False Positives and False Negatives

Definition of Terms

What are false positives and false negatives?
The meaning of false positives and false negatives in the context of WET testing depends on the assumed objective of WET tests. WET tests may be viewed as indicators either of instream effects or of toxic amounts of chemicals in a sample. Frequently, these terms have taken on the following inappropriate connotations:

False Positive = WET test that indicates toxicity with no associated ecological effect in the water body that was sampled, and

False negative = WET test that does not indicate toxicity even though there is impairment of the waterbody that was sampled.

In the context of NPDES permitting however, these definitions are incorrectly applied. The lack of an observed instream impact may well be expected unless critical flow conditions exist instream at the time of the analysis. Likewise, WET tests of a discharge may not predict an instream effect when impairment exists due to habitat alteration. A variety of literature has discussed these factors, and the reader is referred to numerous discussions of the topic (de Vlaming and Norberg-King, 1999; Diamond and Daley, 2000; Diamond, *et al*, 1999; Dickson et al, 1995; La Point and Waller, 2000; Parkhurst, 1995; Schimmel and Thursby, 1995).

Alternatively, if WET tests are viewed as biological detectors of "toxics in toxic amounts" or the lack or imbalance of physiologically necessary ions then the terms "false positive" and "false negative" suggest an analogy with analytical testing. In analytical testing, false positives are an indication of the presence of an analyte above detection limits when none is present, and a false negative is the failure of the procedure to detect analytes that are actually present above detection limits. According to this view of WET testing:

False Positive = an indication of toxicity when there is no physical or chemical characteristic of the sample (e.g. presence of a toxic amount of a chemical, absence or imbalance of an essential chemical) that would normally or typically cause an organism's response (e.g. death), and

False Negative = the failure of the test organisms to respond when there is a physical or chemical characteristic of the sample that would normally or typically cause an organism's response.

The following discussion focuses on WET tests as detectors of toxics in toxic amounts.

What is toxicity?

For this discussion it is important to establish a working definition of toxicity. Toxicity is an actual physiological process that affects a biological system. In the context of compliance permitting, toxicity is a statistical concept that describes a statistically significant effect of an effluent treatment compared to a control. Regardless of the presumed objective of WET tests, the WET test system will register toxicity anytime there is:

1) A statistically significant reduction in the endpoint measured between a sample (e.g. the permitted instream waste concentration dilution and the control) or
2) A value of a point estimate that is less than some predetermined value.

Accordingly, a false positive is statistically significant negative effect, or observed value of an endpoint, that is not "real" (is spurious or artifactual) or for some reason is not due to the presence of a toxics in toxic amounts (or lack of an essential constituent). Similarly, a false negative represents the failure of the test organisms to respond when there is a physical or chemical characteristic of the sample that should cause an organism response (e.g. death) or when there is a lack of test sensitivity to detect a statistical significance.

In making these considerations, the concentrations or effects of individual chemicals are not necessarily reliable as predictors of the expression of toxicity. The greatest value of WET analyses is their unique capability to measure a matrix effect of the entire solution. They account for additivity, antagonism and synergy for all components of the solution without the need to analytically identify or quantify each component.

General Comments

Why do false positives and false negatives warrant our attention?
The basic concepts of false positives and false negatives are appropriate for WET testing because, as with any monitoring or test system, it is useful to know how often the system can be expected to give "wrong" answers. Because unreliable WET results may lead to unnecessary regulatory action or undetected environmental harm, this issue has received some attention in the literature (Moore et al, 2000; Warren-Hicks *et al*, 2000, Grothe *et al* 1996); USEPA 2000).

How can we identify false positives or false negatives?
It is not possible to identify any particular isolated test outcome as a false positive or a false negative (EPA, 2000). However, it is possible to incorporate additional information that will aid in the interpretation of test results. For example, Quality Assurance/Quality Control (QA/QC) information can be used to assess the overall quality of the test system (test facilities, culture procedures, test maintenance, technician skill, etc.) and statistical issues such as test sensitivity can be evaluated through further analysis of test data. In addition, test results can be evaluated with respect to their repeatability and the magnitude of the effect.

