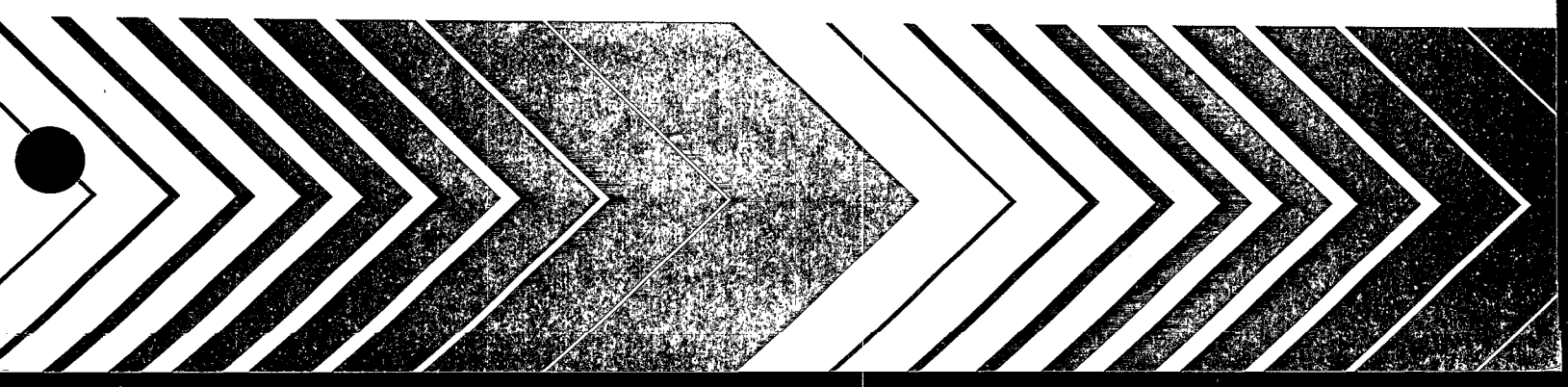
United States
Environmental Protection
Agency

Office of Research and
Development
Washington DC 20460

EPA/600/R-97/114
July 1999

EPA

A Review of Single Species Toxicity Tests: Are the Tests Reliable Predictors of Aquatic Ecosystem Community Responses?



EPA/600/R-97/114
July 1999

A Review of Single Species Toxicity Tests: Are the Tests Reliable Predictors of Aquatic Ecosystem Community Responses?

By

Victor de Vlaming[1]
Teresa J. Norberg-King[2]

[1] State Water Resources Control Board
901 P Street
PO Box 944213
Sacramento, California 94244-2130

[2]Mid-Continent Ecology Division
6201 Congdon Boulevard
Duluth, MN 55804-1636

Office of Research and Development
U.S. Environmental Protection Agency
Duluth, Minnesota 55804

Notice

This document has been reviewed according to U.S. Environmental Protection Agency Policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

The views expressed in this document are those of the individual authors and do not necessarily reflect the view and policies of the U.S. Environmental Protection Agency or the State Water Resources Control Board.

Abstract

This document provides a comprehensive review to evaluate the reliability of single species (also referred to as indicator species) toxicity test results in predicting aquatic ecosystem impacts, also known as the ecological relevance of laboratory single species toxicity tests. Since aquatic ecosystem biological assessments have been performed to determine whether toxicity test results are predictive of biological community impacts, the strengths and limitations of these validation tools have been assessed. Ecological relevance has been analyzed in studies on ambient waters, effluents, and other types of aqueous media. Furthermore, the effectiveness of laboratory single species toxicity tests with individual chemicals in predicting biological community impacts and/or environmental adverse effect concentrations is evaluated. Merits of published criticisms of the predictive effectiveness of single species used in laboratory toxicity tests are analyzed. Also, the question of whether single species used in laboratory toxicity tests are more sensitive than most natural populations is discussed. Alternatives to single species toxicity tests are explored. A preponderance of evidence reveals that laboratory single species toxicity test results are reliable qualitative predictors of aquatic ecosystem community impacts.

Foreword

The US Environmental Protection Agency (USEPA) has begun a long-term process aimed at restoring and maintaining the chemical, physical, and biological integrity of the Nation's waters. One major element in this effort was removing the discharge of toxic materials in toxic amounts to surface waters. Through the policy designed to reduce or eliminate toxics discharge and to assist in achieving objectives of the Clean Water Act (CWA), USEPA issued technical direction in the Technical Support Document for Water Quality-Based Toxics (TSD) guidance (March 1984 Policy for the Development of Water Quality-Based Permit Limitations for Toxic Pollutants; 49 FR 9016). Through these directives, the Agency described its integrated toxics control program. The integrated program consists of the application of both chemical-specific and biological methods to address the discharge of toxic pollutants. USEPA continued with the development of the toxics control program by developing effluent toxicity test methods, and these methods are being used to assess the quality of surface waters, effluents, stormwater, as well as other types of aqueous media. The use of toxicity tests for biological monitoring provides tools that can be used to assess the combined effect of mixtures and unknown constituents in a water sample to be evaluated, which in turn provides a direct evaluation of the attainment of protection to the aquatic life.

Many uses for laboratory toxicity tests are to determine compliance with enforceable water quality standards and effluent limits. The concept behind, and the intent of, the single species toxicity tests (also referred to as indicator species tests) is to assess the probability of impacts on aquatic ecosystems. To be effective water quality monitoring tools, toxicity test results should have a predictive relationship with aquatic ecosystem impacts. USEPA (1991) reported that the results of indicator species toxicity tests are effective predictors of aquatic ecosystem impacts. This comprehensive literature review was undertaken to provide a critical examination of the relationships among ambient water toxicity, effluent toxicity, and effects on organisms in ambient waters.

Contents

Contents (continued)

List of Tables

List of Figures

Acknowledgments

This document was peer reviewed by numerous individuals and this version of the document incorporates reviewer recommendations. These review comments were considerably valuable in improving the quality, accuracy and clarity of the literature review.

The reviewers of the early drafts of this document who offered helpful suggestions are:
Larry Ausley (North Carolina DENR, Raleigh, NC),
Tom Dean (Coastal Resources Associates, Vista, CA),
Debra Denton (USEPA, Region 9, San Francisco, CA),
Regina Donohoe (California EPA, Sacramento, CA),
Chris Foe (Central Valley Regional Water Quality Control Board, Sacramento, CA),
Jeff Miller (Aqua-Science, Davis, CA),
Don Mount (Ascl Corporation, Duluth, MN), and
Michael Perrone (State Water Resources Control Board, Sacramento, CA).

The critiques and suggestions of the following individuals were particularly valuable:
Brian Anderson (University of CA-Santa Cruz, Monterey, CA),
Gordon Anderson (deceased, formerly of Santa Ana Regional Water Quality Control Board, Riverside, CA),
Rodger Baird (City Sanitation Districts of Los Angeles, Whittier, CA),
Peter Chapman (EVS Consultants, Vancouver, BC),
JoAnne Cox (State Water Resouces Control Board, Sacramento, CA),
Carol DiGiorgio (Department of Water Resources, Sacramento, CA), and
Mike Marcus (The Cadmus Group, Albuquerque, NM).

John Cairns, Jr. (Virginia Tech, Blacksburg, VA) provided an abundance of the relevant literature for this review.

We appreciate the peer reviews conducted by Robert Spehar and Jo Thompson, USEPA, Office of Research and Development, Mid-Continent Ecology Division, Duluth, MN.

Without the assistance and cooperation of Robert Holmes (California State University, Humboldt, Arcata, CA) and M. Perrone (Water Resources Control Board) this document could not have been produced.

