data is not mandatory. Laboratories must perform reference toxicant testing, but the development of control charts is discretionary.

The second shortcoming with EPA's reference toxicant procedure is that it does not promote reduction of variability within each laboratory. The 95% confidence interval for each control chart is established on the basis of the data produced by each individual laboratory (i.e., by calculating the standard deviation of the test results and multiplying by a factor of 1.96). The greater the variability of the laboratory's test results, the wider the calculated 95% confidence interval is around the mean. Variability increases differences between tests and decreases comparability and reliability of test results. Wider confidence intervals, therefore, represent lower test result reliability and performance. The consequence is that a reference toxicant test

result will be considered acceptable in some laboratories (i.e., those routinely performing with high variability) — but unacceptable in others.

Third, EPA states that reference toxicant results that fall outside the control chart limits "does not necessarily invalidate associated test results."[226] EPA leaves to the regulator decisions regarding how much of a deviation is acceptable. Absent any objective criteria for making that determination, the discretion eliminates whatever benefit the control chart concept otherwise might offer.

To the extent EPA considers unacceptable reference toxicant test results as a basis for invalidating the results of all toxicity tests performed concurrently with, and possibly within some period prior to, the reference toxicant test, the acceptability of reference toxicant results has direct bearing on compliance determinations. This emphasizes the need for MQOs for reference toxicity tests performance and control chart characteristics.

Further, reference toxicant test performance is used as a surrogate for WET test reliability. The variability and reliability of effluent tests theoretically parallel that of reference toxicant tests. EPA therefore must have an objective basis (DQI) for determining what constitutes acceptable reference toxicant results, and thus the quality of effluent data suitable for compliance determinations. Absent such a standard, as is the case with the WET test methods, permittees would be subjected to compliance determinations based on data whose reliability depends on the particular laboratory involved, as opposed to the actual quality of the effluent tested. Unless this problem is adequately addressed in the Part 136 rulemaking, it will result in unjustified enforcement actions, and evidentiary problems for government lawyers seeking to

[226] Proposed Method Manuals Changes at 65.

enforce permit requirements when data are available from more than one laboratory and/or analyst.

The Coalition recommends that EPA modify the proposed methods to require that the results of reference toxicant testing be compared against an acceptability range developed on an interlaboratory basis for each reference toxicant and a defensible approach for calculating the range.

EPA already has informally moved to some extent in this direction. For example, in its 1985 Acute Methods Manual, EPA stated that it would provide reference toxicants along with the expected LC_{50} values. The Agency has not followed through on that effort. The Coalition strongly supports this approach for all test methods. Now that EPA formally is finalizing its biological methods into Part 136, it needs to provide specific instructions on QA/QC protocols and DQIs/MQOs including published ranges for acceptable performance for each of the reference toxicants deemed acceptable.

4. Tracking Long-term Trends.

EPA recommends that laboratories track their test performance over time. Specifically, the Agency suggests that laboratories record the average response and variability for both laboratory dilution water (non-toxic culture water) and for reference toxicant conditions. These data can be used to assess the general health of test organisms. EPA's Section 518 Report states:

> The health of the test organisms or biological system being monitored is a unique attribute without a complimentary attribute for analytical methods. The health of a toxicity test organism has a profound effect on the quality of the data, and must therefore be considered a key criterion for assessing adequacy … The health of test organisms and biological systems cannot be 'calibrated' before the experiment in the same way as analytical instrumentation …

> There are no knobs to turn to adjust for these factors to achieve consistent performance during a test method. For these reasons, the biological procedure must include biological standards (e.g., standard reference toxicants) in order to ensure data integrity.[227]

Unfortunately, while EPA recommends that such data be collected, they do not require it. And the Agency fails to establish any QA/QC criteria by which to evaluate laboratory performance. Even the labs that voluntarily track such information measure their performance only against their own historical results. There is no independent standard by which to compare the performance of one lab against another. Thus, even if WET tests, when evaluated in an interlaboratory study, may exhibit variability similar to chemical test methods, the difference in standardized QA/QC procedures causes the variability to go largely unchecked when toxicity testing is performed in practice.

B. EPA's Proposed Methods Provide No Basis For Ensuring Representativeness of Test Results Within And Between Laboratories.

EPA uses the ANSI/ASQC definition for representativeness:

> The measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition.[228]

In regards to WET testing, EPA's guidance on DQIs states that WET data quality cannot be determined without information on whether:

> individual measurements of the characteristics of interest (WET) accurately reflect the conditions in the sampling unit, and whether an adequate number of units were measured to reflect the

[227] Section 518 Report at 3-11.

[228] DQI Guidance at 67.

population of interest.[229]

In the context of WET as an analytical method of measure, a sampling unit might include the organisms tested, the lab testing the organisms, and the effluent sample being tested. All affect the representativeness of a WET result. As stated above for comparability, there are no DQIs and MQOs for representativeness in the proposed methods relevant to testing labs or the organisms they use. The checklist provided in EPA's guidance on DQIs identifies where the proposed methods fall short regarding representativeness. Questions to be answered before one can conclude that the proposed WET methods provide representative results include:

- Was the population of labs and/or organisms defined prior to testing?
- Is there a statistical basis for sampling labs, organisms and effluent?
- Were MQOs for precision, accuracy and sensitivity set/achieved?
- Were appropriate methods of data analysis used?

It is clear that the currently proposed WET methods do not address these issues and, therefore, representativeness and quality of individual WET results cannot be adequately determined.

C. EPA's Proposed Methods Do Not Provide For An Adequate Assessment Of Bias In Test Results Within And Between Laboratories.

Please see Section II.D. on interferences resulting in test bias.

D. EPA's Proposed Methods Do Not Provide For An Adequate Assessment Of Sensitivity In Test Results Within And Between Laboratories.

EPA defines sensitivity in its 2001 DQI Guidance as:

[229] DQI Guidance at 75-76.

> the capability of a method or instrument to discriminate between measurement responses representing different levels of the variable of interest.[230]

EPA goes on to say that detection limits and sensitivity are closely related and used synonymously, and that sensitivity:

> is often a crucial aspect of environmental investigations that must make comparisons to particular action levels or standards.[231]

EPA has placed a great deal of emphasis on sensitivity as a data quality attribute and holds that users of methods should establish their objectives for sensitivity before beginning data collection. EPA never has established objectives for WET test sensitivity in the NPDES program. The guidance states that those responsible for selecting analytical methods for a project or program (in the case of the WET program, EPA and States) must:

> determine the levels of sensitivity needed to generate data adequate for decision, establish MQOs based on this evaluation, be sure that the indicator of sensitivity used to evaluate a particular method appropriately reflects the performance of the method in the particular matrix of interest . … [and] should always consider the needed sensitivity of a measurement prior to requesting laboratory analyses.[232]

Users of WET methods, including regulators and permittees, cannot follow this guidance because the levels of sensitivity and MQOs required for the WET program have not been defined for any matrix.

The concept of sensitivity is particularly relevant to WET testing where results are being directly compared to NPDES permit triggers, WLAs, or limits; yet DQIs and MQOs for

[230] DQI Guidance at 3.

[231] DQI Guidance at 9.

[232] DQI Guidance at 42.

assessing WET data sensitivity have not been proposed or established in the proposed methods. Therefore, users of the data generated using the proposed methods cannot adequately assess the quality of data within the context of its use. EPA states in its DQI Guidance that sensitivity is a function of instrument (method for WET) precision and slope of the calibration curve. EPA has data assessing the precision of the WET methods (the recent interlaboratory study) but has not established MQOs for precision in the WET methods. EPA did not report on the slope of concentration-response curves found for each method, matrix, and toxicant(s). Therefore, the impact of slope on method sensitivity has yet to be assessed; precluding development of MQOs for slope.

Sensitivity also can be assessed through replicate analyses at different concentrations of effluent or by using the confidence intervals of a modeled response. EPA has not addressed sensitivity using the confidence interval and avoided establishing MQOs using replicate analysis by changing how data are interpreted (PMSD percentiles). The PMSD percentile circumvents the real issue, which is intratest variability. MQOs for intratest and intratreatment variability have not been considered in the proposed methods. In short, sensitivity DQIs and MQOs are non-existent in the WET methods.

EPA states in its DQI Guidance that the

> sensitivity indicators of primary interest to EPA are indicators that relate to limits of detection.[233]

EPA has neglected to establish detection limits for WET methods. It therefore is not possible for users of the WET methods to adequately assess the quality of data generated using these

[233] DQI Guidance at 42.

methods. EPA is well aware of this criticism of its WET methods, yet it has chosen not to deal with the issue even after EPA's own Quality Staff reinforced the importance of this DQI. EPA's DQI Guidance openly acknowledges that detection limits and sensitivity vary by matrix and often among laboratories. The proposed EPA WET methods, however, fail to even broach the issue. EPA does not explain its rationale for this departure.

Given that the sensitivity DQI about which EPA is most concerned is a detection limit and that detection limits are not established in the methods, EPA has proposed methods that will not serve their appropriate and defensible use. EPA has published what it considers to be detection limits (for QA/QC purposes) for all of the chemical-specific test methods it has promulgated in Part 136, so it seems arbitrary for the Agency to have deviated from that longstanding practice in the WET rule. EPA has claimed in the past that detection limits cannot be derived for WET tests. However, the WET Coalition has developed an approach for calculating WET detection limits, and it is attached to the comments for EPA's consideration.[234]

E. EPA Must Adopt Performance Criteria Supporting Regulatory Use of WET Data.

The Agency has failed to adopt performance criteria that support the regulatory use of toxicity data. These criteria are necessary to define the reliability of procedures not only used to perform tests (since WET is a method defined parameter) but also those used to process and interpret test results. For example, when EPA's software is used to calculate confidence limits for the IC_{25}, it will not allow the upper confidence limit to exceed 100% effluent. The point estimate, instead, is forced to a lower concentration where the upper confidence limit is equal to

[234] Risk Sciences, Developing A Detection Level for Whole Effluent Toxicity (WET) Tests (2002).

100% effluent. As a result, the final promulgated method is significantly biased toward overestimating the potential for toxicity. This example illustrates the importance of developing and implementing acceptability criteria for statistical procedures used to analyze WET data.

Other examples of toxicity test elements requiring performance criteria to support the regulatory use of WET data include variability (control CV, PMSD) and characterization of the cause-effect (concentration-response) relationship. Variability has been documented numerous times in the peer-reviewed literature to influence the statistical analysis of toxicity test data. There are currently no performance criteria in the test methods addressing variability either within or between tests. Furthermore, EPA states in the preamble of the proposed rule that:

> the concentration-response relationship established between the concentration of a toxicant and the magnitude of the response is a fundamental principle of toxicology …. Use of this concept can be helpful in determining whether an effluent sample causes toxicity and in identifying anomalous test results.[235]

Although the Coalition does not disagree with the basic thought put forth by EPA on this topic, we believe that EPA has underestimated the importance of this relationship in correctly determining the toxic potential of a sample. Despite its importance, EPA has not provided sufficient performance criteria to allow a user to determine whether such a relationship has been satisfactorily demonstrated.

The above examples do not represent all aspects of the test methods that require performance criteria, but they should serve their purpose in characterizing the importance of developing and adopting performance criteria for all steps in the promulgated toxicity test methods.

[235] 66 Fed. Reg. at 49,799 (col. 3).

F. EPA's QA/QC Protocols Must Be Mandatory.

All QA/QC protocols and DQOs for test methods to be used in NPDES compliance determinations must be mandatory under all circumstances and must be clearly defined and presented as requirements. Aside from the TAC, none of the other test conditions in the protocols are mandatory. Some are explicitly discretionary, without justification. For example, EPA requires reference toxicant testing of batches of test organisms obtained from an outside source, but only recommends such testing for laboratories that culture their own organisms. Coalition experience has been that the source of the organisms has little impact on the reliability of those organisms in detecting the toxic potential of a sample. In fact, some organism distributors can provide much higher quality organisms than in-house cultures.

Other test conditions appear to be mandatory, by virtue of the use of terms such as "must" or "shall," but in fact are rendered discretionary with the following statement:

> Deviations in test conditions (from the specifications in the summary of test condition tables) must be evaluated to determine the validity of test results. Test condition deviations may or may not invalidate a test result depending on the degree of the departure and the objective of the test. The reviewer should consider the degree of the deviation and the potential or observed impact of the deviation on the test result before rejecting or accepting a test as valid.[236]

In short, EPA first imposes conditions it presumably deems essential to ensure reliable test results (i.e., test results at least as reliable as would be predicted by the validation study on which EPA relied to conclude that the test method was suitable for Part 136), and then authorizes regulatory authorities to "accept" test results, on a case-by-case basis, even when they originate from a laboratory that deviated from the test conditions in the method protocol. EPA grants that

[236] *See, e.g.*, Proposed Method Manuals Changes (§ 12.2.4.2) at 58.

discretion without any objective standards for its exercise.