Repeatability: In a reliable test system, false positive and false negative results are expected to be infrequent random events so that effects observed consistently are likely to be due to the presence of toxics in toxic amounts. However, it is possible that consistently committed errors could give the appearance of consistent results. QA/QC practices are crucial in preventing, minimizing and identifying these errors.

While repeatability is a fundamental criterion used to interpret scientific data, it is not, strictly speaking, possible to repeat a WET test. Effluent and environmental samples are unique and a repeated test involves test conditions and test organisms that differ, if only slightly, from the original test. In addition, matrix/toxicant characteristics can change during sample storage in the time lag between the original test and a repeated test. Nonetheless, follow-up testing is a valuable and useful tool for verification of test results. In the absence of significant procedural errors, effects that are observed in follow-up tests are unlikely to be false positives. Failure to observe relatively large effects in follow-up tests can, in itself, be informative. Small, marginally significant effects can be expected to be more difficult to verify through follow-up testing due to test variability and changes in the sample during storage.

Magnitude of the Effect: Another indication of the reliability of a WET test result is the magnitude of the observed effect. Very large effects are less likely to be "false" than small effects. Again, it is clearly possible for procedural errors to cause a large effect, thereby giving the impression of a highly toxic sample. Poor technician practices in the setup and maintenance of the test can introduce errors (e.g. contamination from glassware, dissolved oxygen problems due to overfeeding, careless handing of organisms, improper use of analytical balances) resulting in a false indication of toxicity. The objective of QA/QC is to detect, prevent and correct not only gross errors like those just listed but more subtle sources of error as well (e.g. bias due to non-random test design, variation in organism health due to inconsistent feeding, handling, and culture protocols). Effective and conscientious application of QA/QC and thorough training of technicians are essential elements in the prevention of procedural errors. Assuming that sound QA/QC practices are applied and gross procedural errors are not influencing results, a test showing complete mortality in all test concentrations after a few hours is very likely a repeatable result due to the presence of a toxic chemical. Test results that are marginally significant may be more likely to be due to factors unrelated to the presence of toxic amounts of chemicals. Similarly, observed differences that are very small and statistically non-significant are less likely to represent false negatives than statistically non-significant differences that are large. Means of identifying false positives and false negatives a are discussed in further detail below.

Continuing the analogy with analytical testing, it would be possible to assess the frequency of false positives and false negatives in WET testing through a properly designed and executed series of reference toxicant and "blank" testing. In theory, such tests could provide an estimate of error rates due to at least some of the causes described below.

Which is more important, or of greater concern; false positives or false negatives?
The relative importance of false positives vs. false negatives depends on how WET results are used. For example, from the perspective of a permitee, false positives can result in unnecessary additional testing and sampling expenses. In future analyses they can falsely indicate a history of toxicity. In contrast, from a risk assessment perspective, false negatives could represent undetected ecological effects (Crane and Newman 2000).

Reasons that Toxicity May Be Observed

Is there anything besides the presence of toxic chemicals (or lack of essential constituents) that can cause mortality or poor organism performance?
Statistically significant test results are determined by the individual test replicate response. The individual replicate and organism response can be influenced not only by sample characteristics but also by how the replicates were handled at the start of the test and maintained during the test. Intentionally, there are only a very few variables at play in any given analysis. Only effects on the test organisms due to the sample are of interest. An attempt is made through experimental design to control other relationships within the test system so that factors other than sample characteristics do not influence the test results. This can be of particular importance in test modifications to control specific, known causes of toxicity such as the use of headspace CO2 to assess the effects of pH on ammonia toxicity or sample sterilization to control pathogen interference. In a well executed and randomized toxicity test there are relatively few uncontrolled variables other than toxicity due to the sample.

Causes of Toxicity Not Related to Characteristics of the Sample

Technician error can clearly cause a false indication of toxicity. Lack of technician expertise and poor practice in culturing and handling test organisms can compromise the condition and health of test organisms. Technician errors include not only simple mistakes that occur even with competent, conscientious personnel, but also systematic flaws in the conduct of the tests and culture operations. These errors affect the "other" aspects of the test replicate that were mentioned above (diluent, organisms, food, test surroundings, etc) and can cause statistically significant test results (i.e. toxicity) for reasons not related to characteristics of the sample. All of the fundamental aspects of WET testing from culturing to performing the test to data analysis must be sound, consistent and performed by personnel who have expertise and experience in WET testing. See Burton *et al.*, (1996) for an extensive discussion of factors that may affect effluent toxicity and test variability.