Acronyms and Abbreviations

A. punctulata	sea urchin, *Arbacia puntulata*
AEC	Acceptable Effluent Concentration
C. dubia	cladoceran, *Ceriodaphnia dubia*
C. variegatus	sheepshead minnow, *Cyprinodon variegatus*
C. parvula	red algae, *Champia parvula*
CETTP	Complex Effluent Toxicity Testing Program
CWA	Clean Water Act
EC	Effect Concentration
FIFRA	Federal Insecticide, Fungicide and Rodenticide
IC	Inhibition Concentration
IWC	Instream Waste Concentration
km	kilometers
LOEC	Lowest Observed Effect Concentration
m	meters
M. bahia	mysid shrimp, *Mysidopsis bahia*
M. berylina	inland silverside, *Menedia berylina*
NOEC	No Observed Effect Concentration
NPDES	National Pollutant Discharge Elimination System
P. promelas	fathead minnow, *Pimephales promelas*
POTW	Publicly Owned Treatment Works (wastewater treatment plant) and also referred to as WWTP
r	Correlation coefficient
RWC	Receiving Water Concentration
S. capricornutum	green algae, *Selenastrum capricornutum*
STP	Sewage Treatment Plant
TSCA	Toxic Substances Control Act
TSD	Technical Support Document (cf., USEPA, 1991)
WET	Whole Effluent Toxicity
WWTP	Wastewater Treatment Plant, also referred to as a POTW

Definitions

Accuracy is the degree of difference between observed values and known or actual values. This is appropriate for chemical and physical measurements, but not biological systems. Toxicity is relative rather than absolute and the organisms measure toxicity without a reference organism in a reference toxicant solution.

Acute Toxicity is a test to determine the concentration of effluent or receiving waters (or ambient waters) that produces an adverse effect on a group of test organisms during a short-term exposure (e.g., 24, 48, or 96 h). The endpoint is lethality. Acute toxicity is measured using statistical procedures (e.g., point estimate techniques or a t-test). Acute toxicity is usually defined as TUa =100/LC50.

Acute-to-Chronic Ratio (ACR) is the ratio of the acute toxicity of an effluent or a toxicant to its chronic toxicity. It is used as a factor for estimating chronic toxicity on the basis of acute toxicity data, or for estimating acute toxicity on the basis of chronic toxicity data.

Additivity is the characteristic property of a mixture of toxicants that exhibits a total toxic effect equal to the arithmetic sum of the effects of the individual toxicants.

Ambient Toxicity is measured by a toxicity test on a sample collected from a surface water.

Bioassay is a test used to evaluate the relative potency of a chemical or a mixture of chemicals by comparing its effect on a living organism with the effect of a standard preparation on the same type of organism. Bioassays frequently are used in the pharmaceutical industry to evaluate the potency of vitamins and drugs.

Criteria Continuous Concentration (CCC) is the USEPA national water quality criteria recommendation for the highest instream concentration of a toxicant or an effluent to which organisms can be exposed indefinitely without causing unacceptable effect.

Criteria Maximum Concentration (CMC) is the USEPA national water quality criteria recommendation for the highest instream concentration of a toxicant or an effluent to which organisms can be exposed for a brief period of time without causing an acute effect.

Chronic Toxicity is defined as a long-term toxicity test in which sublethal effects (e.g., reduced growth or reproduction) are usually measured in addition to lethality. Chronic toxicity is defined as TUc = 100/NOEC or TUc = 100/ECp (ICp). The ICp and ICp value should be the approximate equivalent of the NOEC calculated by hypothesis testing for each test method.

Coefficient of Variation (CV) is a standard statistical measure of the relative variation of a distribution or set of data, defined as the standard deviation divided by the mean. Coefficient of variation is a measure of precision within (intralaboratory) and among (interlaboratory) laboratories.

Critical Life Stage is the period of time in an organisms life span in which it is the most susceptible to adverse effects caused by exposure to toxicants, usually during early development (egg, embryo, larvae). Chronic toxicity tests are often run on critical life stages to replace long duration, life-cycle tests since the most toxic effect usually occurs during the critical life stage.

Effect Concentration (EC) is a point estimate of the toxicant concentration that would cause an observable adverse effect (e.g., survival or fertilization) in a given percent of the test organisms, calculated from a continuous model (e.g., USEPA Probit Model).

Hypothesis Testing is a technique (e.g., Dunnett's test) that determines what concentration is statistically different from the control. Endpoints determined from hypothesis testing are NOEC and LOEC. Null hypothesis (Ho): The effluent is not toxic; Alternative hypothesis (Ha): The effluent is toxic.

Inhibition Concentration (IC) is a point estimate of the toxicant concentration that would cause a given percent reduction in a non-quantal biological measurement (e.g., reproduction or growth) calculated from a continuous model.

Instream Waste Concentration (IWC) is the concentration of a toxicant in a riverine system after mixing. Also referred to as the receiving water concentration (RWC). The IWC or RWC is the inverse of the dilution factor.

LC50 is the toxicant concentration that would cause death to 50% of the test organisms.

Lowest Observed Effect Concentration (LOEC) is the lowest concentration of toxicant to which organisms are exposed in a test, which causes statistically significant adverse effects on the test organisms (i.e., where the values for the observed endpoints are statistically significant different from the control). The definitions of NOEC and LOEC assume a strict dose-response relationship between toxicant concentration and organism response. If this assumption

were always the case, there would be no issue concerning the endpoint definitions because the NOEC would always be a lower concentration level than the LOEC. However, this strict dose-response relationship does not exist with all toxicants. When this occurs the test must be repeated or the lowest NOEC should be reported for compliance purposes.

Minimum Significant Difference (MSD) is the magnitude of difference from control where the null hypothesis is rejected in a statistical test comparing a treatment with a control MSD is based on the number of replicates, control performance and power of the test.

Mixing Zone is an area where an effluent discharge undergoes initial dilution and may be extended to cover the secondary mixing in the ambient waterbody. A mixing zone is an allocated impact zone where water quality criteria can be exceeded as long as acutely toxic conditions are prevented.

No Observed Adverse Effect Level (NOAEL) is a tested dose of an effluent or a toxicant below which no adverse biological effects are observed, as identified from chronic or subchronic human epidemiology studies or animal exposure studies.

No Observed Effect Concentration (NOEC) is the highest tested concentration of toxicant to which organisms are exposed in a full life-cycle or partial life-cycle (short-term) test, that causes no observable adverse effect on the test organism (i.e., the highest concentration of toxicant at which the values for the observed responses are not statistically significant different from the controls). NOECs calculated by hypothesis testing are dependent upon the concentrations selected.

Point Estimation Techniques are used to determine the effluent concentration at which adverse effects (e.g., fertilization, growth or survival) occurred, such as Probit, Interpolation Method, Spearman-Karber. For example, concentration at which a 25% reduction in fertilization occurred.

Precision is a measure of mutual agreement among individual measurements or enumerated values of the same property of the sample; can be described by the mean, standard deviation and coefficient of variation. The precision is usually discussed by test consistency or repeatability both with a laboratory (intralaboratory) and among several laboratories (interlaboratory) using the same test method and reference toxicant.

Receiving Water Concentration (RWC) is the concentration of a toxicant or the parameter toxicity in the receiving water (i.e., riverine, lake, reservoir, estuary or ocean) after mixing. Isopleths of effluent concentration can be established by dye studies or modeling techniques is determining CMC and CCC.

Significant Difference is defined as statistically significant difference (e.g., 95% confidence level) in the means of two distributions of sampling results.

Test Acceptability Criteria (TAC) are defined for toxicity tests results to be acceptable or valid for compliance, the effluent and the concurrent reference toxicant controls must meet specific criteria as defined in the test method (e.g., *Ceriodaphnia dubia* survival and reproduction test, the criteria are: the test must achieve at least 80% survival and average 15 young/female in the controls).

Toxicity Tests are laboratory experiments which employ the use of standardized test organisms to measure the adverse effect (e.g., growth, survival or reproduction) of effluent or receiving waters.

Toxic Unit Acute (TUa) is the reciprocal of the effluent concentration that causes 50% of the organisms to die by the end of the acute exposure period (i.e., TUa = 100/LC50).

Toxic Unit Chronic (TUc) is the reciprocal of the effluent concentration that causes no observable effect on the test organisms by the end of the chronic exposure period (i.e., TUc = 100/NOEC).

Toxic Units (TUs) are a measure of toxicity in an effluent as determined by the acute toxicity units or chronic toxicity units. Higher TUs indicate greater toxicity.

Toxicity Identification Evaluation (TIE) is a set of procedures to identify the specific chemical(s) responsible for effluent toxicity. TIEs are subset of the Toxicity Reduction Evaluation (TRE).

Toxicity Reduction Evaluation (TRE) is a site-specific study conducted in a stepwise process designed to identify the causative agents of effluent toxicity, isolate the sources of toxicity, evaluate the effectiveness of toxicity control options, and then confirm the reduction in effluent toxicity.

Whole Effluent Toxicity (WET) is the total toxic effect of an effluent or receiving water measured directly with a toxicity test.