Thus, a regulator may "accept" a test result showing an excursion of a WET permit limitation, even though the data point came from a laboratory that deviated from test conditions in the method protocol. Yet, absent objective standards for evaluating such deviations, a different regulator in an adjacent state, or even in the same office, might have concluded that the test result must be "rejected." Given the obvious unfairness to permittees, EPA must eliminate that arbitrary feature from all of its test protocols.

If EPA concludes that certain test conditions are essential to ensure reliability, those conditions must be mandated. If EPA concludes that certain test conditions can be waived on a case-by-case basis, it must provide: (1) evidence in the Part 136 rule confirming that such deviations will not affect reliability, and (2) objective standards in each test protocol for regulators to use in deciding how much of a deviation is acceptable. Finally, as to test conditions that EPA decides to mandate, it should consider expanding the flexibility built into those conditions (e.g., ranges) to the extent that can be justified without compromising reliability.

G. EPA Must Clearly State QA/QC Defining Test Validity.

Given the consequences of toxicity test results in the regulatory process, EPA must not only impose QA/QC requirements to ensure the quality of test results but also must specify what must be done when QA/QC procedures demonstrate that a problem exists (i.e., that the "system is out of control"). The proposed protocols are either silent or ambiguous about this issue. For example, in the 1993 acute manual, EPA states:

> If the toxicity value from a given test with the reference toxicant falls well outside the expected range for the test organisms when using the standard dilution water, the sensitivity of the organisms

and the overall credibility of the test system are suspect.[237]

This could mean several things: (1) that the results from actual toxicity tests run concurrently with the reference toxicant tests always will be deemed invalid, (2) that the results from actual toxicity tests run during some period prior to the unacceptable reference toxicant test are invalid, or (3) that the validity of test results is left to the discretion of the regulatory authority and/or the lab. EPA must specify what it expects labs and/or regulatory authorities to do with QA/QC results so that the regulated community has an opportunity to comment on the factors defining data validity and whether it is "true and accurate". This information is critical, given it will directly influence the source of data on which compliance determinations will be based and the legal status of permittees following DMR certification of WET results.

H. EPA Should Modify The QA/QC and Other Conditions In Its Test Protocols.

As discussed elsewhere in Section III, EPA needs to modify its test protocols to be definitive regarding which QA/QC and other conditions are mandatory (i.e., necessary to ensure reliable performance),[238] and which are discretionary (i.e., not necessary to ensure reliable performance). EPA needs to modify the test conditions that will be mandatory to be as flexible as possible (e.g., broad ranges) without compromising the reliability necessary for Part 136 methods. For example, the sample temperature upon receipt at the laboratory is required to be 4°C, but the requirement could be made more flexible by specifying instead that a range of 2°-6° is acceptable.

As evidenced by the large percentage of data points EPA accepted for use in its

[237] Acute Methods Manual (Section 4.15.4) at 16.

[238] *See* EPA Settlement Agreement language earlier in Section III.

Interlaboratory Study, even though the data came from laboratories that deviated from the Part 136 protocols, EPA apparently agrees that the current test conditions warrant modifications. Over the past several months, the WET Coalition has offered EPA specific examples of changes that should be explored. The list of suggestions is provided as an attachment to these comments. The WET Coalition urges EPA, prior to issuing its final rule, to evaluate the list and make those changes for which it has sufficient evidence to confirm that the change will not compromise reliability.

IV. ADDITIONAL COMMENTS AND SUGGESTIONS

A. New Statistical Methods And Approaches Are Needed

The disadvantages of the statistical approaches currently supported in the WET methods and used in the NPDES program have been studied and documented since the use of some of these methods began in the early 1980s. Hypothesis tests, as proposed, suffer from significant variability in sensitivity within and between labs. Their endpoints (NOEC, NOAEC) cannot be added, subtracted, multiplied, or divided in the reasonable potential and limit derivation/compliance processes and are driven by which dilutions are tested. Furthermore, they do not include the calculation of confidence limits. The proposed point estimate techniques assume linear response regardless of whether the data support this assumption, they bias endpoints due to the linear assumption and smoothing of data, and they do not reliably calculate confidence limits used to judge the reliability of the endpoints. None of the current EPA proposed models for continuous data are parametric. Some methods, such as ICp, assume that there is no fixed mathematical relationship among tested concentrations but require a monotonic relationship between concentrations. This requirement forces data into a model that cannot reliably represent the data.

It is critical that the toxicity data, generated from a wide range of tests and biological endpoints, fit the statistical methods and assumptions included in the methods. This is not the case for the proposed methods, which are fairly rigid in their construct and assumptions. This is the root of many of the problems experienced with the current approaches. The 1989 version of the acute manual stated that only 7% of reviewed effluent toxicity tests where test organism lethality occurred were amenable to Probit analysis, for example. Comments also have been submitted by attendees of the 1995 SETAC WET workshop regarding the shortcomings of EPA's proposed statistical approaches. Contemporary experience supports these observations.[239]

The models used in the WET methods must be flexible enough to represent a wide range of concentration-response curves independent of the concentrations tested. The Coalition believes that EPA must revisit the approaches being proposed based on objective criteria that support their use in the NPDES program. Criteria used to select statistical approaches must include adaptability to different data sets/concentration-response curves, minimization of assumptions required to support their use, an ability to always calculate defensible confidence limits, an ability to use all data generated in a test to calculate confidence limits and statistical endpoints, computer resources required to run programs, software requirements, training and experience required to support use of each approach, etc. It is clear that these criteria were not used to select the statistical approaches currently proposed.

It is possible that some of the current procedures can be adjusted to address their shortcomings. However, the Coalition also urges EPA to allow explicitly the use of alternative point estimate and hypothesis test approaches that support reliable use of WET data in the

[239] Grothe, D.R., et al. (eds.). 1996. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts. SETAC Press, Pensacola, FL.

NPDES program (given that most statistical approaches cannot be applied to all types of data sets). Alternative approaches may include, but are not limited to, new statistical endpoints, such as percent effect,[240] and new procedures, such as general linearized models ("GLMs"). These approaches address most of the significant disadvantages associated with those proposed without introducing additional disadvantages. Statistical approaches have evolved a great deal since the first draft of these methods was proposed, but the WET methods have not taken advantage of these advances. EPA must provide an opportunity for stakeholders to use these more powerful and reliable tools within the context of WET NPDES testing; this can be accomplished only by providing this opportunity within the text of the methods.

B. Dose Response.

The premise of toxicity test design is that there is a relationship between the concentration of exposure and specific biological effects. It is this relationship that supports the calculation of statistical endpoints such as NOECs, LC_{50}s and IC_{25}s. Failure to adequately characterize this relationship significantly reduces the reliability of these statistical endpoints and will preclude their use in a regulatory context. See correspondence of Norberg-King to Region X EPA, June 5, 1989:

> The dose response curve is the basis for the validity of a toxicity test. The control serves as the starting point from which the dose response is evaluated. If a dose response is not obtained, the toxicity can not be inferred.

EPA also states in the preamble:

> The concentration-response relationship established between the

[240] *See, e.g.,* Risk Sciences, Regulating Whole Effluent Toxicity Using 'Percent Effect' As the Test Endpoint (2001).

> concentration of a toxicant and the magnitude of the response is a fundamental principle of toxicology.
>
> Use of this concept can be helpful in determining whether an effluent sample causes toxicity and in identifying anomalous test results.[241]

Despite the importance of the curve to which EPA refers when addressing concentration response relationships, EPA did not provide a procedure in the proposed WET methods to characterize the dose-response relationship and did not provide criteria to document that a concentration-response relationship exists. EPA only proposes to include review procedures for interpreting data given varying concentration-response scenarios, rather than decision criteria to judge the presence or absence of such relationships. EPA claims that use of this review would ensure that a valid concentration-response relationship is demonstrated. This is not true, as the guidance only aids in the interpretation of multi-dilution tests. The methods do not require explicitly that such a relationship exist prior to use of data in a regulatory context. Hence, the methods are incomplete and incapable of rendering a definitive and defensible assessment of WET.

In its preamble, EPA stated that:

> [e]ight of the 23 test results considered anomalies or inclusive had erroneously indicated toxicity in blank samples. These results would have been reported as false positives if the concentration-response review procedures had not been used.[242]

EPA developed guidance on interpreting multi-dilution tests (EPA 821-B-00-004) but failed to develop DQIs and MQOs to define the absence or presence of concentration-response

[241] 66 Fed. Reg. at 49,799 (col. 3).

[242] 66 Fed. Reg. at 49,800 (col. 1).

relationships in the methods. Aside from the guidance's failure to provide definitive procedures and benchmarks for identifying valid response relationships, this guidance, unfortunately, is biased to interpret data in an unreasonably conservative fashion rather than in a way that is scientifically supportable. For example, if a 0.5 dilution series is tested and all concentrations show no effect except for the 25% and 100% dilutions, the guidance suggests that the LOEC is 25% and the NOEC is 12.5% even if no statistical difference was noted with the 50% effluent concentration.

The Coalition does not support EPA's guidance on concentration-response relationships and holds that this guidance will not meet its intended goal. EPA must delete reference to this guidance in the methods, as it relates to concentration-response relationships, and replace it with specific criteria consistent with these comments. The Coalition recognizes that EPA's guidance is non-binding, but it likely will be used as such unless EPA states explicitly that WET test users may rely on other defensible approaches to defining the presence/absence of concentration-response relationships and interpreting these relationships.

Given the crucial role of concentration-response relationships as acknowledged by EPA, the use of WET data based upon questionable concentration-response relationships is indefensible and results in unjustified findings of reasonable potential and/or triggers limit exceedences. The methods must include a requirement for a valid concentration-response relationship before each WET test result can be used in a regulatory context, as well as procedures, DQIs, and MQOs that reliably establish the presence of this relationship.

The importance of the concentration-response relationship is emphasized when the permittee must demonstrate no toxicity at a 100% effluent concentration (which is a very

common NPDES permit requirement). One commonly assumes that, if toxic impact are measured in 100% effluent, and all other concentrations do not suggest toxicity, the toxicity is due solely to the effluent. Toxicology teaches that greater impact (or any impact at all) should be found at higher concentrations, so the assumption superficially seems justified. This is also the logic used in example #7 of EPA's guidance on concentration-response relationships.[243] The bias introduced by this assumption is obvious, however, when we look at how the interpretation changes when a dilution other than 100% suggests toxicity and all other concentrations, including the 100% effluent concentration, do not. In this case, we often assume that the single indication of toxicity is an outlier and the organism response is not representative. EPA reinforces this logic in example #5 of its guidance on concentration-response relationships. The responses at both concentrations could be unrepresentative, but the interpretation changes depending on which dilution is under consideration. It is just as probable for the 100% response to be an outlier as it is for any other tested concentration. The Coalition believes that at least two adjacent dilutions in the dilution series must show impact before a concentration-response relationship can be confirmed and an endpoint calculated. If this does not occur using low dilution factors, higher dilution factors should be used to bring more tested concentrations closer to the concentration allegedly causing toxicity (100%). This may require dilution factors as high as 0.9 or 0.95.

C. Proposed Use Of PMSD Percentiles To Interpret Hypothesis Test Results.

As discussed below, the WET Coalition has some serious concerns and reservations regarding EPA's proposed use of PMSD percentiles to interpret WET tests.

[243] WET Testing Guidance at 4-14.

1. 10^{th} and 90^{th} Percentile PMSDs are Arbitrary.

The percent minimum significant difference (PMSD) is a statistic representing the percent difference between tested effluent and controls in a single test that can be detected as statistically significant, usually at an alpha of 0.05. EPA currently proposes to use PMSD percentiles to interpret the results of WET tests when the hypothesis test approach is used to statistically derive test endpoints (NOEC, LOEC). The 10^{th} percentile of a PMSD population, as characterized in EPA's 2001 WET Variability Guidance, is used as a measure of method precision and a censoring point for interpreting the results of hypothesis tests. For example, the 10^{th} percentile of the chronic *C. dubia* reproduction test PMSD is 11%. EPA proposes that any effluent concentration resulting in a difference of 11% or less from the controls in the average number of juveniles produced in a *C. dubia* chronic test will not be interpreted as a reportable difference even if the hypothesis test finds that the difference is statistically significant. This approach recognizes the inability to be acceptably precise at or below the 10^{th} percentile PMSD. EPA also is using the 90^{th} percentile of each PMSD population corresponding to a particular biological endpoint to interpret results on the other end of the PMSD distribution. EPA reports the *C. dubia* 90th percentile PMSD for reproduction as 35%. Any effluent concentration resulting in a difference of 35% or greater from the controls in the average number of juveniles produced in a *C. dubia* chronic test will be interpreted as a reportable difference even if the hypothesis test does not find that the difference is statistically significant. The 10^{th} and 90^{th} percentile PMSD approaches are proposed for the sublethal endpoints of the *C. dubia*, P. promelas, *M. bahia,* and *M. menidia* chronic tests.