Bias in assigning organisms to treatments is a potential source of false positives and false negatives and can be controlled by proper randomization (Davis *et al*, 1998). Consistent bias in organism selection can potentially lead to consistent false indications of toxicity. With some test systems the effects of inadequate randomization may be subtle. With others, randomization of test chambers is extremely important to avoid systematic errors in organism response. For example, in the *Selenastrum capricornutum* (*Raphidocelus subcapitata*) algal growth test, small variations in light intensity across the growth chamber will result in measurable effects on algal growth rates. Scrupulous randomization of test chambers is necessary to eliminate this source of bias in that test system.

Davis *et al* (1998) performed a study of the potential effects of various randomization schemes on the results of WET tests. Using computer-generated data, and assumptions about the distribution of characteristics related to organism vigor, they simulated the sub-sampling that occurs when test organisms are placed in test beakers. An interesting result of the study was that, while biases due to non-random organism assignment could be substantial, relatively modest randomization efforts seemed to significantly reduce bias.

Are some kinds of WET tests more prone to false positives or false negatives than others?

Effects of some of the factors mentioned above are exacerbated by small sample size. The number of replicates, number of organisms per replicate and the number of treatments in a WET test determine sample size. In any experimental design it is possible for the performance of one organism to affect the statistical outcome, but with some tests, the potential effect of a single individual is particularly pronounced. For example, the 7-day chronic *Ceriodaphnia dubia* test typically has 10 females per test concentration and statistical significance can depend on the life or death of a single individual test organism. To illustrate, with no control mortality in this test method, 40% mortality in a test exposure is statistically significant while 30% mortality is not. With one dead control organism, 50% mortality is statistically significant while 40% mortality is not. Small sample sizes may increase uncertainty and increase the probability of false positives and false negatives. In a properly randomized test system these effects must be considered real. That is, they must be classified as "toxicity" because they are statistically significant effects. Increasing sample size can minimize the relative effects of random anomalies. USEPA (2000) provides information on the number of replicates needed to detect a given effect depending on the variability among replicates.

Are false positives and false negatives still possible in the absence of technician error or sub-standard laboratory performance?

In addition to induced sources of error, false positives and false negatives can occur for purely statistical reasons.

Type 1 and Type 2 Errors. With any hypothesis test (including hypotheses associated with point estimates), only two possibilities exist with respect to the null hypothesis: Either it is true or it is false. A Type 1 error (false positive) occurs if a difference is declared when the null hypothesis is true (i.e., declaring an effluent toxic when it is not toxic). A Type 2 error (false negative) occurs when the null hypothesis is false but no difference is declared (i.e., declaring an effluent not toxic when it actually is toxic). Type 1 and Type 2 erros are dependent on each other (e.g., as alpha increases, beta decreases), assuming that the sample size (number of treatments, number of replicates), the amount of difference to detect, and the variability are held fixed. A full discussion of Type 1 and 2 errors can be found in most statistics textbooks (Zar 1999) and will not be repeated here. The crucial point for this discussion is that these errors are the result of sampling error. They are an inherent part of any experimental system involving experimental units that are variable and they can affect both hypothesis tests and point estimates. With hypothesis tests, sampling errors can result in unusually large, or small, observed differences between the control and treatment and/or unusually low, or high, variation among replicates. These errors result in unusually sensitive (false positive) or insensitive (false negative) tests. With point estimates sampling errors can cause unusually large, or small, confidence intervals around the endpoint estimate and can affect the statistical significance of the dose response. In properly randomized experiments, the magnitude of either Type 1 or Type 2 errors can be controlled.