Section 1

1.0 Introduction

The Clean Water Act (CWA), Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), and Toxic Substances Control Acts (TSCA) are the federal legislation mandating those potential hazards of chemicals and wastewaters be assessed. In particular, the CWA aims at preventing the release of toxic concentrations of chemicals, regardless of whether they originate from point or nonpoint sources, into the nation's surface waters by stating "it is the national policy that the discharge of toxic pollutants in toxic amounts be prohibited."

As part of the effort to implement the above CWA policy, the USEPA incorporated toxicity-based discharge limits into National Pollutant Discharge Elimination System (NPDES) permits. To support this approach, USEPA published a Technical Support Document (TSD) (USEPA, 1991) and short-term toxicity test methodologies (USEPA, 1994a; 1994b; hereafter referred to as the USEPA toxicity tests). The intent of these toxicity tests is to rapidly and reliably estimate the potential chronic effects of toxic chemicals in ambient water and wastewater, stormwater and other water matrices on aquatic life.

For freshwater ecosystems, USEPA has focused on three species for short-term tests designed to estimate the degree of chronic toxicity in a water sample (USEPA, 1994a). These freshwater methods include a fish, larval fathead minnow (*Pimephales promelas*), a zooplankton (*Ceriodaphnia dubia*), and an alga (*Selenastrum capricornutum*). The marine and estuarine short-term tests estimate chronic toxicity (USEPA, 1994b) with two fish species, sheepshead minnow (*Cyprinodon variegatus*) and the inland silverside (*Menidia berylina)*, a red alga (*Champia parvula*), an east coast mysid (*Mysidopsis bahia*), and a sea urchin (*Arbacia punctulata*).

USEPA states "whole effluent toxicity (WET) is a useful parameter for assessing and protecting against impacts upon water quality and designated uses caused by the aggregate toxic effect of the discharge of pollutants" (in the TSD; USEPA, 1991). Four data sets were the focus of support for the reliability of the USEPA toxicity tests results in predicting aquatic ecosystem community responses: USEPA's Complex Effluent Toxicity Testing Program (CETTP) studies (USEPA, 1991), the South Elkhorn Creek, Kentucky study (Birge et al., 1989), the Trinity River, Texas study (Dickson et al., 1989), and the North Carolina study performed by Eagleson et al. (1990).

The eight CETTP studies include: Scippo Creek, Ohio (Mount and Norberg-King, 1985); Ottawa River, Ohio (Mount et al., 1984); Five Mile Creek, Alabama (Mount et al., 1985); Skeleton Creek, Oklahoma (Norberg-King and Mount, 1986); Naugatuck River, Connecticut (Mount et al., 1986a); Back River, Maryland (Mount et al., 1986b); Ohio River, West Virginia (Mount et al., 1986c); and Kanawha River, West Virginia (Mount and Norberg-King, 1986). In these studies the 7-d *Ceriodaphnia* and/or early life stage larval fathead minnow toxicity test results from surface water and/or effluents were compared with data from aquatic ecosystem community surveys (bioassessments) to determine whether the toxicity test results were effective predictors of instream biological responses. USEPA concluded (USEPA, 1991) that the four data sets "comprise a large database specifically collected to determine the validity of toxicity tests to predict receiving water community impact. The results, when linked together, clearly show that if toxicity is present (in discharges) after considering dilution, impact will also be present."

Criticisms of the CETTP and associated studies, as well as their conclusions, have been published (see Section 6 below). In a broader sense, there have been questions regarding the reliability of single species (frequently described as indicator species) toxicity test results in predicting aquatic ecosystem responses (impairments). Moreover, there are questions regarding the validity of, and the uncertainty associated with, extrapolations from single indicator species toxicity test results to aquatic ecosystem responses. USEPA also bases their chemical-specific water quality criteria on laboratory single species toxicity test estimates of chronic toxicity, yet the validity and reliability of these criteria are less frequently questioned.

The central aspect of the uncertainty appears to be whether the indicator species toxicity test results, obtained under controlled laboratory conditions, can be reliably translated into responses by complex and multivariant aquatic ecosystem communities. For example, laboratory effluent toxicity test results could overestimate biological community responses if aquatic ecosystem physical/chemical or biotic factors mitigated (e.g., altered chemical bioavailability) effluent toxicity. On the other hand, some aquatic ecosystem physical/chemical and biotic factors could act as stressors which exacerbate the effects of toxic chemicals such that laboratory toxicity test results underestimate instream biological responses. There is also the concern that indicator species toxicity test

results do not represent the range of sensitivities and the different levels of biological organization which exist in aquatic ecosystems. These, as well as other, concerns regarding the reliability of single species toxicity test results in predicting aquatic ecosystem biological responses will be considered in this document.

Regulatory agencies have tended to rely on single species, especially USEPA toxicity tests, test results on surface water and wastewater samples to estimate potential toxicity threats to aquatic ecosystem communities. Since there have been criticisms of the predictive effectiveness of single species tests, the intent in this review is to evaluate and summarize the published literature, as well as other available reports, on the ecological relevance of laboratory single species toxicity test results. This review examines, but is not limited to, the CETTP and associated studies (e.g., Birge et al., 1989; Dickson et al., 1989; Eagleson et al., 1990). Various aspects of the reliability of single species toxicity test results as predictors of biological community impacts have been reviewed by many authors (as noted in Section 3). This report is a comprehensive review of the literature in this area.

The following sections address:

- the intent of single species toxicity tests,
- the procedures (bioassessments) used to "validate" indicator species toxicity test results,
- the concepts of false positives and false negatives,
- the USEPA CETTP and associated studies,
- criticisms of the CETTP studies,
- single species tests with effluent,
- single species tests with individual chemicals or small groups of chemicals,
- comparisons of single species and multiple species test results, and
- alternatives to single species toxicity tests.

The vast majority of the literature in this area relates to toxicity tests with freshwater species and ecosystems. There is a paucity of studies which attempt to relate laboratory toxicity test results with bay and estuary or ocean impacts, nonetheless, the few relevant studies (Section 11) are summarized in this review. The conclusions (Section 2) of this report are weighted toward freshwater toxicity test results as predictors of aquatic ecosystem community responses.

2.0 Intent of Single Species Toxicity Tests

Before summarizing and discussing data which relate to how reliably the USEPA toxicity tests (and other single species toxicity test results) predict ecosystem responses, a consideration of the intent of these tests seems warranted.

A criticism of the single species tests has been that their results are invalid predictors of aquatic community responses because only qualitative (i.e., statistically significant toxicity test results indicate some degree of biological community response/impairment) rather than quantitative relationships are established between toxicity test results and ecosystem community responses. Quantitative in this context refers to a case in which some level or percent response in toxicity test results can be directly correlated with a specific level/percent response in instream biological communities. However, these tests were not designed to be quantitative predictors of ecosystem responses. The USEPA toxicity tests and other indicator single species tests were intended to be screening tools (i.e., to indicate the potential for wastewater or ambient water samples to cause biological community impacts, characterizing relative ecosystem effects) and "early warning" signals (a measurement which indicates the potential for aquatic ecosystem impairment prior to actual damage to biological communities (USEPA, 1991; USEPA, 1994a). The toxicity tests are applicable to ambient water samples regardless of the sources (i.e., point or nonpoint) of contaminants.

Because the USEPA toxicity tests were intended to be early warning signals of biological community impacts, the results of a single toxicity test should not constitute a violation of a water quality standard, or of an effluent limitation. Unfortunately, such misuses have occurred and these cases may be major contributors to the criticisms leveled at the USEPA toxicity tests.

3.0 Validation Procedures: Ecological Surveys/Bioassessments

3.1 Bioassessments

The method generally used for "validating" the reliability of single species toxicity test results in predicting aquatic ecosystem impairments (and "safe" concentrations) has been to perform ecological surveys (biological assessments), and then compare these data to toxicity test results with water samples from the same ecosystem sites or to data from effluent toxicity tests. Bioassessments can consist of estimates of species composition, diversity, and density of aquatic organisms.

Because bioassessments play a crucial role in this "validation" process, there are considerations regarding these procedures which must be explored. From these surveys, a judgement is made as to whether or not the aquatic ecosystem or a part of it is impacted. Bioassessments are not *de facto* better or easier to interpret than other types of measurements.