The selection of the 10^{th} and 90^{th} percentiles of PMSDs to interpret data is purely arbitrary and is not adequately justified by EPA. EPA must show that these percentiles are

defensible censoring points for interpreting data based on approaches and concepts used for other methods common to the NPDES program.

2. EPA Failed to Provide PMSDs for all Biological Endpoints and Tests.

In addition to the above recommended improvements needed in the PMSD approach, the concept needs to be expanded in applicability. EPA has only proposed this approach for a small fraction of the total number of WET test biological (survival, growth, reproduction) endpoints, rather than for all biological endpoints. The purpose of the 10^{th} percentile approach, for example, is to address test results that may be a function of atypically high levels of statistical power rather than the presence of actual effluent toxicity. Statistical differences from controls due to effluent exposures is not likely to be found when the impact relative to controls falls below the 10^{th} percentile of the PMSD population. This is logical since statistical differences at or below the 10^{th} percentile level only occur 10% or less of the time between all labs. Lack of a proposal for a 10^{th} percentile for all endpoints infers that either hypothesis tests have no quantitative limits to predicting toxicity due to an effluent exposure or the 10^{th} percentile for the PMSDs for these other endpoints is less than, say, 1%. EPA's Variability Guidance and the 2001 Interlaboratory Study both confirm that the 10^{th} percentile PMSD is greater than 1% for all WET methods studied. It is also unrealistic to expect, given the variability documented for WET tests (comparable to chemical tests according to EPA), that these tests are capable of reliably predicting toxicity (with acceptable precision) due to an effluent exposure at "any" percent difference from controls, regardless how small. This is comparable to saying that EPA Method 200.7 for metals is capable of reliably measuring any quantity of copper regardless of how low it is. This is obviously not the case and cannot be defensibly claimed for WET tests. The Coalition does not support the proposed use of PMSDs as detection limits. However, if the

Agency insists on this approach, failure to provide these limits for any method or endpoint proposed precludes the use of those methods/endpoints in a regulatory setting. The percentile approach (both upper and lower) must be applied to all biological endpoints proposed by EPA for use in the NPDES program or those methods and endpoints without limits must be removed from the promulgated list of methods and endpoints.

3. EPA Failed to Provide PMSDs for all Statistical Endpoints and Tests.

EPA not only neglected to propose criteria for interpreting some test biological endpoints (survival, reproduction, growth), but it also failed to address some statistical endpoints. For example, NOAECs are being used exceedingly in NPDES permits to substitute for LC_{50} requirements typically when the initial dilution is not abundant instream. Acute tests analyzed with hypothesis testing techniques suffer from increasing degrees of uncertainty as the differences between control and effluent exposure organism response decreases, not unlike that observed for chronic tests. The Coalition does not support the proposed use of PMSDs as detection limits. However, in order to use the PMSD approach in the context of NPDES permitting, EPA must include in the proposed methods interpretive criteria for acute test NOAEC that are analogous to those proposed for some chronic tests.

4. EPA Failed to Develop Tools to Address Uncertainty in Point Estimates.

The PMSD percentile concept was developed to specifically address interpretation issues specific to hypothesis test endpoints, such as the NOEC and LOEC. However, EPA has not provided any language in the proposed methods to address limits of precision and uncertainty pertaining to point estimates like the LC_{50} or IC_{25}. Point estimates have a degree of uncertainty often times characterized by 95% confidence limits around the estimate. These limits indicate that there is a 95% probability that the actual estimate falls between these limits. Point estimates,

as a statistical approach, have limitations in their ability to differentiate organism response in effluents from that occurring in controls. When the confidence interval overlaps a 100% effluent exposure, it is not possible to determine reliably that the point estimate is actually different from 100% effluent. An effluent cannot reliably be deemed toxic if its point estimate is not different from 100% effluent, given that a point estimate of 100% is, by definition, not toxic. One would conclude in this example that a difference in the point estimate of the effluent and that representing no toxicity could not be established with a specific level of statistical confidence. The same conclusion would be drawn when the confidence interval of a tested effluent's point estimate overlaps an effluent concentration selected as a permit limit. In this example, one could not conclude that there was a difference in the effluent point estimate and the effluent concentration predicting impact (the limit) with a specific level of statistical confidence. Both point estimates and hypothesis test endpoints have limitations of precision and reliability; the proposed methods must include DQIs and MQOs addressing these limitations.

5. EPA Failed to Update its PMSD Limits Using the Interlaboratory Study Results.

The database used to characterize PMSD percentiles proposed in the methods was not created to specifically represent each population of the test endpoints, but was assembled based on Agency requests for data from labs and regulatory agencies. This database was not necessarily assembled to accurately represent the current capabilities and performance levels of labs in all parts of the country. However, the recently completed Interlaboratory Study of the proposed WET methods was designed to characterize the current capabilities of WET labs. In some cases, the differences in PMSD percentiles are small between the two databases. In other cases, they are large. It would not seem defensible, however, to combine the databases because they were collected under different circumstances. It will be necessary to confirm that the data

sets represented the same populations before the data sets could be combined.

6. EPA Failed to Adopt the 90th Percentile PMSD as a TAC.

The Coalition also is concerned that EPA is using the 90th PMSD percentile simply as a tool to determine the LOEC rather than as a DQI. EPA's approach is designed to address tests with high PMSDs, low power, and low sensitivity. Without an approach to directly deal with tests of low sensitivity, one may conclude no toxicity when toxicity is present (false negative). EPA states in section V of the preamble:

> Application of the PMSD approach is intended to control the within-test variability in WET methods.[244]

The PMSD approach included in the proposed methods does not control variability because it is not a method performance requirement. EPA merely interprets the data independent of any statistical tests, which addresses the symptom (poor sensitivity) and avoids the real problem, which is intra-test variability. The fundamental difference between the EPA approach and one defining a DQI is that labs can continue operating at their current level of performance, regardless of how poor it may be, without any incentive for improvement in within-test precision. However, if specific levels of PMSD were used as WET test TAC, there would be an incentive for labs to improve their precision within and between tests because tests not meeting this TAC would be deemed unacceptable. EPA's approach only favors laboratories, while neglecting to address the needs of the data users, i.e., permittees and regulators. EPA must include TAC in the methods that control precision within WET tests, rather than procedures designed to merely sidestep the issue.

[244] 66 Fed. Reg. at 49,811 (col. 2).

7. The 10th Percentile PMSD is no Substitute for a WET Detection Limit.

Finally, and most importantly, the 10th percentile PMSD is not an adequate or defensible substitute for a WET MDL (which is the term EPA uses for detection limit). EPA states in the preamble:

> The purpose of the lower PMSD bound is to avoid declaring as 'significant' toxic effects that are smaller than those that can generally and routinely be detected by the method as currently conducted by qualified laboratories.[245]

By definition, the 10th percentile PMSD cannot be "generally and routinely" attained; the 10th percentile PMSD can only be attained 10% of the time.

EPA defines the MDL as:

> the minimum concentration of an analyte (substance) that can be measured and reported with 99% confidence that the analyte concentration is greater than zero ....[246]

While the Coalition takes issue with the MDL in terms of how it is calculated (Part 136 App. B) and often used in practice, the definition itself is suitable for a detection limit concept. The PMSD concept, as proposed, is inconsistent with EPA's detection limit concept. The similarities in the concepts presented by EPA in both cases are striking and indisputable. However, there is a significant difference between EPA's 10th percentile PMSD and an MDL. EPA has not provided a link between the PMSD percentile chosen and a confidence level that the effect represented by that PMSD is different than that experienced without toxicity (dilution water) across laboratories. For example, data or analyses have not been provided to show that one can be 99% confident that an 11% difference (the 10th percentile for this test and endpoint)

[245] 66 Fed. Reg. at 49,812.

[246] 40 C.F.R. § 136.2(f).

from a control in a chronic *C. dubia* reproduction test is statistically greater than a zero percent difference and that a 11% difference can be "generally and routinely" detected. The fundamentals of the PMSD approach proposed by EPA contradict the concept of MDLs in that EPA claims to select MDLs based on the ability of representative labs to reach routinely that level of sensitivity using the method. The MDL adopted for a method, and its reliability and attainability, are based on a review of MDLs achieved by labs in an interlaboratory study. Labs can reach the 10^{th} percentile PMSD only ten percent of the time, while EPA expects an MDL for a 40 C.F.R. Part 136 method to be met most of the time. By EPA's definition, the lower PMSD is to address apparent effects that can "generally and routinely be detected" by labs. The lower PMSD can be attained only 10 percent of the time; this cannot be considered routine performance. The MDL concept applied to PMSD probably would result in a censoring point specific to detection for WET methods above the 75^{th} percentile PMSD, rather than the 10^{th} percentile.

EPA additionally states in 40 C.F.R. Part 136, Appendix B on calculating the MDL that:

> The MDL obtained by this procedure is used to judge the significance of a single measurement of a future sample.

Likewise, the 10^{th} percentile PMSD is designed to judge the significance of a single measurement of a future sample. Based on these observations, the Coalition can only assume that EPA proposed this approach to address detection/sensitivity issues pertaining to the WET methods.

Given these shortcomings the Coalition cannot support the use of the 10^{th} percentile PMSD as a censoring point for detection in the WET methods.

8. EPA's PMSD Limits Only Apply To Tests Conducted In The Same Way As Those Used To Develop Those Limits.

The PMSD percentiles apparently were developed based only on tests that met the assumptions for running Dunnett's hypothesis test, because this is the only procedure in EPA's methods that generates a MSD. However, if Dunnett's assumptions are not met when assessing the PMSD in practice, other statistical approaches must be used. When this occurs, the PMSD may change. Therefore, the basis of the PMSD for a particular test may be different than that used to calculate the PMSD limits, making the use of these limits unsupportable.

This also raises the question of whether EPA used Dunnett's to calculate PMSDs even when the statistical assumptions of Dunnett's were not met. This approach may invalidate the PMSD limits that EPA presented in its guidance. PMSD also will be dilution series specific. Intratest variability is a function of dilution series, as dilution series decreases one would expect intratest variability, in many cases, to decrease. If most of EPA's data base used to calculate the PMSD limits is based on tests using a 0.5 dilution factor, and the PMSD is a function of intratest variability, it would be inappropriate to compare a PMSD from a test with a dilution series of, say, 0.75 to the PMSD limits EPA has derived. Intratest variability may be greater in this higher dilution series than if a 0.5 dilution series were run; therefore, it could not be considered a member of the PMSD population that EPA developed. This concern also applies to all changes that EPA is making to the current methods. If these changes impact PMSD, then the PMSD distributions characterized for each test and endpoint are no longer representative and appropriate. EPA must adequately address this issue and each change proposed in the rulemaking.

D. Test Acceptance Criteria And DQIs.

1. North Carolina Additional TAC.

The State of North Carolina has incorporated a number of changes to the chronic *C. dubia* reproduction test, as published by EPA, that are intended to minimize variability between tests. Changes were made to this particular test because it is the most frequently used chronic test that this State uses in its NPDES program and presumably the most sensitive endpoint. These changes include:

a. Only 1-3 broods are used to calculate reproduction for each treatment; neonates from four or more broods are not included in the reproductive count for each replicate. This improves comparability between treatments and the control as well as between tests.

b. Males can only represent 20% of the controls in each test. Since the presence of males infers stress on the population used to support the test, a higher percentage of males would suggest that the organisms used in the test were stressed to an unacceptable degree. This source of stress could bias test results to predict more toxicity than is actually associated with the tested sample, but it also may reduce the sensitivity of the test by reducing the total number of reproducing females in the controls. Fewer females result in fewer neonates; this decreases the probability of finding statistically significant differences between controls and treatments.

c. The CV for reproduction in controls cannot exceed 40 percent. As the CV in the controls increases, the power of the toxicity test to detect differences from the controls decreases. This TAC limits within test variability, which in turn will impact inter-test variability.

d. This test is always terminated seven days after it is started, $\pm$ 2 hrs. North Carolina's experience has been that the third brood is routinely attained during this time period and longer durations are not necessary. Shorter durations may bias the test to predict more toxicity because effluent concentrations may not reach the third brood due to the impact of acclimation to the tested matrix. Longer durations are not necessary. This standardizes test exposure and should control inter-test variability.

e. North Carolina uses a Practical Sensitivity Criterion ("PSC") for this endpoint. Any effluent concentration that is statistically different from a control and less than 20% different from the control is not considered to be definitively impacted relative to the control. The PSC sets a limit on test sensitivity (detection limit) and controls inter-test variability by equating all treatments, regardless of when the test is done, in terms of impact. Although the PSC is a better representation of the detection limit than the 10^{th} percentile PMSD, it still cannot be attained most of the time by labs across the country, which is a requirement of a detection limit established by EPA.

f. Pass/fail tests (one treatment versus control) using hypothesis statistics in acute and chronic tests also use an alpha of 0.01 rather than 0.05 as suggested by EPA's methods. This decreases the statistical rate of false positives five-fold and increases certainty in conclusions of pass/fail tests by an equal ratio.