In contrast with false positives and false negatives due to technician error or poor randomization, Type 1 Type 2 errors are repeatable only in the sense that they occur with a given frequency that can be controlled by the investigator. Typically, WET test statistical analysis protocols fix Type 1 errors at a standard level (alpha = 0.05) and allow Type 2 errors to vary. In the typical WET test, false positives due to sampling error should be rare (i.e. $P < 0.05$) while false negatives may be more common (USEPA 2000).

A factor that determines the magnitude of Type 1 and Type 2 errors (in addition to the investigator's choice of the error rates) is the inherent variability of the replicate response in the test. With a fixed Type 1 error, Type 2 errors increase and statistical power decreases with test organism variability. This variability has two components: Genetic and environmental. With monoclonal cultures (e.g. *Ceriodaphnia dubia*), genetic variability should be negligible so that variability is primarily an organism culture issue (but see Soares *et al* 1992). With sexual test organisms, differences among individual responses are clearly due in part to genetic differences. These differences could, in theory, be minimized through culture techniques such as controlled inbreeding. However, measures to reduce genetic variability of sexual test organisms are likely to have other undesirable consequences such as inbreeding depression. In general, vigorous, healthy organism cultures are characterized by a high degree of homogeneity. This further emphasizes the importance of sound randomization and technical competence and expertise in conducting WET tests.

Errors Due to Test Organism Variability

The subject of false positives and false negatives is closely related to the issue of WET test variability. This issue has received considerable recent attention. In a comprehensive review of reference test data from a variety of sources USEPA (2000) produced guidance for evaluating and controlling WET test variability. Some recommendations included :
1. Paying close attention to technician skill and experience,
2. Implementing procedures to evaluate test sensitivity and minimize within-test variability based on the PMSD. Laboratories should plot control charts for PMSD, and plot the individual raw test data and the average treatment responses to examine possible causes for excessive variability.
3. Applying rigorous and consistent QA/QC oversight in all WET testing.

Part of WET test variability (within and among laboratories) is due to variation in organism vigor and health. Even in well-controlled culture operations with rigorous QA/QC, the vigor of test cultures varies through time. This variability is one source of temporal inter-test variability. Cultures may, at times, produce test organisms that meet performance criteria for testing but are still susceptible (relative to the organisms that the culture usually produces) to relatively small changes in their living conditions. During these periods of lower organism vigor, small departures from culture conditions can result in problems with the test system such as culture crashes and invalid tests. This is one reason that laboratories take considerable pains to control and keep constant as many culture activities as possible.

If a culture is producing test organisms that are susceptible to relatively small departures from culture conditions, then samples that should not be toxic (i.e. contain no toxics in toxic amounts or ionic imbalances) might indicate toxicity simply because they depart from culture conditions. Conversely, it is possible for organism cultures to produce organisms that are unusually vigorous and resistant to toxic stresses. These episodes of unusually low or high test organism vigor can result in tests that are unusually sensitive or insensitive. Because these episodes are relatively uncommon, they may tend not to be repeatable and can be viewed as false positive or false negative results. Normally implemented reference toxicant testing in each laboratory may be able to identify these types of unusual conditions and minimize their effects in compliance WET testing. There still remains though an acceptable range of sensitivity among organisms within a laboratory that can affect their range of response to a toxicant.

It is not known, at this time, to what extent these kinds of false positive and false negative results occur. A properly designed and executed series of "blank" and reference toxicant testing could provide an estimate of false positive and false negative rates due to all causes. EPA (2000) concluded that hypothesis test procedures prescribed in EPA's WET methods will provide adequate protection against false conclusions that and effluent is toxic. In addition, EPA (2000) provides guidance for identifying tests that detect small statistically significant differences.

Summary

Toxicity is an actual physiological process that affects a biological system, but in the context of compliance permitting it is a statistical concept that describes a statistically significant effect of an effluent treatment compared to a control. A false positive is an indication of toxicity when there is no physical or chemical characteristic of the sample (e.g. presence of a toxic amount of a chemical or mixture of chemicals, absence of an essential chemical or mixture of chemicals) that should cause toxicity or when there is an unusually high degree of statistical sensitivity. A false negative represents the failure of the test organisms to respond when there is a physical or chemical characteristic of the sample that should cause an organism response (e.g. death) or when there is an unusually low degree of statistical sensitivity.