Bioassessments are subject to most of the same pitfalls as other biological and toxicological studies, including poor design and careless performance. Sound experimental design and careful conduct are crucial, requiring a thorough understanding of the complexity of aquatic ecosystems, as well as confounding factors (e.g., current

velocity, depth, light penetration, shading, temperature, substrate, organic matter, nutrients) which can affect site selection so that they can be "controlled" or accounted for. Moreover, sites within a stream should be chosen to minimize differences among them with respect to physical and chemical parameters. The idea is to minimize factors which can influence ecosystem parameters so that any change can be ascribed to toxic chemicals.

To be effective in "validation" of toxicity test results, ecological surveys must be able to clearly distinguish between contaminant-caused effects and all other effects on aquatic populations. Aquatic ecosystem biological surveys are not, by themselves, sufficient to determine toxic chemical impacts because biological community structure and function are influenced by a host of other factors (e.g., dissolved oxygen, temperature, physical parameters, habitat conditions).

Limitations (LaPoint, 1994; 1995) in bioassessment studies have included failure to consider seasonal variations (frequently sampling is only a one or two time event), poor selection of endpoints (endpoints should be reliable, having ecological relevance), poor sampling procedures, lack of sample replication, failure to consider nonchemical stressors, failure to identify cause of change, use of inappropriate procedures and statistics which are not standardized, and failure to provide early warning of impairment. Many ecological assessments have been characterized by a high degree of variability (greater than in chemical and toxicity measurements), imprecisions, and lack of repeatability (e.g., LaPoint, 1994; 1995).

Many bioassessments provide qualitative (not quantitative) data; for example, macroinvertebrate surveys with kick-nets are qualitative and, usually, are not replicated. Most of the ecological surveys associated with field "validation" of single species test results have consisted of "simplistic field designs" and "superficial study" of the natural system (Neuhold, 1986; Chapman et al., 1987; Luoma, 1995). Neuhold (1986) contends that measurements such as biomass and population numbers, which are frequently the basis of ecological surveys, are too insensitive as endpoints because they take considerable time to change enough to "clear" the background noise level. Interpretations of bioassessment data are frequently controversial. For example, according to LaPoint et al. (1996), biological assessment of contaminant(s) effects is more difficult than laboratory single species toxicity tests with regards to the possible ecological significance due to the large number of aquatic species potentially responding in the system. Clements and Kiffney (1996) state, "Most importantly, inability to establish a direct cause-and-effect relationship between contaminants and selected endpoints greatly limits instream biomonitoring."

There is yet to be agreement on meaningful ecological endpoints or the amount of change in ecological measurements which represent impairment. It has been difficult to identify and measure subtle damage in aquatic ecosystems. No procedures/protocols for performing ecological surveys on large waterways, such as major rivers, have been published. Developing scientifically valid biological assessment methods for such systems is needed, but it seems unlikely that regulatory agencies will have the budgets to fund such large efforts. In relation to these bioassessment concerns, the difficulties surrounding "validation" of laboratory single species toxicity test results have been reviewed (Cairns, 1983; 1988a; Livingstone and Meeter, 1985; Chapman, 1995a,b). These considerations should be remembered when using bioassessment measurements in evaluating the predictive accuracy of the single species or multiple species toxicity test results.

The intent here is not to malign bioassessments, but to draw attention to the fact that they are not *de facto* conclusive. On the other hand, well designed and performed bioassessments are powerful tools, crucial to environmental monitoring and assessment. The advantages of using ecological surveys and, in particular, macroinvertebrate surveys as water quality indicators have been thoroughly discussed in an informative book edited by Davis and Simon (1995).

3.2 Can Laboratory Single Species Tests Be Validated?

Mount (1995) suggests that it is impossible to conclusively establish that ecosystem impairments are caused by ambient water or effluent toxicity. This is because there are many stressors and other confounding factors at work in natural ecosystems. Proving cause in complex, poorly understood ecosystems will be difficult at best. Recently, Chapman (1995a) wrote, "Basically, I consider the perceived need for validation of the laboratory by field studies to be incorrect dogma." A reactive toxicity test can confirm ecosystem impairments, but proactive tests can only be "validated" by waiting for ecosystem effects to appear. Furthermore, absence of biological community effects can never be fully proven. While recognizing and addressing various short-comings in ecotoxicology, Chapman (1995a) warns against perpetuating the "established" validation dogma; he points out, as did Mount (1995), that field studies can never validate laboratory studies since there is no certainty that effects observed (or not) in field studies were caused by effects measured in the lab.

Another inherent problem in "validating" that single species toxicity test results can be reliably extrapolated to ecosystem responses is the status of many aquatic ecosystems. Given the ever expanding number of aquatic

population declines (e.g., Herbold et al., 1992; Obrebski et al., 1992; Bailey et al., 1994) and number of extinct, endangered, and threatened species, it is clear that many aquatic ecosystems are partially to seriously impaired. If single species test results are to be early warnings (predictive of future events), proactive in function, they cannot be "validated"' in all circumstances with existing ecological conditions. The point is not to discontinue study of the relationship, but rather to understand the limitations of the procedures used to "validate" the predictiveness of single species tests.

3.3 To What Extent Should These Tests Be Validated?

Without partial disturbance of healthy or relatively healthy aquatic ecosystems, it may not be possible to "validate" that extrapolations from laboratory toxicity tests reliably predict aquatic ecosystem responses. Depending on scale, biological surveys, especially if repeated through time, may be destructive to aquatic ecosystems. The question is, should toxic chemicals be released into ecosystems to repeatedly establish a link between laboratory toxicity test results and ecological impairments?

Since unequivocal demonstration that effluent or ambient water toxicity is the sole cause of ecosystem impairments may not possible, it seems sensible to question how much effort, time, and money should be expended to "validate" a quantitatively accurate correlation between single species toxicity test results and instream biological responses.

USEPA's toxicity tests were designed as screening tools to provide early warning of potential environmental impacts. For this and other reasons mentioned above, it has been difficult to establish a quantitative correlation between the results of these tests and ecological responses in all aquatic ecosystems. As Chapman (1995b) suggests, we can never be sure that a proactive prediction (based on laboratory toxicity test results) is correct without allowing for potential environmental degradation. Possibly, surrogate aquatic ecosystems will allow us to establish a better link between laboratory test results and ecosystem responses, while minimizing impacts on natural aquatic systems.

4.0 False Positives and False Negatives

In this review, the concepts of "false positives" and "false negatives" will emerge when comparing the results of single species tests with ecological survey measurements. Caution is essential in the application of such concepts.

There has been a tendency to label any one statistically significant toxicity test result which does not match with an ecological endpoint as a false positive. This may be an inaccurate designation. A single effluent or ambient water sample can contain toxic levels of chemicals but, due to effluent or ambient water variability, the duration, magnitude, and frequency of the toxicity are not sufficient to elicit a measurable biological community response. More sampling and testing could reveal this. On the other hand, the toxic sample could be an early warning, signaling toxicity of a magnitude, duration, and frequency to cause adverse ecosystem responses. The false positive designation is also based on the assumption that the measure of ecosystem integrity is accurate.

A false negative designation has sometime been applied to cases in which statistically significant toxicity is absent from an effluent or ambient water sample, but instream impairment is indicated. Such a designation is not necessarily true. For example, one sample may not typically characterize effluent or ambient water toxicity. More frequent sampling and testing could reveal that toxicity is of sufficient magnitude, duration, and frequency to evoke biological community responses. On the other hand, assuming that the bioassessment measurement reliably demonstrated impairment, the impact could be a consequence of nonchemical, non-wastewater related causes. The presence of bioaccumlative toxic chemicals in a water sample could lead to a false negative designation because short term toxicity tests are not designed to detect such substances. In addition, there are biological endpoints in aquatic ecosystems which are not represented in indicator species toxicity tests.

Following a systematic analysis, Luoma and Ho (1993) concluded that false negative predictions (finding no statistically significant toxicity in laboratory single species tests when, in truth, there is biological community degradation) are just as probable as false positive predictions. Luoma and Ho contend that "false negatives may be common in toxicity tests, but they are difficult to detect. The main reason is that the ecological tests included in many validation studies are insensitive. Typically, validations are conducted only at one point in time, make inadequate replication, consider ambiguous community structure indices, or do a poor job of documenting exposures." Caution should be exercised when describing the relationship between a single toxicity test result and an index (which represents the integration of many types of stresses over time) of ecosystem integrity, as a false positive or negative.