Each of these adjustments improves the confidence that all stakeholders can have in the results of WET tests. Aside from reservations expressed above specific to the PSC, the Coalition supports these approaches and recommends that they be included in the final rule.

2. Increase TAC for *C. dubia* reproduction and *P. promelas* Growth.

The Agency has requested comment and recommendations on increasing the chronic *C. dubia* reproduction and P. promelas weight TACs. However, EPA did not provide justification for considering these changes, why only these TACs should be changed, or a proposal and justification for the actual changes. EPA must provide sufficient information in the proposed rule or its preamble to address these questions before the Coalition can provide meaningful comment.

For example, EPA infers in the preamble of the proposed rule that these changes would improve the performance of the WET test methods. EPA therefore is implying that "performance" increases as the reproductive potential of daphnids and the size of larval fish at the end of a chronic test increase. It is unclear exactly what EPA means by improved performance; without a clear understanding, meaningful comment cannot be provided. EPA needs to elaborate on what aspects of performance it is referring to in the proposed rule. It is unclear whether EPA is referring to test precision, sensitivity, comparability, representativeness, bias, or some combination of these data quality attributes. It is also unclear from EPA's request whether it is inferring that these changes would improve the ability of WET tests as an analytical method or as a means to predict instream effects. Again, without more information to clarify the

Agency's intent and data supporting its position, it is not possible to provide definitive comment or recommendations.

Finally, the criteria for choosing the original TAC must be provided by EPA in order to determine if the changes proposed are in agreement with these criteria, assuming they are technically and scientifically defensible.

3. Increase Minimum Number Replicates in Chronic Fish and Sea Urchin Tests.

The Coalition supports an increase in the minimum number of replicates for these tests to a total of four per control and concentration tested. This change is required to support the use of the non-parametric hypothesis tests outlined in the procedures when the minimum statistical assumptions of parametric tests cannot be met. This change may or may not improve performance, depending on how the Agency defines "performance".

4. Increase Minimum Number of Replicates in *C. dubia* Chronic Test.

The Agency has requested comment and recommendations on increasing the chronic *C. dubia* reproduction TAC based on a desire to improve the performance of the test method. However, EPA did not provide a definition for "performance", a justification for considering this change, and why this test should be modified in this way. EPA must provide sufficient information in the proposed rule or its preamble to address these questions before the Coalition can provide meaningful comment. Data and appropriate analysis supporting the assertion that this improves the performance of the method in a specific way must be provided by the Agency. Furthermore, the Coalition cannot support an increase of replicates in this test or any other test that currently requires four or more replicates without calculation of a defensible, interlaboratory detection limit below which uncertainty of the data is too great to conclude a difference from

control response.

E. ICp Issues

The currently proposed chronic test methods recommend the use of the ICp model to calculate point estimates for sublethal endpoints. This model suffers from a number of problems that may influence the calculation of the desired endpoint. Therefore, it is possible that the reported result of a chronic toxicity test may be a function of the model's deficiencies rather than the tested effluent's quality. Most of these problems stem from two of the primary assumptions of the model, that the responses are monotonically non-increasing and that they follow a piecewise linear response pattern. When the responses are not monotonically non-increasing, the model smoothes the data by averaging mean responses from concentrations adjacent to one another in the dilution scheme. For example, assume that the following results were reported for a *C. dubia* chronic reproduction test:

Effluent conc. (%)	# neonates (avg.)	smoothed#
Control	20.5	20.5
6.25	20.2	20.2
12.5	20.1	20.1
25	18.4	14.5
50	4.6	14.5
100	20.5	14.5

The current method does not include a procedure for testing and removing outliers; therefore, the 50% effluent concentration response is considered valid. Following the procedure for monotonically increasing responses, the data is smoothed. Although the response in most concentrations, including 100% effluent, was nearly identical to the control, the smoothing procedure reduced the average number of juveniles in the three highest concentrations to a level inferring a 29.3% reduction in relation to the control. The model therefore would predict a IC_{25} slightly less than 25% effluent even though all treatments equal to and less than the 25% concentration resulted in a 8.9% reduction, or less, in reproduction relative to the control. This is a good example of how the response of a single concentration (one that does not follow the concentration-response relationship defined by the other tested concentrations) results in the calculation of an IC_{25} that is contradicted by the raw data. The data actually suggest that 25% reduction in response occurred at only one concentration and that significant impact was not even measured in undiluted effluent. The ICp model is not capable of objectively dealing with

such data sets and will overestimate the potential for toxicity.

Another problem with the ICp model stems from its assumption of linear response between concentrations. This is particularly troubling when using the 0.5 dilution scheme because the model is forced to assume the shape of the concentration-response relationship between concentrations that can be very widely separated. This is best illustrated when comparing responses between the 50% and 100% effluent concentrations. Compliance with a trigger or limit between these concentrations may be totally a function of the response in only one of the treatments, and this treatment may not be representative of its population. If the response in the 50% is slightly higher or lower or the response in the 100% treatment is slightly higher or lower, the slope of the line connecting these two responses changes as does the IC_{25} (assuming that the response at 100% effluent is $\geq$ 25% less than the control). The information provided with this type of dilution series (the most common series used in the NPDES program) is insufficient to justify the assumption that the response is linear with concentration in the 50%-100% effluent range. Over such a wide concentration range it is very possible that the concentration-response relationship is not linear and that the concentration resulting in a 25% reduction in response is different than that calculated by the ICp model. Again, the model is dictating the test result rather than the quality of the effluent tested.

The first two examples above illustrate yet another issue of concern to the Coalition. One dilution can affect the entire IC_{25} calculation even though data from four other dilutions and the control are available. This is a significant shortcoming of the ICp model (as well as hypothesis tests). Most of the data generated in each and every test is not used to calculate the IC_{25} endpoint; only those concentrations bracketing a 25% reduction in response relative to controls are used to calculate the IC_{25}. Although responses from multiple dilutions help establish the

concentration-response relationship across dilutions spanning a 16x span (in this example), one cannot assume that the presence of a relationship across the entire test translates to a predictable relationship between any two dilutions, particularly those as different as the 50% and 100% concentrations. The nature of the relationship between concentration and response is interpolated between tested concentrations; therefore, the extent of interpolation and the uncertainty in that interpolation increase as the magnitude of the difference in two adjacent tested concentrations increase. The shape of the curve between, say, 50% and 100% will be a function of the organisms tested, their testing conditions, and the concentrations of toxicants causing response at each of these dilutions. Since this information cannot be reliably predicted (because the toxicants are often unknown or acting in synergy or antagonistically), the shape and nature of the curve is unknown, at best.

Yet another issue with use of the ICp model is that it is biased towards calculations of lower IC_{25} values when the upper confidence limit exceeds 100% effluent. When this occurs, the model decreases the IC_{25} estimate to a level proportional to the difference between the actual upper confidence limit and 100%. This adjustment can be quite significant and can result in much lower IC_{25}s than that actually represented by test data. The frequency of the bias will increase as the IC_{25} approaches 100% effluent.

Finally, the ICp model biases estimates of the IC_{25} to lower values when tested concentrations of effluent out-perform the response of the controls. This is a very common occurrence in chronic toxicity tests. The bias occurs due to data smoothing required to meet the monotonic non-increasing assumption of this model. This may cause, as stated by the EPA chronic methods, “a large upward adjustment in the control mean.” A larger control measurement increases the probability that a 25% difference can be found and will decrease the

concentration (indicating greater toxicity) at which a 25% difference is predicted.

Thus, the model inaccurately represents the data. If the control mean is artificially adjusted upward, it will increase the difference between the control and some of the tested concentrations, increasing the probability of finding a 25% reduction at some tested concentration. Additionally, this adjustment could possibly change the slope of the line between the two concentrations bracketing the 25% reduction. The slope of the line in this part of the concentration-response relationship plays a role in defining the IC_{25}.

It is clear that the ICp model is not robust enough to address data sets commonly experienced by labs executing tests for permittees in fulfillment of NPDES requirements. There are models available (GLM, for example) that do not adjust the control mean, do not assume linear relationships between concentrations, do not adjust results based on the relationship between confidence interval and 100% effluent, and use all of the data for all of the tested concentrations to represent the concentration response relationship. The problems with the ICp model must be addressed prior to finalizing the chronic WET methods or EPA must replace it with a defensible point estimate approach.

F. Study Report And SOP For Shipping Large-Volume Samples At Less Than 4°C.

The report prepared by Dyncorp I&ET for EPA provides information on how to prepare a WET sample for shipment prior to testing.[247] The report and SOP assumes that a target temperature range must be met when a sample reaches its testing destination for the sample and resulting test to be valid. However, EPA did not test the necessity of the temperature range that

[247] DynCorp I & ET, *Study Report and Recommended Standard Operating Procedure (SOP) for Shipping Large Volume Samples at Less Than 4°C* (September 24, 2001).

it requires in the method.[248] The record is silent regarding the use of the 4° C level. Therefore, the relevance and importance of meeting the temperature range has not been addressed; EPA merely assumes that it is necessary for tests to be valid. EPA dedicated significant resources to complete the referenced study but failed to test the principle assumption of this testing requirement.

There are some practical issues associated with EPA's recommendations that are not dealt with in the SOP. The pumping approach described by the study uses 100 feet of tubing immersed in an ice bath to cool samples to the target temperature range, but did not test the impact of the tubing or the pump on sample quality. EPA simply assumed there was no impact. The Agency also failed to provide specifications for the type of tubing or pump to use or a cleaning/conditioning procedure for this equipment prior to use. EPA also must address whether new tubing or used and properly cleaned tubing can be used. Another aspect of this approach that was not recognized in the report is that one must carry 40 pounds of ice to the sampling location to immerse the tubing. This may be difficult, given that many sampling locations are remote or are somewhat inaccessible.

The freezer approach to modifying temperature also may pose problems for facilities that are staff-limited, a common contemporary problem. This approach assumes that facilities have sufficient staff to closely monitor samples as temperatures are reduced. The study only looked at times to reach 4^{o}C and did not look at times to reach freezing. The times recorded to reach the maximum acceptable temperature for samples using a freezer are not predictable, nor are the times to freezing. It is quite possible that, when a sample does not reach the maximum allowed

[248] *See, e.g.*, Acute Methods Manual, Section 8.5.7.1.

temperature, staff would continue to work around their facility for at least another hour before checking the sample again for status. A one-hour "round" for staff to visit and monitor several process sites in a facility is common practice and will preclude more frequent monitoring of sample temperature. The Coalition understands that frozen samples cannot be used for WET testing. The sample could be frozen by the next time it is checked, invalidating the sample. Without data on time to freezing, the potential for this problem cannot be quantified and addressed appropriately.

The Coalition requests that these issues be considered before releasing this study and its recommendations in final form. Additionally, defensible temperature ranges, based on correlations between temperature and WET sample quality, for samples received by labs must be included in the methods. The current requirement is arbitrary, must be revisited, and must not be included in the final methods unless defensible data linking temperature and WET sample quality in a quantitative fashion is provided to the Coalition. The methods also must be adjusted to address a typographical error that occurs throughout the document. The methods frequently require samples to be stored at $4^{o}C$; however, the actual requirement is to hold temperatures between zero and $4^{o}C$. The methods must be changed to address this oversight.

G. Applicability Of Methods Published By Voluntary Consensus Standards Setting Organizations.

The Coalition does not support the automatic adoption or use of WET methods published by voluntary consensus standard bodies in the NPDES program. These methods do not provide the level of detail required of methods used in this context and often do not provide the DQIs and MQOs necessary to ensure the reliability of information in a regulatory context. This approach will raise the same types of questions being raised with the methods under consideration in this

rulemaking. These include whether the methods have been field and lab validated and whether the methods meet the requirements previously established for those intended to be used in a regulatory context. The Coalition recommends that EPA consider adopting test methods developed by voluntary bodies on a case-by-case basis, only after receiving and reviewing the underlying validation studies for acceptability and requesting public comment on the methods and the validation studies in a Part 136 rulemaking.

H. EPA Inappropriately Changed the Calculation Approach For The Chronic Growth Endpoints.

For the following reasons, the WET Coalition takes issue with the approach EPA has prescribed for calculating chronic growth endpoints.

1. EPA Changed the Chronic Growth Endpoint Calculation Procedure Without Inviting Public Comment.

EPA adopted a procedure in the 1995 version of the chronic methods for calculating the growth endpoint that was different than the procedure proposed for comment in 1989. The new procedure calculates growth based on the number of organisms starting a test, rather than those surviving the test (1989 approach). The procedure included in the recently proposed methods is now a biomass endpoint (even though it is termed a "growth" endpoint in the title of each method). Even though the "biomass" approach is substantially different from the 1989 proposed approach, as discussed below, EPA failed to include the change, or invite public comment, in its 1989 proposal. Instead, it proposed the chronic methods with the earlier calculation approach and then changed that approach in its 1995 final rule. EPA cannot make such substantial changes without following the procedural requirements applicable to rulemaking in the Administrative Procedure Act. When EPA changed the calculation procedure in 1995, it did not provide that time, and has not since, any data in the record to support the change.