Statistically significant test results can be due to characteristics of the sample or to aspects of the test system unrelated to the sample (e.g. Diluent, organisms, food, maintenance and conduct of the tests, sampling error.) In the absence of technician or experimental error or biased sampling, the false positive and false negative rates should equal the Type 1 and 2 error rates, respectively. In the absence of technician error and biased sampling, repeatable results are likely not false positives or false negatives. In the absence of technician error and biased sampling, large effects are less likely to be false positives than small effects.

A properly designed and executed series of reference toxicant and "blank" testing could provide an estimate of the false positive and false negative rates due to all causes. Potential problems with false positive and false negatives are seen with both hypothesis test and point estimates for the same fundamental reasons. Some efforts are underway to incorporate known sources of variability to establish upper and lower limits for WET test sensitivity (based on the PMSD) to reduce the frequency of false positives and false negatives (USEPA, 2000).

Literature Cited

Burton GA, Arnold WR, Ausley LW, Black JA, DeGraeve GM, Fulk FA, Heltshe JF, Peltier WH, Pletl JJ, Rodgers JH. 1996. Pp 131 – 156 In: Grothe DR , KL Dickson, DK Reed-Judkins eds. *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving Stream Impacts*, SETAC Press, 1996.
Crane M, Newman MC. 2000. What level of effect is a no observed effect? Environ Toxicol Chem 19: 516–519.
Davis RB, Bailer AJ, Oris JT. 1998. Effects of organism allocation on toxicity test results. Environ Toxicol Chem17:928-931.
Diamond J, Daley C. 2000. What is the relationship between whole effluent toxicity and instream biological conditions? Environ Toxicol Chem 19:158–168.
Dickson KL, Waller WT, Kennedy JH, Ammann LP, Guinn R, Norberg-King TJ. 1995. Relationships between effluent toxicity, ambient toxicity, and receiving stream impacts: Trinity River dechlorination case study. Pp 287-308 In: Grothe DR , KL Dickson, DK Reed-Judkins eds. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving Stream Impacts, SETAC Press, 1996.
Grothe DR , Dickson KL, Redd-Judkins, DK. 1995. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving Stream Impacts, SETAC Press, 1996.
La Point TW, and Waller WT. 2000. Field assessments in conjunction with whole effluent toxicity testing. Environ Toxicol Chem 19:14–24.
Moore TF, Canton SP, Grimes M. 2000. Investigating the incidence of Type I errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia*. Environ Toxicol Chem 19:118–122.
Parkhurst BR. 1995. Predicting receiving stream impacts from effluent toxicity. Pp 309-321 In: Grothe DR , KL Dickson, DK Reed-Judkins eds. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving Stream Impacts, SETAC Press, 1996.
Schimmel SC, Thursby GB. 1995. Predicting system impacts from whole effluent toxicity: A marine perspective. Pp 322 – 330 In: Grothe DR , KL Dickson, DK Reed-Judkins eds. Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving Stream Impacts, SETAC Press, 1996.
Soares A, Baird DJ, Calow P. 1992. Interclonal variation in the performance of *Daphnia magna* Straus in chronic bioassays. Envir Toxicol Chem 11:1477-1483
USEPA . 1999. A review of single species toxicity tests: Are the test reliable predictors of aquatic ecosystem community responses? Eds: de Vlaming V and TJ Norberg-King., ORD Duluth, MN. EPA/600 /R-97/114
USEPA. 2000. Understanding and accounting for method variability in whole effluent toxicity applications under the National Pollutant Discharge Elimination System. Eds: Denton DL, Fox J, Fulk F, Greenwald K, Narvaez M, Norberg-King TJ, Phillips L.Office of Wastewater Management. Washington , DC . EPA\833-R-00-003
Warren-Hicks WJ, Parkhurst BR, Moore DJ, Teed RS, Baird RB, Berger R, Denton DL, Pletl JJ. 2000. Assessment of whole effluent toxicity test variability: Partitioning sources of variability. Environ Toxicol Chem 19:94–104.
Zar JH. 1999. Biostatistical analysis. 4th ed. New Jersey : Prentice Hall.