5.0 Field Studies

5.1 CETTP Studies

The eight CETTP and three related studies examined the relationship between 7-d *Ceriodaphnia* and/or larval fathead minnow early life stage toxicity test results on surface water or wastewater and instream survey indices for zooplankton, benthic macroinvertebrates, and/or fish populations. The intent of these studies was to determine

how effectively toxicity test results on ambient waters or effluents corresponded with ("predicted") estimates of aquatic ecosystem community health. In the eight CETTP studies there were 80 sites in eight different watersheds where instream bioassessment indices were compared to surface water toxicity test results.

The intent is not to summarize and evaluate each of these CETTP studies separately since they have, as a group, been the subject of recent analyses (Dickson et al., 1992; Marcus and McDonald, 1992). The approach is to summarize two of the studies (Birge et al. 1989; Eagleson et al., 1990) which have been associated with the CETTP and then the two analyses (Dickson et al., 1992; Marcus and McDonald, 1992) of the CETTP studies. This summary is followed by an evaluation of those two analyses by an independent statistician. The final portion of this section is a review of four CETTP studies which, according to Marcus and McDonald (1992), do not evidence a statistically significant canonical correlation between toxicity test results and instream indices of biological community health.

Marcus and McDonald (1992) make the interesting observation that, "There is, unfortunately, an excess emphasis by many investigators and reviewers on significance in assessing statistical results. The question of primary concern is not whether there is high or low frequency of significant correlations, but what the degree of correlation between pairings of each laboratory and field variable is." Examination of the CETTP studies reveals a distinct qualitative correspondence between ambient water toxicity and ecosystem variables. In most of the CETTP studies there appeared to be biological impairments in a gradient below discharge points (which showed toxic effluents) compared to upstream sampling sites. In most cases, ambient water toxicity at a site was associated with biological community impairments. Because of small sample sizes in the CETTP studies, routine correlative parametric statistics were not applied to compare bioassessment and toxicity test data.

Statistics are frequently used to demonstrate "proof" of effect, the threshold of effects being arbitrary. McBride et al. (1993) conclude that routine application of significance tests does not extract the maximum information from environmental data. These authors discuss the advantages of equivalence tests where the investigator must state what degree of difference is considered a practical difference. In an equivalence test the null hypothesis is that the difference in means is greater than some practically significant value which the tester must state in advance. They recommend that environmental managers and scientists focus attention on statistical power (the probability of rejecting the null hypotheses of no difference in the test groups when in fact it is false--ideally the level of power should be high) and decide what is a practical difference. This practical difference concept could apply to both the bioassessment and toxicity data.

In a book on ecological risk estimation Bartell et al. (1992) write, "It might also be easier to design experiments or to monitor natural systems for qualitative endpoints rather than having to demonstrate statistical differences between quantitative results. The large variances that typify ecological experiments may argue for adopting more qualitative endpoints." Statistics are a valuable tool in our attempt to understand biological and ecosystem operations. As we endeavor to comprehend the biological world, it may be useful, however, to remember that statistical significance does not guarantee biological significance and biological significance does not always equate with statistical significance.

5.2 Associated Studies

5.2.1 South Elkhorn Creek Study

Birge and associates (Birge et al., 1989; 1990) performed ecological assessements on a stream, which received a point-source discharge. Results of single species toxicity tests were compared to ecological endpoints. One objective was to assess the reliability of the laboratory test results in predicting ecological responses. Ecological measurements included macroinvertebrate species richness, abundance, diversity, and functional group analysis. Toxicity in effluent and ambient water samples from the different stream sites was assessed using a fathead minnow embryo/larval 8-d test.

The point-source discharge was from a wastewater treatment plant (WWTP) into Town Branch Creek. Town Branch Creek entered into South Elkhorn Creek about 14 km below the WWTP outfall. There were three control sites, one above the WWTP outfall on Town Branch Creek and two on South Elkhorn Creek above the confluence with Town Branch Creek. There were seven sampling sites at various distances downstream of the discharge point. The most distant station was 67.8 km downstream of the WWTP outfall.

Embryo-larval survival in water samples from all three control (upstream of the WWTP) was greater than 90%. Toxicity tests with WWTP effluent generated data on effect concentrations (expressed as percent effluent). Hydrology of the creek was studied so that percent dilution of effluent could be predicted at each sampling site. Toxicity at sites downstream of the discharge point reflected the toxicity predicted by the effluent toxicity test data considering instream dilution. That is, instream toxicity was reliably predicted by effluent dilution data. These data are significant in that they demonstrate that the major modification of effluent toxicity was stream dilution; physical and chemical characteristics of the stream did not appear to mitigate toxicity to any great extent.

Ambient water samples collected at the three sites immediately below the point of discharge showed statistically significant toxicity, whereas none of the reference sites yielded significant toxicity. A decreasing gradient of toxicity downstream of the discharge point was evident. Both the fish and invertebrate data suggested adverse impacts at the three sites immediately below the discharge point. Below these heavily impacted sites there tended to be a gradient of increasing diversity of both fish and macroinvertebrates downstream of the discharge point. The correlation coefficient (r) between embryo/larval survival in water samples from the stream sites and estimated percent effluent at those sites was -0.87 (i.e., the greater the percent effluent at a site the lower the embryo-larval survival). The number of fish species ® =-0.83) and number of invertebrate taxa r =-0.94) were also inversely correlated with the estimated percent effluent at a site. The correlation coefficients between embryo-larval survival in water samples from the stream sites and number of invertebrate taxa was r = 0.96 while the value for the number of fish species was r = 0.92. All of these correlation coefficients were statistically significant. Results of this study illustrates the laboratory toxicity test results were very reliable predictors of instream biological community responses. Data from this study were included in the statistical analysis by Dickson et al. (1992).

5.2.2 North Carolina Study

Effluent toxicity test results were compared to indices of aquatic ecosystem community health at 43 sites on rivers and streams in North Carolina (Eagleson et al., 1990). Toxicity tests were performed with both municipal waste treatment and industrial facilities effluents. The 7-d *Ceriodaphnia* test, was used to estimate chronic toxicity in effluents. Instream biological responses were gauged by surveys of benthic macroinvertebrates above and below points of discharge. Attempts were made to reduce habitat type confounding factors, as well as other physical confounding factors. Care was taken to compare the results of toxicity tests and field responses at low and average flow conditions; toxicity decay was also incorporated into the comparisons.

Results of this study revealed that, if proper consideration was given to effluent dilution, the USEPA toxicity tests results can be reliable predictors of ecological effects. Comparisons of upstream and downstream sites with regard to biological indices were made with the nonparametric Wilcoxon signed-rank test. If a site downstream of an effluent discharge point was identified as a statistically significant response (degradation) compared to the reference site above the discharge point, the site was classified as "instream impact measured." If there were no differences between upstream and downstream site biological indices measurements, the site was classified as "no instream impact measured." When an effluent sample, diluted to the appropriate instream waste concentration (IWC), resulted in a statistically significant response compared to controls, the sample was designated as "instream impact predicted." If an effluent sample did not produce statistically significant toxicity, it was designated as "instream effect not predicted."

The classification system described in the above paragraph was combined into a contingency table which is best illustrated in Figure 1.

Toxicity test predictions were accurate in 88% of the cases. If non-effluent anthropogenic factors contributed heavily to instream biological impacts, one might expect a high frequency of "false negatives." However, there were only 5% false negatives. If habitat differences or other physical factors between the reference site and the sites downstream contributed to those sites being classified as impaired, one would expect a more equal distribution between the two different categories with impacted sites. However, the distribution in the two categories where instream impact was measured was very unequal (i.e., two tests showing no toxicity and 29 testing positive for toxicity, cf., Figure 1).