When asked why the change was made, EPA staff responsible for the methods responded that the change was intended to standardize how such endpoints are calculated relative to other methods. EPA offered no other explanation for its decision. However, at the time the methods were promulgated, only one method (*C. dubia* reproduction) used this approach to calculate its endpoint. Therefore, EPA changed four chronic methods to accommodate a feature found in a single method.

2. EPA's Change in the Chronic Growth Endpoint Calculation Procedure Lacks Scientific Support.

Given that EPA did not present evidence in support of these changes, the Coalition concludes that the changes made in the 1995 rule were arbitrary. EPA did not even investigate the impact of the changes on test results further evidence that the changes were arbitrary. Peer-reviewed literature outlining the impacts of these changes on test results has been published since 1995. Markle et al. (2000) illustrated convincingly that the change in how the *P. promelas* growth endpoint is calculated reduces the IC_{25} in every case, thereby inferring more toxicity even though the quality of the effluent is identical. The impact of this change was not predictable with hypothesis test endpoints. The paper also observes that there was no trend in improvement in inter-test precision. Both of these observations are supported by information presented in SETAC's "Wild, Wild WET" course. Further, the SETAC course materials suggest that the MSD may be increased using the new approach. This means that the statistical sensitivity of tests using hypothesis statistics is decreased. Therefore, increased precision and sensitivity do not appear to have driven the change that EPA adopted in calculating WET chronic method growth endpoints. In fact, the changes decreased the reliability of the tests.

Decreases in test performance coupled by bias in conclusions regarding toxicity are

reasons supporting a decision not to change how the growth endpoint is calculated. EPA must change the procedure for calculating chronic test growth endpoints back to one based on the number of organisms surviving in a replicate unless convincing evidence in support of the changes can be provided and data supporting the change can be provided for each method.

I. EPA Inappropriately Changed The Fish Age Requirements On Acute Test Endpoints.

For the following reasons, the WET Coalition takes issue with the approach EPA has prescribed for fish age requirements in acute tests.

1. EPA Changed the Fish Age Requirements in Acute Tests Without Inviting Public Comment.

EPA adopted requirements in the 1995 version of the acute methods for the age of fish used in toxicity tests that were substantially different than those proposed for comment in 1989. EPA changed acute tests to require the use of younger fish in the 1995 methods without proposing to do so or inviting public comment in its 1989 proposed rule. Because the change was substantial, as discussed below, EPA failed to provide the public participation opportunity required by the APA per rulemaking. When EPA changed the fish age requirement in 1995, it did not provide any data in the record at that time, and has not since then, to support the changes. When asked why the changes were made, EPA staff responsible for the methods claimed that the changes were intended to standardize this variable between tests. EPA offered no other explanation for its decisions.

2. EPA's Change in Fish Age Lacks Scientific Support.

Given that EPA did not adopt criteria to support these changes and their reasoning for the changes, the Coalition concludes that the changes made in the 1995 rule were arbitrary. EPA did

not even investigate the impact of the changes on test results – further evidence that the changes were arbitrary. Peer-reviewed literature outlining the impacts of these changes on test results has been published since 1995. Markle et al. (2000) illustrated convincingly that younger *P. promelas* used in acute tests were more variable in their response, thereby decreasing the reliability of data and decreasing statistical power in hypothesis tests. Tests of 1-14 day-old fish showed that 1 day-old fish were the least sensitive. Tests of 14-90 day-old fish showed that 14 day-old fish were the least sensitive. Therefore, even if EPA made the change with hopes that the tests would be more reliable and sensitive, the data do not support this. Decreases in test performance coupled with bias in conclusions regarding toxicity support a decision not to change the age requirement for the acute fish tests. The Coalition's position is that EPA must change the methods' requirements for fish age in the acute tests back to that proposed in 1989 unless defensible criteria for the changes can be provided and data meeting the criteria can be provided for each method.

J. Methods Are Not Clear That All Endpoints Are Required Of Each Test.

The chronic WET methods proposed often include multiple biological endpoints, including survival, reproduction and/or growth. However, the methods do not communicate to the user whether all are required to be evaluated when conducting tests. Regulatory agencies routinely assume all endpoints are required, simply because the endpoints are listed as part of each test's title. The intent would be clear if EPA were proposing to use these WET tests as water quality criteria. However, EPA states in the preamble to the rule:

> EPA's promulgation of WET test procedures are not water quality criteria recommendations under section 304(a). When States develop and implement water quality standards, including narrative water quality criteria, States should translate those criteria into measurable expressions of toxicity. The test methods themselves

> are not per se translators of the narrative criterion: 'no toxics in toxics amounts.' The test methods are merely the measurement tools according to which such criteria may be translated.[249]

Since the methods are not translators of narrative criteria, EPA has left the responsibility of deriving and adopting translators with each regulatory agency overseeing its respective permit programs. These regulatory agencies must choose specific tools to translate narrative criteria. There is no requirement, therefore, for specific endpoints (biological or statistical) to be calculated unless a permit or regulatory agency requires them in regulation. EPA has approved numerous permits in Region VI that do not include sublethal endpoints in the regulatory process.

The test methods must state clearly that each biological and statistical endpoint listed for that method is a different protocol, and that it does not intend for all of those endpoints to apply when a test method is included in a permit or regulation. The specific choice of endpoint must be made by the regulator in a manner that is consistent with the water quality standards it is responsible for implementing. It is the Coalition's understanding that promulgation of methods with multiple biological or statistical endpoints does not mandate that all of these endpoints must be calculated and reported even if the respective methods are identified generally in permits or regulation. Each method must contain language clarifying this fact.

K. Requirement To Measure Chlorine After Sampling.

Section 8.5.3 of the methods requires that total residual chlorine must be measured immediately following sample collection if the effluent has been chlorinated. This requirement is unnecessary for two reasons. First, chlorine will likely not be present in effluent if that facility also has dechlorination processes in place. Even though the wastewater is chlorinated, it will not

[249] 66 Fed. Reg. at 49,796.

have measurable concentrations of chlorine. Second, the time period between testing and sample collection can be up to 36 hours. Even with refrigeration, the chlorine contact of a sample will diminish over time. Therefore, the concentration measured after sampling will have no relationship with that upon testing. Additionally, the methods do not indicate what the purpose of the measurement is and what should be done if chlorine is measured. It would seem defensible to require chlorine measurement upon testing, when dechlorination at the discharge site does not take place, to determine if chlorine will impact test results. The requirement currently in the methods must be deleted or changed to apply only when the sample is being prepared for testing and when the effluent being tested has not been dechlorinated at the discharge site.

L. Intralaboratory And Interlaboratory CVs In Methods.

EPA is including intralaboratory, matrix-specific CVs in the methods proposed in the Federal Register notice. The intralaboratory CVs obtained for each matrix are almost entirely based on the analysis of only two samples per matrix per laboratory. EPA states in its WET Variability Guidance (Appendix B) that:

> Methods in Tables B-1 through B-3 that are represented by fewer than three laboratories or fewer than 20 tests are not shown in Tables B-7 and B-8, because characterizing method variability using so few tests and laboratories would be inadvisable.

Based on the criterion, many of the intralaboratory CVs presented would be discarded. This criterion aside, CVs based on only two results for a matrix and laboratory cannot be regarded as representative. Sample sizes per matrix and laboratory must be larger for these numbers to be reliable enough to include in the promulgated methods. Review of the Variability Guidance suggests that, on average, the fewest number of tests per laboratory and reference

toxicant used to characterize within lab variability was seven. This indicates that the intralaboratory CVs presented in the proposed methods are not defensible and cannot be used to represent intralaboratory variability. These CVs must be removed from the methods prior to promulgation of the methods or replaced with CVs based on data sets meeting MQOs selected to support the intended use of those CVs.

EPA also combined CVs across all sample types to calculate average CVs for each test endpoint in Table 1 of both chronic manuals. There is no reference in the methods to whether EPA first confirmed that the CVs between sample types were similar before they were pooled. The CVs presented in these tables may not fairly represent the CVs to be expected for individual sample types. EPA must check and confirm this assumption as valid before using these values in a regulatory context (promulgation).

M. Methods Do Not Address Variability And Uncertainty In Point Estimates.

EPA's proposed methods provide some guidance, albeit insufficient given the intended use of WET data generated using these methods, regarding variability and uncertainty on a test-by-test basis when hypothesis test endpoints are used. This guidance includes application of PMSD percentiles when interpreting NOECs and NOAECs. However, EPA has entirely ignored the issue of uncertainty regarding the interpretation of point estimates for individual tests. Point estimates also suffer from variability, resulting in uncertainty usually represented by confidence limits. Confidence limits are also a function of which statistical approach is used. The level of uncertainty at a particular point estimate can cloud and preclude interpretations of data relative to pre-selected benchmarks (permit limits). For example, one cannot definitively conclude that an IC_{25} of 85% with confidence limits of 60%-110% is unacceptable if the goal is an $IC_{25} > 100\%$ effluent. There is a 95% probability that the real IC_{25} for this sample is in the 60%-110% range,

but that is the limit to the conclusions that can be drawn from the statistical analysis. EPA has not formally recognized this uncertainty in the methods, but it is clear that this is a significant deficiency in the proposed methods. As they have done for hypothesis tests, EPA must include procedures in the methods to determine when the uncertainty in point estimates is unacceptable and when differences from benchmarks (IWC) cannot be discerned due to variability. The proposed methods are not acceptable unless uncertainty relative to point estimates is adequately addressed at promulgation.

EPA also has provided language intended to address situations where point estimate techniques do not provide reliable confidence limits. Rather than providing direction on how to calculate reliable limits under these circumstances, EPA merely states that this happens and abandons the issue. That is not sufficient. EPA must provide point estimate techniques that will calculate defensible confidence limits that are required to assess the reliability of endpoints and the uncertainty associated with decisions using those estimates. This deficiency is due to the fact that EPA does not have appropriate criteria (DQIs and MQOs) for selecting statistical approaches that provide defensible results. The Coalition is opposed to the inclusion of procedures in the proposed WET methods that cannot provide reliable and defensible confidence limits needed to ensure that data is of acceptable quality for its intended use.

N. Requirement For A Specific Dilution Series.

Sections 8.10.3 of the chronic and 9.3.3 of the existing acute methods state that definitive tests use dilutions that bracket the receiving water concentration and concentrations are selected based on very specific equations. Although the Coalition agrees that it is preferable to bracket the receiving water concentration with test concentrations whenever possible, the language of these sections contradicts the language of the newly proposed methods:

> ... test concentrations should be selected independently for each test based on the objective of the study, the expected range of toxicity, the receiving water concentration, and any available historical testing information on the effluent.[250]

The sections of the proposed methods cited do not use the word "must" when referring to which dilutions are required, but words such as "recommended" or "should" are also not used. Regulatory agencies have interpreted this language as a requirement. In light of the modifications proposed by EPA, the Coalition believes that these sections must express the language quoted above from the proposed rule's preamble. Otherwise, there is a conflict between the new language (section 9.3 of acute methods and section 8.10 of the chronic methods) and the pre-existing language.

Sections 8.10.2 of the chronic methods and 9.3.2 of the acute methods also appear to conflict with the newly proposed language by inferring that a "geometric series" be approximated when selecting dilutions. The term "geometric" is very vague but has been interpreted by regulatory agencies in very specific ways. This prevents users of the methods from conducting the tests in a way supported by the new language. Given the newly proposed language, there is a conflict within the methods on this issue. Additionally, EPA's Supplementary Information Document on WET methods (1995) states that:

> none of the statistical methods recommended by EPA for the analysis of toxicity test data require the use of log or geometric dilution series in the toxicity tests.

Given the intent of the new language to support selection of test dilutions different from examples currently presented in the methods, sections 8.10.2 of the chronic methods and 9.3.2 of the acute methods must be modified to remove reference to "geometric" series.

[250] Proposed Method Manuals Changes at 75.

The Coalition supports inclusion of the new language on selection of dilutions for WET testing only with the addition of text specifying that the dilutions selected must support calculation of defensible confidence intervals. This requirement must be mandatory for point estimate test results to be reliable.

O. Intratest Outlier Management.

EPA's "Clarifications Regarding Flexibility in 40 C.F.R. Part 136 Whole Effluent Toxicity (WET) Test Methods" memo, dated April 10, 1996, discusses replicates as outliers in the *Champia* test:

> As with any toxicity test the analyst may have to determine the single 'odd' replicate is an outlier. If the overall mean number of cystocarps is at least 10, with the low or high 'odd' replicate excluded, then whether or not a single outlier is present does not effect the determination of control acceptability.