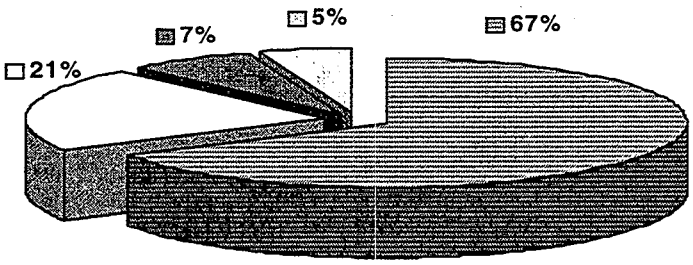
Although these investigators did not apply statistical analyses to this contingency table, results from Fisher Exact and Chi-Square tests showed statistical significance ($P<0.001$). Moreover, one must reject the null hypothesis that toxicity test results do not predict biological responses. Even though there is some potential that confounding factors (see Section 6.6 below) influenced biological measurements, it appears that the dominant impairments were due to wastewater constituents. Assuming that the ecological indicators were accurate, the results of this study provide a strong case that the *Ceriodaphnia* (even though not indigenous to these stream ecosystems) toxicity tests results were reliable qualitative predictors of aquatic ecosystem impairments. A more powerful statistical design would have included more non-impacted sites, but prior to the ecological surveys the nature of sites was unknown.

5.3 Review of CETTP Studies

5.3.1 Dickson et al. Analysis

In 1992 Dickson and colleagues published the results of a study undertaken to statistically analyze data from the eight CETTP studies, the South Elkhorn Creek, Kentucky study (Birge et al., 1989), and a study on the Trinity River, Texas (Dickson et al., 1989). The intent of the Dickson et al. (1992) study was to apply a statistical method and classification approach to all of the above mentioned data to elucidate relationships between surface water toxicity test results and ecosystem community responses.

After entering data from all of the studies listed above into a database, a canonical correlation analysis was performed to examine the relationship between ambient



Toxicity test predicts instream impact; instream survey measures impact [29/43 = 67%]

Toxicity test predicts no instream impact; instream survey measures no impact [9/43 = 21%]

Toxicity test predicts instream impact; instream survey measure no impact [3/43 = 7%]

Toxicity test predicts no instream impact; instream survey measures impact [2/43 = 5%]

Figure 1. Summary of Eagleson et al., 1990. The categories represent the four possible outcomes when comparing laboratory effluent toxicity test results to ecological survey data collected at 43 stations.

water toxicity and estimates of biological community condition. Canonical correlation tests for significant relationships between two matrices of data. The bioassessment metrics can be explored and meshed into a variate for ecosystem condition which in turn is compared to the toxicity variate composed of the toxicity data (e.g., from both *Ceriodaphnia* and larval fathead minnow tests). A goal of canonical correlation is to identify a combination of the predictor variables (i.e., toxicological responses) and response variables (biological community indices) which have the strongest correlation among all possible combinations. The output of canonical correlation includes indicators of the relative importance (sometimes defined as weights) of each variable to the overall correlation.

There were two major goals in the Dickson et al. (1992) study: 1) ascertain whether or not statistically significant correlations existed between the surface water toxicity variable and the biological community variable and 2) use the results of the canonical correlations to identify important variables. Using the toxicity test and biological community indices variables, a classification system was developed to determine the reliability of toxicity test results as predictors of instream community responses.

A major aspect of the analysis was data collected in the Trinity River study (Dickson et al., 1989). In that study the relationship between ambient water toxicity test results and biological community response was scrutinized at 11 sites along the river. Reference sites were located above a WWTP discharge point and the remaining sites were below the outfall. The relationships between ambient water toxicity and biological community indices were examined through time, with sampling and testing in six separate months. Assessments of ambient water toxicity consisted of the *Ceriodaphnia* and larval fathead minnow short-term test estimates of chronic toxicity. Instream biological community assessments included fisheries data (richness, evenness, and an index of biotic integrity) and benthic macroinvertebrate data (richness and evenness).

Separate canonical correlations were performed with the toxicity variable compared to the fisheries indices and to the benthic macroinvertebrate indices; the toxicity variable was not correlated with a consolidated bioassessment variable consisting of combined fisheries and macroinvertebrate indices. Statistically significant ($p<0.001$) coefficients of determination (r^2 represents the proportion of variation in one variable determined by the variation of the other) were observed for both canonical (range of r^2 was 0.38 to 0.59) and robust canonical (range r^2 was 0.38 to 0.94) analyses in all six months of the Trinity River study for the fisheries and macroinvertebrate measurements. These findings imply that the matrix of toxicity test results were effective predictors of instream biological community responses.

Unfortunately, detailed information on canonical correlation for the CETTP studies was not presented by Dickson et al. (1992). In fact, r^2's were presented for only the Five Mile Creek (Mount et al., 1985) and Kanawha River (Mount and Norberg-King, 1989) studies. Data showing the relative contributions (i.e., weights) of each of the toxicological and each of the biological community variables were not presented for the two CETTP studies. Statistically

significant r^2's from the robust canonical correlations were noted for the Five Mile Creek ($r^2 = 0.81$, $p = 0.0005$) and Kanawha River ($r^2 = 0.81$, $p<0.00001$) data. These correlations suggest that toxicity test results from ambient water samples were reliable predictors of instream biological community responses.

Based on the canonical analyses, fish species richness was shown to be an important aquatic ecosystem response variable. Therefore, Dickson et al. (1992) selected fish richness as the ecological response variable in all studies where it was available for the next phase of the analysis. However, two CETTP studies, Kanawha River (Mount and Norberg-King, 1986) and Ohio River (Mount et al., 1986c) did not include fish surveys, so benthic macroinvertebrates richness was substituted as the biological response variable.

The next step in the analysis was to develop a classification system to judge whether or not a site was predicted to be impacted based on ambient water toxicity data, and whether or not a site was observed to be impacted based on instream community metrics.

For the ambient water toxicity data, a low value for test performance (e.g., low *Ceriodaphnia* neonate production or low larval minnow growth) was used to classify a site as "impact predicted" and a high value for test species performance was classified as "impact not predicted." For instream biological community variables a low value (e.g., low species richness) resulted in that site being classified as "impact observed," whereas a high biological community value classified the site as "impact not observed." Rather than establish arbitrary thresholds (cutoffs) for classification of ambient water toxicity results into categories, the natural variability of the measured parameters was incorporated into the system. Because the measure of toxicity consisted of the sum of a subset of the toxicity variables, with each of these variables standardized, and with the assumption that the majority of the observations were normally distributed, the authors reasoned that the sum of a set of these variables should have an approximately normal distribution. Assuming independence of variables, the authors reasoned that the sum divided by the square root of the number of variables being summed should have a standard normal distribution. For these reasons, Dickson et al. (1992) concluded that a classification scheme could be defined such that a site would be classified "impact predicted" if the normalized toxicity measure fell below a threshold obtained from percentiles of the standard normal distribution.

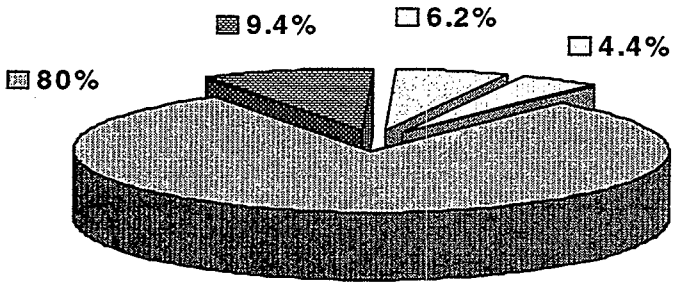
Controversy surrounds the biological metrics and the amount of change or difference in these metrics which represents impairment. Therefore, as with the toxicity test data, Dickson et al. (1992) used the biological community data to determine a classification. Sites that revealed a biological response below a corresponding Poisson distribution percentile (representing counts of the number of fish and invertebrates at a site) were classified as "impact observed."

Two misclassification errors were possible in this scheme, which were 1) misclassifying a nonimpacted site as impacted or 2) misclassifying an impacted site as nonimpacted. The percentiles selected for threshold depends on which misclassification error is of greater concern. If the desire is to keep the error rate of classifying an impacted site as nonimpacted low, then one might select a 95th percentile threshold. To keep the error rate of classifying a nonimpacted site as impacted low, the 5th percentile could be the threshold.

The classification scheme described above produced two-way contingency tables for predicted and observed impacts at aquatic ecosystem sites. Fisher's test was used to evaluate the accuracy of toxicity test predictions of instream impacts. The classification scheme was applied to the CETTP and Trinity River data sets, as well as to the combined data sets. Using both the 95-95 and the 5-5 percentile cutoffs, strong, statistically significant qualitative relationships were demonstrated between ambient water toxicity and instream biological response (impairment). The contingency table for the 95-95 percentile threshold using the combined data sets is reproduced below.

Figure 2 shows the data from a contingency table summarizing the data analyzed by Dickson et al (1992). The total percentage of sites in all of the CETTP and Trinity River studies where toxicity test results reliably predicted instream biological findings was 84.4%. Fisher's Exact test revealed that toxicity test results effectively ($p = 0.003$) predicted instream biological responses. The low percentage (6.2%) of "false negatives" suggests that factors other than toxicity were not major contributors to biological community impacts.

These data can be grouped and examined in a different manner. Grouping by whether sites were biologically impacted or not yields totals of 136 and 22, respectively. For a stronger statistical design, a much larger number of potentially unimpacted sites would be necessary. However, the condition of sites was unknown prior to the biological surveys. Looking only at the impacted sites, ambient water toxicity tests predicted impacts correctly in 93 % of the cases with 7% "false negatives." Examination of the non-impacted site data reveals that toxicity tests were reliable predictors in 32% of the cases, with 68% "false positives." This potential (see discussion on false positives/negatives above) high rate of "false positives" is disturbing and confirms that the results of a single toxicity test should not be used to characterize wastewater or an ambient water toxicity.



Toxicity test predicts instream impact; instream survey measures impact [128/160]
Toxicity test predicts instream impact; instream survey measures no impact [15/160]
Toxicity test predicts no instream impact; instream survey measures impact [10/160]
Toxicity test predicts no instream impact; instream survey measures no impact [7/160]

Figure 2. Summary of Dickson et al., 1992 analysis. The categories represent the four possible outcomes when comparing laboratory toxicity test results on ambient water samples from stream sites with ecological survey data from the same sites. Total number of stream sites is 160.

The procedures used in this study required that gradients of ambient water toxicity and of biological community responses exist. The statistical analyses performed revealed that the frequency of observing instream impairments when toxicity test results predicted an impact was significantly greater than the overall frequency of impairments observed. The analysis by Dickson et al. (1992) provides a compelling qualitative relationship between ambient water toxicity and indigenous species responses. Toxicity test endpoints identified as effective qualitative predictors of aquatic ecosystem responses were *Ceriodaphnia* neonate production and larval fathead minnow growth.

5.3.2 Marcus and McDonald Analysis

Marcus and McDonald (1992) also analyzed the CETTP and Elkhorn Creek (Birge et al., 1989) data sets using canonical correlation. In this analysis, the null hypothesis was that no correlation existed between the matrix of instream biological community measurements and the matrix of toxicity test results (i.e., neonate production by *Ceriodaphnia* and larval fathead minnow growth).

The results of their analysis showed a statistically significant canonical correlation occurred in four of the eight CETTP site studies (Scippo Creek: $r = 0.93$, Naugatuck River: $r = 0.78$, Back River: $r = 0.996$, and Kanawha River: $r = 0.79$) as well as in the Elkhorn Creek (Birge et al., 1989) data set $r = 0.99$). This translates to five of the nine data sets (streams/rivers) analyzed. Relatively high values were found for the canonical correlation coefficients. For all but two of the nine data sets the coefficients indicated a greater than 50% ® > 0.7) relationship between the sets of laboratory toxicity test results and the instream biological variables. Marcus and McDonald (1992) emphasized these high correlation coefficients and downplayed statistical significance. Except for the Naugatuck River study (Mount et al., 1986a), canonical variable weights (weights refer to the relative importance of each variable to the overall correlation) were not shown in this publication, hindering the ability of the reader to interpret the statistical analysis.

Marcus and McDonald (1992) concluded that, "Although future improvements will be made in these test methods, and better methods may be developed, we conclude that at this time these two toxicity test methods (i.e., the *Ceriodaphnia* and larval fathead minnow short term estimates of chronic toxicity) can be potentially useful assessment tools for screening and monitoring."

In the analysis of the CETTP data, Marcus and McDonald (1992) found that the ambient toxicity measures often showed greater relationships to instream biological measurements than expected by chance. They observed that potentially important relationships appeared often. "Our analyses of the CETTP data indicate that results from the tests of ambient water toxicity often contain potentially important biological information about relationships contained in variables of these field variables." (Marcus and

McDonald, 1992). In other words, qualitative relationships appeared often in the CETTP data.

Marcus and McDonald (1992) reported that *Ceriodaphnia* neonate production generally had the greatest incidence of significant correlations to biological community measures (greatest potential for predicting impairments).

Based on simple correlation analysis of the CETTP data Parkhurst (1996) suggested that ambient toxicity did not show a strong relationship with measures of instream biological communities. Nonetheless, a statistically significant relationship was noted between ambient water toxicity and instream biological indices in five of nine CETTP and associated studies. Furthermore, only sublethal endpoints from the toxicity tests were used in the correlation analysis; that is, Parkhurst (1996) omitted lethality data from his analysis.

5.4 Independent Evaluation of Statistical Analyses

The appropriateness of the statistical methods and an evaluation of the major differences used by both Dickson et al. (1992) and Marcus and McDonald (1992) was conducted by Smith (1994). In this review, Smith (1994) was not convinced that the canonical analyses were the optimal statistical approach to examine the CETTP and Trinity River data as canonical correlation assumes linear relationships. Note that Marcus and McDonald (1992) did address the linearity question within and between the sets of variables. Smith suggests that there are many cases where biological community parameters have been shown to have nonlinear relationships with toxicity. Furthermore, canonical correlation focuses on linear combinations of toxicity and instream response variables that correlate maximally. Smith concluded that, "It is possible that the most ecologically meaningful relationships between the toxicity tests and instream responses are not represented by maximal correlations."

As indicated above, Marcus and McDonald (1992) did not provide canonical variable weights in their publication, except for one analysis, rendering interpretation of their appraisal difficult. Dickson et al. (1992) presented canonical variable weights for the Trinity River study, but not for the CETTP studies. Smith observes that examination of the canonical variable weights, especially for two (February and June) of the six months of Trinity River sampling data, call into question the usefulness of the canonical procedure for analyzing the CETTP and associated studies data. More specifically, toxicity test variable weights and bioassessment variable weights sometimes had opposite signs (plus or minus). When both toxicity and bioassessment weights have the same sign (plus, plus or minus, minus) the data indicate that increased larval fathead survival/growth and/or *Ceriodaphnia* survival/neonate production were correlated with greater diversity/density in the biological community measurements. However, when the signs were opposite, the indication is that an increase in one variable was accompanied by a decrease in the other variable (e.g., higher growth/reproduction with lower diversity/density). Biologically, the opposite signs appear inconsistent with the expected relationship between toxicity and community parameters.

Smith suggests analyzing the CETTP and associated studies using separate analyses of the toxicity and instream response data, possibly ordination techniques. He commented that, "These analyses would provide insight into the relationships and patterns shown by toxicity data alone and the instream response data alone. From these analyses, I would produce interpretable variables summarizing the different patterns observed (for both toxicity and response variable sets separately). I would then correlate the summary variables for the toxicity tests with the summary variables for the instream responses using multiple regression. Each regression analysis would involve an instream response summary variable as the dependent variable, and the toxicity test summary variables as the independent variables. Bivariate plots would be useful to see the nature of the relationships and determine if nonlinearity needs to be taken into account when using the analytical tools."

Assuming that the variables used to classify impairments were sufficient, Smith indicated that the conclusions made by Dickson et al. (1992) from their classification system were reasonable.

Data from only two CETTP studies were analyzed in common by the two different groups of investigators. Both sets of authors reported statistically significant canonical correlations (but the correlation coefficients differed) for the Kanawha River data set. Dickson et al. reported a statistically significant canonical correlation coefficient for the Five Mile Creek data set, whereas Marcus and McDonald (1992) did not. The differences between the two analyses probably related to the fact that Dickson et al. (1992) did not "mesh" the bioassessment fish and macroinvertebrate data whereas Marcus and McDonald did.

From Smith's review, it is clear that there are various ways to statistically analyze studies which attempt to examine the relationship between toxicity test results and instream biological community responses. Data can be used or grouped in various arrays such that the outcome of an analysis can be very different.