EPA's methods provide limited text on how to identify and treat intratest outliers. The acute methods do not make any specific recommendations on a method to test for outliers, and none of the methods provide any instruction regarding what to do if an outlier is identified other than to conduct analyses with and without outliers and report both results. The methods do not address how to determine which result should be reported in fulfillment of permit requirements. EPA's memo of 1996 does provide guidance on how to respond to identification of a test outlier, and this must be included in the *Champia* method if it is promulgated following this comment period (although the Coalition does not support the promulgation of this method). It would seem that the logic applied to the *Champia* method should apply to other methods when the number of original replicates per treatment is four or more (because four replicates are used in the *Champia* test and EPA agreed that one replicate could be dropped under these circumstances).

The reliability of the proposed methods assumes that replicate response is representative of the population of responses for a treatment or control. Positively identified outliers contradict this assumption and must be removed to restore representativeness (a data quality attribute). EPA must include specific procedures in all of the methods that identify outliers and support their removal prior to completing all statistical analyses required of the method.

P. pH Control.

The WET Coalition appreciates EPA's willingness to address the issue of pH "drift." For the reasons that follow, however, the Agency has not yet adequately resolved the problem.

pH has been documented numerous times in peer-reviewed literature to impact the toxic potential of chemicals to aquatic life. The best example of this is ammonia, probably the most ubiquitous toxicant identified in effluent samples. As pH increases, the most toxic fraction of ammonia (unionized ammonia) increases. Therefore, if pH in a test does not represent the pH of an effluent under certain conditions (IWC) and the effluent contains significant (> 5ppm, total) ammonia, it is probable that the unionized ammonia concentration in the test is greater than that experienced by organisms instream. pH can increase in test replicates due to exchange of carbon dioxide in the effluent with the atmosphere or due to sample manipulations like those required to increase the salinity of samples prior to testing using estuarine/oceanic organisms. As magnitude of exposure increases, response (mortality, decrease in reproduction/growth) increases. At some point, the stress due to this artifact becomes large enough in a test concentration to result in a measurable difference from controls either through combination with other stressors or alone. The only way to address this artifact is to control pH in the test.

Numerous stakeholders commented on this concern when these methods were first

proposed for promulgation in 1989. The methods did not address the need for pH control or provide procedures on conducting pH control when adopted in 1995, despite the fact that numerous comments were made and EPA acknowledged the comments in their "Supplementary Information Document On WET Methods".

EPA has proposed new language in the methods to address pH during WET testing. The Coalition has significant concerns regarding the proposed changes. First, the acute methods provide no option for pH control. This is unacceptable; the probability that pH drift or shift due to sample manipulation required of the methods will influence test results is significant and cannot be ignored in the methods. The decision to use pH control should not be based on whether acute or chronic tests are run or which species is tested. The issue is whether pH is influencing the results of tests relative to their intended purpose. Any test that experiences sufficient pH drift or shift to influence a test result must include pH control. EPA must include provisions for pH control in the acute methods.

EPA states that a regulator authority "may" allow for pH control during tests. However, if the permittee has sufficiently illustrated that pH drift or shift has influenced test results, the authority "must" allow control of pH in tests unless it conflicts with the objective of the test. If the purpose is to estimate the impact of effluent instream, then there is no doubt that pH control must take place for tests to be reliable indicators of instream impact. The word "may" in this context must be changed to "must".

Control of pH must not be limited to only specific methods. For example, the impact of pH drift or shift on algal tests or tests with other species can be significant and influence the results of the tests. Language must be provided in the methods supporting pH control when the

permittee has shown through parallel testing that pH drift or shift influences test results within the context of their use. Furthermore, language must be provided to support other defensible approaches to pH control, like the use of organic buffers.

EPA provides guidelines in the methods for when it believes artifactual toxicity is due to pH drift or shift. These include when pH change is large (more than one pH unit) and/or the concentration of the pH-dependent toxicant is near its threshold for toxicity. These guidelines do not recognize the fact that pH changes can be subtle enough to contribute to the total toxic response observed in a test but may not be the sole reason for toxic response. Therefore, pH drift or shift may influence test results even though there may be smaller changes in pH or concentrations of pH-sensitive toxicants are not near their thresholds for toxicity. The ability of pH drift or shift to contribute to toxicity, rather than be the sole reason for toxicity, must be communicated clearly in the methods.

The methods state that the pH should be maintained at the pH of the receiving water when tests are designed to address toxicity of effluent instream. This statement is in error and may introduce uncertainty in test results. The target pH in this scenario must be the pH at the edge of the appropriate mixing zone. If chronic tests are conducted, the pH instream after chronic dilution has taken place must be used rather than using any pH found in the stream. Only when no dilution is allowed should the pH of the receiving water be used as a target pH. The same comment applies to acute tests. The language must be changed to "… pH must be maintained at the pH of the respective IWC."

The proposal states that when the objective of the test is to estimate end-of-pipe toxicity, the pH should be maintained at the pH of the sample after warming to test temperature. This

statement is in error and will introduce greater uncertainty in test results. The pH must be maintained at the pH of the sample upon completion of collection. The pH of a sample can be affected by both refrigeration and holding time, which take place between sampling and testing. The most representative pH in this case is the one measured upon completion of sampling. The language must be changed to "… pH should be maintained at the pH of the sample upon completion of sample collection."

The text of the methods requires that pH drift in the uncontrolled test be substantially greater than in the controlled test. This requirement misses the point of conducting side-by-side tests and introduces unnecessary confusion when interpreting the test results. It is unclear what "substantially" means and how it will be interpreted between permittees, labs and regulators. This requirement is not necessary because the defining criterion of whether pH control is necessary is whether toxicity is thereby reduced. This criterion is already found in the methods and is sufficient to delineate when pH control must be allowed. Including both requirements (pH control reduces toxicity and pH drift must be greater in the uncontrolled test) in the methods can actually cause confusion when they conflict. For example, it is not clear whether pH control must be used when pH drift in uncontrolled tests is not substantially greater than in controlled tests but pH control reduces toxicity. Additionally, the proposed language states that drift in uncontrolled tests "must" be at least twice that observed in controlled tests. EPA has not provided any data to support this and, again, this language provides opportunity for conflict with the requirement that pH control reduce toxicity. The word "substantially" must be removed from the methods as it relates to pH control, and the sentences requiring pH drift in uncontrolled tests to be at least twice that of controlled tests also must be removed from the methods.

The new language of the methods provides regulatory agencies with the authority to

request more information or additional testing before pH control is allowed. The methods must be more definitive on this topic; otherwise, regulatory agencies will use this text to justify unreasonable requests of permittees and labs. EPA must include in the methods the specific type and quantity of information required to justify pH control in tests.

EPA states in the methods that the daily cycle of upward pH drift and renewal may, in rare circumstances, cause artifactual toxicity even in the absence of pH-sensitive toxicants. EPA did not provide any data in the proposed rule to support the statement that this is a rare occurrence. EPA must either provide the data to support this statement or remove "In rare circumstances" from the method text.

The text of the proposed changes seem to use a 5 ppm total ammonia concentration as a benchmark above which toxicity can be expected with pH drift or shift. EPA should clarify that concentrations greater than 5 ppm are not necessarily toxic instream because this depends on site-specific factors, such as dilution and receiving water chemical characteristics. The 5 ppm total ammonia benchmark can only be used relative to toxicity in the test vessel rather than toxicity instream.

Q. Method Changes To Address Pathogen Interference.

EPA has proposed new language for the chronic *P. promelas* test addressing situations where pathogens outside of the discharger's control are influencing the results of effluent tests. There are a number of issues that must be addressed within the text of the method before these changes can be adopted in a promulgated method. First, regardless of the treatment, there is a significant probability that tests will be impacted to varying degrees by pathogens when they are present. The methods presented are designed to reduce the level of impact, but they cannot

preclude it. The method must provide a procedure for accounting for impact due to pathogens even when the new language of the method is followed.

There is concern that a change in test design from four replicates of ten organisms each to twenty "sub-replicates" with two organisms each, to address pathogen interferences, will modify the variability represented by a treatment. This change increases the probability that the exposure conditions of two organisms in one sub-replicate are different than that of another sub-replicate and that sub-replicates will manifest different responses accordingly. Therefore, sub-replicates become replicates. The purpose of replicates is to represent the population of responses for a testing environment independent of the test variable (effluent concentration). Greater range in response per treatment increases within treatment variability and the PMSD for the entire test. Both this issue and the actual impact of pathogens on individual fish may increase the PMSD for tests, which directly impacts the ability of tests to predict potential for toxicity. This also means that the PMSD limits proposed for this test will not accurately reflect variability if a pathogen effect occurs. EPA must provide data showing that PMSD will not change with the new testing approach and provide new PMSD limits in the method prior to promulgation if PMSD does change.

It also does not appear that the method to control pathogen interference has been adequately validated. The modifications suggested to the method were developed in a relatively small scale study (not the interlaboratory study conducted by EPA) and have not been tested in labs across the country. There is also no provision for alternative methods. The Coalition holds that the method must provide language supporting the use of defensible alternatives to those being proposed by EPA and that the proposed procedures are not required prior to testing of pretreated samples. The effectiveness of EPA's proposed changes, in terms of addressing this

issue for NPDES permits nationwide is unknown and must be confirmed before the changes can be adopted in the methods.

R. Method Changes On Dilution Waters.

The proposed methods discuss issues to be considered when selecting a dilution water for WET testing. Generally speaking, the Coalition supports the intent of this new language. However, the Coalition believes the EPA guidance document that EPA proposed to incorporate, titled "Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing" (July 2000) should be more flexible. This guidance infers that receiving waters should be used for dilution water if the objective of the test is to determine the toxicity of the effluent in the receiving system. The Coalition supports this concept, but there can be logistical issues associated with meeting this goal that would ultimately impact the reliability of data generated in this fashion. For example, collection and pre-treatment of receiving waters from estuarine or oceanic sites can be costly and introduce organisms that are not removed through filtration but impact the results of the tests by stressing the target test organisms. The text of the guidance should only make recommendations on approach; the current text infers that you must follow Figure 6.1 for selection of dilution water to be appropriate.

S. Underestimation Of Within Laboratory Test Variability Using Point Estimates In Acute Tests.

The current statistical approaches proposed in the methods for calculating point estimates for acute tests underestimate within laboratory test variability because the individual response of replicates are averaged, in all cases, prior to their use in each of the approaches. This is not the case for the NOAEC determination or any of the chronic test statistical approaches. Averaging of replicate responses prior to analysis removes a component of variability; that between

replicates. Confidence intervals for the point estimate, therefore, will underestimate the uncertainty in that estimate. To address this bias, the acute methods being proposed must either adjust the current procedures to require entry of individual replicate data prior to analysis or replace the current procedures with one or more that are appropriate for acute test designs and endpoints (50% impact for survival, for example) yet allow the entry and use of individual replicate responses.

T. The Primary Objective of NPDES WET Testing.

The proposed methods state that the primary objective of NPDES permit-related toxicity testing is to assess the toxicity of an effluent, independent of any interactions with receiving waters (acute and chronic methods, section 7.1.1). There is a contradiction between this objective and the intended use of WET tests in the NPDES program. WET tests are conducted in this program to determine reasonable potential to exceed water quality standards for toxics or to determine compliance with toxicity limits that are derived from water quality standards. Since water quality standards only apply to ambient water quality, rather than the quality of effluent within a discharge pipe, WET tests conducted in this context, therefore must address ambient receiving water conditions rather than within-pipe conditions. This is a fundamental point that is often forgotten or ignored when implementing WET tests in the NPDES program. Because of this disconnect, regulators believe that an effluent, which is analyzed using the proposed WET methods, will definitively predict toxicity or the lack of toxicity instream. This will not be the case if the factors differing between field and lab exposure are not identified and reviewed for their potential to impact the result of the test. This goes to the heart of the question of whether the WET tests, as currently designed and proposed, will meet the goal that has been established through the use of WET tests – to predict the response of aquatic organisms instream. The

Coalition does not support the language currently found in section 7.1.1 of the methods and holds that it must be adjusted to reflect why and how WET tests are used in the NPDES program.

U. Changes To *S. capricornutum* Test.

EPA proposes to recommend that this test be conducted with EDTA but allows for use of the test without it under particular circumstances. Based on the results of the Interlaboratory Study, this proposal is unacceptable. The completion rates for either set of tests were less than 66%, strongly supporting the conclusion that these tests are not appropriate for use in a regulatory program. EPA claims that these completion rates are due to lab inexperience, but EPA also failed to acknowledge that it prequalified all labs participating in the Interlaboratory Study to make sure that they had acceptable levels of experience to represent lab performance using this method. EPA also states in the preamble:

> Interlaboratory variability of the *Selenastrum capricornutum* Growth Test method was much lower with EDTA (34.3%) than without EDTA (58.5%). When conducted with EDTA, the *Selenastrum capricornutum* Growth Test method exhibited interlaboratory precision similar to other chronic methods evaluated in the WET Variability Study.[251]

Two conclusions can be drawn from these statements. EPA believes that there is a significant difference between a CV of 34.3% and 58.5%, and it believes` that only the CV of 34.3% is similar to that of other chronic methods. Therefore, a CV of 58.5% is different than that of other chronic methods. EPA also claims that the variability of WET methods is similar to that of chemical specific methods. Given that EPA's only criterion for accepting WET methods in 40 C.F.R. Part 136 is the comparability of CVs between these methods and chemical specific methods, one can only conclude a CV of 58.5% is not comparable to that of chemical specific

[251] 66 Fed. Reg. at 49,807 (col. 3).

methods and this method is not acceptable for use in the NPDES program. Additionally, this version of the test resulted in a false positive rate (33.3%) at least one order of magnitude greater than all other methods studied. The *S. capricornutum* test both with and without EDTA clearly are unreliable and must be removed entirely from the options for promulgated methods.

V. *M. bahia* Fecundity Endpoint.

EPA states in the proposed methods:

> While the fecundity endpoint is an optional endpoint, it is often the most sensitive measure of toxicity, and the 7-day mysid test estimates the chronic toxicity of effluents most effectively when all three endpoints (survival, growth, and fecundity) are measured (Lussier et al., 1999).[252]

This conclusion is based on a review of only 22 tests (19% of a total of 115 tests assembled for the Lussier et al. paper). The tests that were not used for this analysis (81% of total) were eliminated based, in part, on failure to meet the TAC of the method and failure to show toxicity. However, these were not the only criteria used to eliminate tests. The paper cited does not state clearly why all of the data collected for the study were not used to perform the analysis that EPA is using to support the position espoused in the methods. The small size of the data set used by the authors to support EPA's conclusions regarding effective identification of toxicity with all three endpoints and the sensitivity of the fecundity endpoint significantly diminishes the reliability and defensibility of EPA's quoted language. Furthermore, bias may have been introduced in the analysis conducted by Lussier et al. when tests were removed without justification. The value of this data relative to a national effort such as the NPDES WET program is also in question since it appears that all of this data was generated in one lab (EPA's

[252] Proposed Method Manuals Changes at 82.

lab), overseen by the same person, and may have never been used to meet NPDES requirements. Coincidentally, the first author of this paper is the same person that developed the methods. This data set cannot be used to represent the performance of the method or the fecundity endpoint, across multiple labs, as it relates to use in the NPDES program.

The quote above must be removed from the method given: (1) the paucity of data used by EPA to support the quote, (2) the questionable methods used to limit this data set, and (3) that the data set used to support the conclusions is limited to one lab, rather than multiple labs, overseen by the same person responsible for developing the test method. EPA must provide sufficient supporting data collected in a statistically defensible fashion that represents labs actually producing data in fulfillment of NPDES requirements in order to include this language in the method.

The reliability of this endpoint was also questioned by WET interlaboratory study peer reviewer Z:

> Measuring a successful fecundity endpoint in only 50% of the tests is extremely poor and raises serious questions about the mysid chronic test method.[253]

and

> … the results seems to show that some of these tests should not be used in the regulatory context because the successful completion rate is too low and CV values are too high.[254]

Given the above, EPA must change the language to clearly state that the fecundity endpoint is not authorized under Part 136.

[253] Peer Review Report at 68.

[254] Peer Review Report at 19.

W. Blocking By Known Parentage.

The US EPA has determined that the number of offspring produced by a test organism is significantly affected by parental source, and parent organisms are known to produce males when under conditions of environmental stress. The proposed rule states that if >50% of the surviving *Ceriodaphnia* from a given "block" (replicates from the same parentage) are identified as males, the entire block is to be excluded from any reproduction analyses. The proposed rule also states that if <50% of the surviving *Ceriodaphnia* from a given "block" are identified as male, only the identified males should be excluded from subsequent reproduction analysis. This approach could bias test results because very different approaches are used when the difference in percentage of males could be small. The decision to base the approach on whether more or less than 50% of the adults are males also appears to be arbitrary. For these reasons, the most defensible and reliable approach would be to exclude any block from reproduction analyses that contains males (due to environmental stress of the parent organism). This change should not result in rejection of a large number of tests if EPA is correct that only 7% of all *C. dubia* reproduction tests included any males.

X. Nominal Error Rates.

The Coalition has a number of concerns regarding the language proposed to address the use of nominal error rates when calculating WET test statistical endpoints. First, EPA is allowing the unconditional use of the 0.01 error rate for only the sublethal chronic *C. dubia* and *P. promelas* endpoints. EPA did not provide its rationale for limiting the use of the 0.01 error rate to just those endpoints. The lower error rate must apply to all tests requiring the use of an error rate.

The Coalition has a number of concerns regarding the language proposed to address the

use of nominal error rates when calculating WET test statistical endpoints. First, EPA is allowing the unconditional use of the 0.01 error rate for only the sublethal chronic *C. dubia* and *P. promelas* test endpoints. For the other chronic freshwater tests, and for the chronic marine tests, the 0.01 error rate can only be used if the WET limits are derived without allowing for receiving water dilution. EPA does not explain why the availability of receiving water dilution is relevant to deciding when the 0.01 error rate is appropriate. The reason for allowing 0.01 rather than 0.05 is to increase confidence in the test results. Even where the WET effluent limit was derived with the benefit of dilution, the need for higher confidence in the test result remains. The main difference, in this context, between permit limits derived with or without dilution is that the user typically can only confirm the presence or absence of a dose-response relationship for the former. But just because a valid dose-response relationship exists does not mean the sample result is valid in other respects. For example, a valid dose-response relationship may be ascertained for a test that failed the TAC, yet EPA would not consider that test to be reliable. Permittees need the 0.01 error rate to improve confidence in the test result, and EPA must change the proposed language to delete the reference to receiving water dilution.

The second concern over the proposed language is that it appears to allow the use of the 0.01 error rate only for testing related to compliance with NPDES permit limitations. WET testing, however, is routinely conducted for other important regulatory purposes. For example, WET testing is performed to determine whether or not the discharge has the "reasonable potential" necessary to trigger the need for NPDES permit limitations. It is essential that EPA clarify that the 0.01 error rate can be used regardless of the purpose of the WET testing.

EPA specifies that the lower error rate can be used only if the conditions in its “nominal error rate” guidance are satisfied. In effect, therefore, EPA is proposing to convert that guidance

into a binding rule. As such, EPA must provide for public comment its technical justification for the conditions it imposes (e.g., to demonstrate adequate sensitivity). The guidance document itself does not offer that rationale. The WET Coalition takes issue with some of those requirements but cannot comment meaningfully without knowing the technical basis for EPA's guidance. EPA does not specify that permittees can use options other than the one in the guidance to show test sensitivity.

By requiring use of its error rate guidance, EPA is proposing to mandate that a particular standard for power be met before an alpha of 0.01 can be used. This is problematic for several reasons. As stated above, EPA has offered no rationale for its power standard requirements and, thus, has deprived the public of the opportunity to comment on that rationale. *See* the Power Analysis Attachment for details on the WET Coalition's concerns. EPA neither requested comment in its proposed rulemaking on adoption of a power standard for toxicity tests nor stated that a power standard was being adopted by the Agency in the methods.

Finally, by requiring that the 90^{th} percentile PMSD be met when a 0.01 alpha is used, EPA is requiring less intratest variability, again defining a standard for performance without first providing its rationale for public comment. EPA must explain formally its rationale and request comments on adoption of standards for intratest variability before it can require use of those standards in the methods.

Y. EPA Has Not Yet Responded To Certain Comments Submitted During The Initial Proposal Of The WET Test Methods.

To ratify the WET test methods, as EPA is proposing, the Agency must address all of the issues that were relevant to its original decision to approve the methods for inclusion in Part 136. In that regard, several commenters on the 1989 proposal raised significant issues to which EPA

did not respond when it issued the contested 1995 rule (e.g., WET detection levels,[255] characterization of intratest variability,[256] pH limits, [257] and standards for intralaboratory variability[258]). Those issues remain as valid today as when they were submitted. For purposes of determining whether to ratify the methods, EPA will need to consider those issues and explain how they were taken into account.

Z. EPA Should Avoid Bias In Presenting Conclusions From The Participating Laboratory Meeting On January 8, 2002.

On January 8, 2002, EPA held a meeting in Chicago reportedly to seek feedback on its WET method proposed rule. The WET Coalition understands that the Agency intends to include in the official rulemaking record a summary of the consensus positions reached by the participants that attended that meeting. To the extent the Agency depends on that event in developing its final rule, it should explicitly recognize the "special interests" associated with the great majority of the participants at that meeting.

In short, EPA's meeting invitation, which was dated December 11, 2001, was addressed exclusively to the laboratories that participated in the Interlaboratory Study. Those laboratories, notwithstanding the scientific integrity they can be expected to exhibit in the performance of their future analytical work for government and stakeholders, cannot be viewed as unbiased for purposes of commenting on the WET proposal. To ensure that the consensus positions arising from the EPA/Participating Laboratory meeting are accorded appropriate weight in the

[255] *See* 304(h) WET, I-B.30.

[256] *See* 304(h) WET, I-B.7.

[257] *See* 304(h) WET, I-B.13.

[258] *See* 304(h) WET, I-B.79.

decisionmaking process, EPA should make clear in the record the following objective facts. First, the livelihood of the participating laboratories depends in large measure on WET testing, and any rulemaking outcome that would restrict the use of those test methods in the regulatory process could subject them to a very substantial adverse economic effect. Second, as discussed in the comments above, a very large majority of the participating laboratories were unable to perform the WET tests without deviating from (1) the mandatory procedures required by the test protocols and EPA's Interlaboratory Study DQOs, and (2) the acceptable procedures for calculating and reporting test results.

The WET Coalition believes that those deviations confirm the lack of ruggedness for many of the proposed WET tests, and contributed to the inadequacy of the database used for drawing conclusions from the Interlaboratory Study.

A summary of the meeting prepared by a member of the WET Coalition that attended is included as an attachment to these comments.[259] EPA should review that summary as part of its decision process.

V. CONCLUSION

As discussed above, WET tests can play a useful role in protecting the environment, but they are insufficiently reliable for use in making NPDES permitting or compliance decisions. For example, WET tests suffer from numerous shortcomings, such as: (1) the lack of adequate validation studies to evaluate their performance and to establish their reliability; (2) their unacceptable performance, based on the validation data that are available; (3) the lack of

[259] Pletl, James, Memorandum re: January 8, 2002 EPA WET Testing Lab Stakeholder Meeting in Chicago, IL (January 9, 2002).

mandatory QA/QC and other conditions in the test protocols as necessary to ensure that test results will be consistent and reliable, regardless of the qualified WET laboratory performing the test; and (4) the lack of evidence showing that the WET test results are capable of reliably measuring the instream "effects" EPA claims those tests can achieve.

The WET Coalition urges the Agency to withdraw the WET test methods from Part 136 until those deficiencies can be remedied. In the event EPA instead decides to ratify the test methods in Part 136, it should do so with an explicit statement that WET methods are not being approved for use in setting or determining compliance with NPDES limitations. EPA must be clear that WET methods can only be applied for purposes, like monitoring, which do not subject dischargers to liability.

REFERENCES

ASTM D2777, *Standard Practice for Determination of Precision and Bias of Applicable Test Methods of Committee D-19 on Water*, § 7.2.3 (1998).

Ausley, L. 2000. Reflection on whole effluent toxicity: the Pellston Workshops. *Environ. Toxicol. Chem.* 19:1-2.

Brent, R., et al. Accuracy of Laboratory Reporting in EPA's WET Interlaboratory Variability Study. Abstract for 2001 SETAC Conference in Baltimore, MD.

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

Chapman, P.M. 1995. Extrapolating laboratory toxicity results to the field. *Environ. Toxicol. Chem.* 14:927-930.

Cruze, R. 1993. Effects of pH variation on chronic toxicity test reliability. Riverside Regional Water Quality Control Plant, Riverside, CA. **WET-IX, C.35**

Davies, Tudor T., U.S. EPA Office of Water, Memorandum to EPA Regional Water Management and Environmental Services Division Directors, *Clarifications Regarding Flexibility in 40 CFR Part 136 Whole Effluent Toxicity (WET) Test Methods* (April 10, 1996). **WET-IX, B.24**

Davies, Tudor T. and Michael B. Cook, U.S. EPA Office of Water, Memorandum to EPA Regional Water Management and Environmental Services Division Directors, *Clarifications Regarding Whole Effluent Toxicity Test Methods Recently Published at 40 CFR Part 136 and Guidance on Implementation of Whole Effluent Toxicity in Permits* (July 21, 1997).

Dhaliwal, B.S., R.J. Dolan, and R.W. Smith. 1995. A proposed method for improving whole effluent toxicity data interpretation in regulatory compliance. *Water Environ. Res.* 67:953-963.

Diamond, J. and C. Daley. 2000. What is the relationship between whole effluent toxicity and instream biological condition? *Environ. Toxicol. Chem.* 19:158-168.

Downey, P.J., et al. 2000. Sporadic mortality in chronic toxicity tests using *pimephales promelas* (Rafinesque): cases of characterization and control. *Environ. Toxicol. Chem.* 19: 248-255. **WET-IX, B.13**

DynCorp I&ET, *Study Report and Recommended Standard Operating Procedures (SOP) for Shipping Large-Volume Samples at Less than 4°C* (September 24, 2001). **WET-IX, B.5**

Gebhart, J.E., J.D. Messman, and G.F. Wallace, *Interlaboratory Evaluation of SW-846 Methods 7470 and 7471 for the Determination of Mercury in Environmental Samples*, U.S. EPA Environmental Monitoring Systems Laboratory, EPA/600/4-88/011 (April 1988). **Attachment to 304(h) WET, I-B.65**

Goodfellow, W.L., et al. 2000. Major ion toxicity in effluents: a review with permitting recommendations. *Environ. Toxicol. Chem.* 19:175-182. **WET-VII, B.1**

Grothe, D.R. and D.E. Johnson. 1996. Bacterial interference in whole-effluent toxicity tests. *Environ. Toxicol. Chem.* 15:761-764. **WET-IX, C.19**

Grothe, D.R., et al. (eds.). 1996. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts. SETAC Press, Pensacola, FL. **WET-VII, B.9**

Hall, Jr., L.W. and J.M. Giddings. 2000. The need for multiple lines of evidence for predicting site-specific ecological effects. *Hum. Ecol. Risk Assess.* 6:679-710.

Hunt, J.W., et al. 1997. Precision and sensitivity of a seven-day growth survival toxicity test using the west coast marine crustacean *Holmesimysis costata*. *Environ. Toxicol. Chem.* 16:824-834. **WET-IX, B.23**

In re Continental Oil Company, et al., EPA General Counsel Opinion No. 72 (October 20, 1978).

In the Matter of: City of Salisbury, Maryland, EPA Administrative Law Judge Division, Docket No. CWA-III-219 (February 8, 2000).

In the Matter of: Metropolitan Dade County, Miami-Dade Water and Sewer Authority (NPDES Permit No. FL0024805), EPA Administrative Law Judge Division, 1996 EPA ALJ LEXIS 80 (October 3, 1996).

In the Matter of: Robert E. Schloesser Respondent Determination, EPA Case No. 91-0134-00, 1992 WL 937653 (November 25, 1992).

Koorse, Steven J. (on behalf of WET Coalition), Letter to Geoffrey H. Grubbs, EPA Office of Water, re: WET Test Rulemaking (September 18, 2001).

Koorse, Steven J. (on behalf of WET Coalition), Letter to Geoffrey H. Grubbs, EPA Office of Water, re: Whole Effluent Toxicity Program (July 16, 2001).

Koorse, Steven J. (on behalf of UWAG and WESTCAS), Comments on EPA's Preliminary Report: Interlaboratory WET Variability Study (December 11, 2000).

Koorse, Steven J. (on behalf of UWAG), Comments on EPA's Proposed Charge to Reviewers: Interlaboratory Study of WET Test Methods (September 15, 1998).

Kszos, L.A., A.J. Stewart, and J.R. Sumner. 1997. Evidence that variability in ambient fathead minnow short-term chronic tests is due to pathogenic infection. *Environ. Toxicol. Chem.* 16:351-356. **WET-IX, C.15**

La Point, T.W. and W.T. Waller. 2000. Field assessments in conjunction with whole effluent toxicity testing. *Environ. Toxicol. Chem.* 19:14-24.

Lazorchak, J.M., P.W. Britton, M.E. Smith, and J.D. Helm. 1997. Summary and Methods Variability Issues of the U.S. EPA Discharge Monitoring Report Quality Assurance Program (DMRQA) Whole Effluent Toxicity Testing (WETT) from 1991-1997. U.S. EPA, Cincinnati, OH.

Marcus, M.D. and L.L. McDonald. 1992. Evaluating the statistical bases for relating receiving water impacts to effluent and ambient toxicities. *Environ. Toxicol. Chem.* 11: 1389-1402.

Markle, P., et al. 2000. Effects of several variables on whole effluent toxicity test performance and interpretation. *Environ. Toxicol. Chem.* 19:123-132. **WET-IX, C.6**

Martin, M., et al. 1989. Experimental evaluation of the mysid *Holmesimysis costata* as a test organism for effluent toxicity testing. *Environ. Toxicol. Chem.* 8:1003-1010. **WET-IX, B.40**

Moore, T.F., S.P. Canton, and M. Grimes. 2000. Investigating the incidence of Type I errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia*. *Environ. Toxicol. and Chem.* 19:118-122. **WET-IX, B.7**

Mount, Donald, National Effluent Toxicity Assessment Center, U.S. EPA Environmental Research Laboratory - Duluth, NETACommunique re: Number of Test Concentrations Needed (January 1990).

Norberg-King, Teresa J., U.S. EPA Environmental Research Laboratory - Duluth, Memorandum to Rob Pederson, EPA Region X, *Review of the Toxicity Results from West Boise and Landers Street POTWs* (June 5, 1989).

Novartis Crop Protection. *An Ecological Risk Assessment of Diazinon in the Sacramento and San Joaquin River Basins* (November 1997).

Office of Management and Budget, *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*, 66 Fed. Reg. 49,718 (Sept. 28, 2001).

Parkhurst, B.R. 1996. Predicting Receiving System Impacts from Effluent Toxicity, in Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts. D. R. Groethe, et al. (eds.). SETAC Press, Pensacola, FL, pp. 309-321.

Parkhurst, B.R., W. Warren-Hicks, and L.E. Noel. 1992. Performance characteristics of effluent toxicity tests: summarization and evaluation of data. *Environ. Toxicol. Chem.* 11:771-791.

Pletl, James, Memorandum re: January 8, 2002 EPA WET Testing Lab Stakeholder Meeting in Chicago, IL (January 9, 2002).

Risk Sciences, Developing A Detection Level for Whole Effluent Toxicity (WET) Testing (2002).

Risk Sciences, Memorandum to Jim Pletl re: Power Analysis (December 26, 2001).

Risk Sciences, Regulating Whole Effluent Toxicity Using "Percent Effect" as the Test Endpoint (2001).

Risk Sciences, Test Sensitivity for Ceriodaphnia Reproduction Using Reference Toxicants: EPA's Whole Effluent Toxicity Interlaboratory Variability Study (2001).

Risk Sciences (on behalf of WESTCAS), Comments on EPA's Proposed Charge to Reviewers: Interlaboratory Study of WET Methods (September 14, 1998).

Settlement Agreement, *Edison Electric Institute, et al. v. EPA*, No. 96-1062 and consolidated cases (D.C. Cir.) (July 24, 1998) ("Settlement Agreement"). **WET-IX, B.21**

Shukla, R., et al. 2000. Bioequivalence approach for whole effluent toxicity testing. *Environ. Toxicol. Chem.* 19:169-174.

Sutfin, Charles S., et al., U.S. EPA Office of Water, Memorandum to EPA Regional Water Management and Enforcement Division Directors, *Certification of "Accuracy" of Information Submissions of Test Results Measuring Whole Effluent Toxicity* (March 3, 2000).

Swygert, Bruce, South Carolina DHEC, Letter to Glenn Stoner, Milliken & Company, re: Third Revised Draft NPDES Permit No. SC0003191 (October 11, 2001).

U.S. Environmental Protection Agency, *Availability, Adequacy, and Comparability of Testing Procedures for the Analysis of Pollutants Established Under Section 304(h) of the Federal Water Pollution Control Act, Report to Congress*, EPA/600/9-87/030 (September 1988) ("Section 518 Report"). **WET-VIII, B.9**

U.S. Environmental Protection Agency, *EPA Quality Manual for Environmental Programs*, EPA 5360 A1 (May 5, 2000).

U.S. Environmental Protection Agency, *EPA Region IX Whole Effluent Toxicity Training Course Manual* (1999).

U.S. Environmental Protection Agency, *EPA Requirements for Quality Assurance Project Plans* (EPA QA/R-5), EPA/240/B-01/003 (March 2001).

U.S. Environmental Protection Agency, *EPA Requirements for Quality Management Plans* (EPA QA/R-2), EPA/240/B-01/002 (March 2001).

U.S. Environmental Protection Agency, *Final Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods*, Vol. 1, EPA 821-B-01-004 (September 2001) ("WET Study Report"). **WET-IX, B.1**

U.S. Environmental Protection Agency, *Final Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods*, Vol. 2: Appendix, EPA 821-B-01-005 (September 2001). **WET-IX, B.2**

U.S. Environmental Protection Agency, *Guidance for Data Quality Assessment: Practical Methods for Data Analysis* (EPA-QA/G-9), EPA/600/R-96/084 (July 2000).

U.S. Environmental Protection Agency, *Guidance for the Data Quality Objectives Process* (EPA QA/G-4), EPA/600/R-96/055 (August 2000) ("DQO Guidance"). **WET-VIII, C.4**

U.S. Environmental Protection Agency, *Guidance on Data Quality Indicators* (EPA QA/G-5i) (September 2001) (Peer Review Draft) ("DQI Guidance").

U.S. Environmental Protection Agency, *Guidance on Evaluation, Resolution, and Documentation of Analytical Problems Associated with Compliance Monitoring*, EPA 821-B-93-001 (June 1993).

U.S. Environmental Protection Agency, *Guidelines and Format for Methods to be Proposed at 40 CFR Part 136 or Part 141*, (July 1996) (Draft).

U.S. Environmental Protection Agency, *Guidelines for Selection and Validation of USEPA's Measurement Methods* (August 1987) (Draft) ("EPA Guidelines"). **Attachment to 304(h) WET, I-B.65**

U.S. Environmental Protection Agency, *Method 1631, Revision B: Mercury in Water by Oxidation, Purge and Trap, and Cold Vapor Atomic Fluorescence Spectrometry*, EPA 821-R-98-002 (May 1999).

U.S. Environmental Protection Agency, *Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing (40 CFR Part 136)*, EPA 821-B-00-004 (July 2000) ("WET Testing Guidance"). **WET-IX, B.11**

U.S. Environmental Protection Agency, *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*, 4th Ed., EPA/600/4-90/027F (August 1993) ("Acute Methods Manual"). **WET-IX, E.8**

U.S. Environmental Protection Agency, *NPDES Permit Writer's Guide to Data Quality Objectives* (November 1990) ("Permit Writer's DQO Guide"). **WET-VIII, C.5**

U.S. Environmental Protection Agency, *NPDES Permit Writers' Manual*, EPA-833-B-96-003 (December 1996) ("Permit Writers' Manual").

U.S. Environmental Protection Agency, *Policy and Program Requirements for the Mandatory Agency-Wide Quality System*, EPA Order 5360.1 A2 (May 5, 2000).

U.S. Environmental Protection Agency, *Proposed Changes to Whole Effluent Toxicity Method Manuals*, EPA 821-B-01-002 (September 2001) ("Proposed Method Manuals Changes"). **WET-IX, B.4**

U.S. Environmental Protection Agency, *Response to Comments: Peer Review of "Preliminary Report: Interlaboratory Variability Study of EPA Short-Term Chronic and Acute Whole Effluent Toxicity Test Methods,"* (September 2001) ("Response to Peer Review Comments"). **WET-IX, D.1**

U.S. Environmental Protection Agency, *Short-Term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Water to Freshwater Organisms*, 3rd Ed., EPA-600-4-91-002 (July 1994) ("Chronic Freshwater Manual"). **WET-IX, E.4**

U.S. Environmental Protection Agency, *Stressor Identification Guidance Document*, EPA 822-B-00-025 (December 2000).

U.S. Environmental Protection Agency, *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 by Versar, Inc. (March 2001). **WET-IX, B.6**

U.S. Environmental Protection Agency, *Supplementary Information Document, Whole Effluent Toxicity: Guidelines Establishing Test Procedures for the Analysis of Pollutants* (October 2, 1995) ("WET SID"). **304(h) WET, II-B.1**

U.S. Environmental Protection Agency, *Technical Support Document for Water Quality-Based Toxics Control*, EPA 505/2-90/001 (March 1991) ("TSD"). **WET-IX, B.35**

U.S. Environmental Protection Agency, *Test Methods for Evaluating Solid Waste,Physical/Chemical Methods, SW-846*, 3rd Ed. (currently being revised, 54 Fed. Reg. 3,212 (January 23, 1989)).

U.S. Environmental Protection Agency, *Understanding and Accounting for Method Variability in WET Applications Under the NPDES Program*, EPA 833-R-00-003 (June 2000) ("WET Variability Guidance"). **WET-IX, B.12**

U.S. Environmental Protection Agency, *Whole Effluent Toxicity (WET) Control Policy*, EPA 833-B-94-002 (July 1994). **304(h) WET, III-B.14**

Water Environment Research Foundation, *WET Testing Program: Evaluation of Practices and Implementation*, Report #D83001 (1998). **WET-IX, B.19**

Water Environment Research Foundation, *Whole Effluent Toxicity Testing Methods: Accounting for Variance*, Report #D93002 (1999) ("WERF Variance Report"). **WET-IX, C.9**