5.5 Review of CETTP Studies in Which A Significant Correlation Was Not Observed

In Marcus and McDonald (1992) canonical analysis, four of the CETTP studies, Ottawa River(Mount et al, 1984), Five Mile Creek (Mount et al, 1985); Skeleton Creek

(Norberg-King and Mount, 1986), and Ohio River (Mount et al., 1986c) did not produce a statistically significant correlation between ambient water toxicity test results and instream biological community parameters. The nature of this type of analysis can obscure valuable pieces of data, as well as informative observations. For this reason, instructive aspects of these four studies are summarized below as the studies provide very useful information which was not revealed by any canonical correlation analysis.

5.5.1 Ottawa River Study

The CETTP Ottawa River study included three discharges. The most upstream was a sewage treatment plant (STP), next was the refinery, and the last discharge was a chemical manufacturing plant. Outfalls from all of these facilities were within a 1.3 km range on the river. Ecological surveys were performed twice (1982 and 1983) at nine different sites on the river. Two sites were upstream of the three outfalls. Sites 2 and 3 were immediately above, and below the STP outfall, respectively. Sites 3 and 4 were immediately above and below the refinery outfall, respectively. Sites 4 and 5 were immediately above and below the chemical plant outfall, respectively. Sites 6, 7, 8, and 9 were approximately 6.8, 13.1, 31.7, and 57.8 km downstream of the chemical plant outfall. If the discharges from these plants were impairing the Ottawa River ecosystem, one would expect that sites 3, 4, 5, and perhaps 6 would appear degraded compared to sites 1 (above discharges) and 9 (most distant downstream). The STP discharge contributed approximately 72 to 82% of the Ottawa River flow. The refinery and chemical plant effluents contributed about 17 to 32% and 8 to 10% of the river flow, respectively.

Seven day early life stage toxicity tests with larval fathead minnows and *Ceriodaphnia* were performed on effluents from the three facilities and on surface water samples collected at the nine river sites. Benthic macroinvertebrate and fish population indices were used to assess instream biological community condition. Table 1 summarizes the toxicity testing data from the Ottawa River tests. Examination of the benthic macroinvertebrate diversity, community loss index, and dominant taxa data suggest that the sites immediately downstream of the outfalls were impaired compared to sites 1 and 9. Sites 3, 4, 5, and 6 appeared the most impacted; recovery was not evident until site 8, (31.7 km downstream of the chemical plant outfall).

The fish population data generally agreed with macroinvertebrate results. In the 1982 sample, sites 4, 5, 6, and 7 were characterized by zero to 8 species and a total of approximately 50 individuals at all four sites. At all the other sites there were 11 to 18 species with the number of individuals in the thousands. For the 1983 sampling, sites 3, 4, 5, and 6 appeared to be the most impacted with 1 to 5 species and low total counts. The remainder of the sites were characterized by 10 to 23 species and high total counts.

Results from this investigation revealed a qualitative correspondence between effluent and ambient water toxicity, as

Table 1. Toxicity testing summary for the Ottawa River study (Mount et al., 1984)

Effluent	
	Larval Fathead Minnow Tests
STP:	No significant toxicity 1982, 1983
Refinery:	Significant toxicity in 50% effluent 1982; 100% effluent 1983
Chem. facility:	No significant toxicity in 1982; significant toxicity in 1% effluent in 1983
	C. dubia Tests
STP:	Significant toxicity in 10% effluent 1983
Refinery:	Significant toxicity in 10% effluent 1983
Chem. facility:	No significant toxicity
Ambient Water Toxicity	
	Fathead Minnow Test
	Significant toxicity at sites 3 through 8 compared to sites 1 and 9 (1983)
	C. dubia Test
	Significant toxicity at sites 3 through 6 compared to sites 1 and 9 (1982); Significant toxicity at sites 3 through 7 compared to sites 1 and 9 (1983).

well as between ambient water toxicity and ecosystem responses at sites downstream of effluent discharge points. Although statistical analyses were not performed, there was distinct correspondence between ambient water toxicity and biological community responses.

5.5.2 Five Mile Creek Study

The Five Mile Creek study included three dischargers, two coke plants and a WWTP. Nine study sites were located along the creek: two sites above the first point of discharge and the remainder downstream of at least one discharge point. Coke plant #1 outfall was 2.3 km upstream of coke plant #2 and in turn , coke plant #2 was 10.7 km upstream of the WWTP outfall.

The 7-d larval fathead minnow and *Ceriodaphnia* toxicity tests were used to assess toxicity in ambient water samples from each of the sites, as well as in effluent samples. Ecological surveys were conducted at all sites in February and October; effluent and ambient water toxicity tests were also performed during these months.

The effluent LOEC for the larval fathead tests was 1% and 3% in October and February, respectively for coke plant #1. The coke plant #2 effluent LOEC for October and February was 30% and 10%, respectively. In February, significant toxicity was seen in the larval fathead test at the two sites below the coke plants. *Ceriodaphnia* tests conducted during February were not reliable because of problems in the culture population. In October, the LOEC in *Ceriodaphnia* tests was 10% effluent for coke plant #1 and 30% effluent for coke plant #2. In general, the WWTP effluent appeared to contain little toxicity in either of the toxicity tests.

There were fewer benthic macroinvertebrate taxa and lower density at sites immediately below the coke plant outfalls, but the data were not conclusive. Likewise, no consistent pattern was seen in zooplankton data. At sites above the coke plant outfalls 4 to 6 (total count >100) and 10 to 12 (total count >1,600) fish species were counted in February and October, respectively. In February, 0 to 2 (total count = 9) fish species were noted at the sites immediately below (500 m) the outfalls of the coke plants. In October, 1 to 8 (total count = 83) fish species were counted below the outfalls of the coke plants. This paucity of fish species and numbers of individuals below the coke plants discharge points suggest that those effluents adversely affected fish life.

To understand this study it is important to note that the two coke plant effluents contributed a relatively low percentage of stream flow--less than 1% for plant #1 and usually not more than 8% for plant #2. In other words, the instream waste concentration (IWC) for these discharges was fairly low. Based on the results of the effluent toxicity tests, acceptable waste concentrations (AECs) were calculated for the effluents of both coke plants. Comparison of the IWCs to AECs revealed that the AEC seldom exceeded the IWC with the exception of the sites immediately downstream of the two outfalls. These were the sites where there was a paucity of fish species and numbers. While the coke plant effluents (as well as ambient water) toxicity qualitatively predicted ecosystem impairments, the effluent tests tended to "underestimate" fish population impairment. That is, some would suggest the effluent toxicity tests yielded "false negatives"!

That the canonical correlation analysis of Marcus and McDonald (1992) did not "recognize" the specifics described above is not surprising since their analysis consolidated toxicity testing as well as bioassessment data. The only significant impairments in Five Mile Creek appeared to be on fish populations immediately downstream of coke plant discharges. In the Dickson et al., 1992 study (where fishery data were correlated with the consolidated toxicity data) a statistically significant canonical correlation coefficient between toxicity and bioassessment data was noted in the Five Mile Creek study.

5.5.3 Skeleton Creek

The Skeleton Creek study consisted of ten sites where ecological surveys (Norberg-King and Mount, 1996) were performed and water samples collected for toxicity analysis. Sites were on Skeleton Creek or its tributary, Boggy Creek. During the study there were two major discharges, a refinery on Boggy Creek and a fertilizer manufacturing facility on Skeleton Creek a short distance below the confluence with Boggy Creek. On Boggy Creek there were two sampling sites above the refinery discharge point and one 200 meters below this point. On Skeleton Creek there was one site above the confluence with Boggy Creek and, thus, above the fertilizer plant discharge point. There were also sites immediately below the confluence with Boggy Creek and 300 meters below the outfall of the fertilizer plant. Five sites were at various distances downstream of the fertilizer plant on Skeleton Creek.

The 7-d larval fathead minnow and *Ceriodaphnia* toxicity tests were used to assess effluents, as well as ambient water samples collected at each of the sites. Larval fathead minnows were more sensitive than cladocerans to the effluents from both the refinery and the fertilizer plant. Ten percent effluent from both facilities yielded statistically significant larval fathead responses.

Statistically significant larval fathead minnow mortality was seen only in the ambient water sample collected immediately below the fertilizer plant outfall. Statistically significant larval fathead minnow growth inhibition was seen only in ambient water samples collected immediately below the outfalls of the refinery and fertilizer plant outfalls.

These toxicity test results were consistent with the fish population data; the site immediately below the fertilizer plant was the only station where there were no fish.

The fact that there were no fish collected at the site below the fertilizer plant and that the ambient water sample from this site was the only ambient sample to cause significant larval fathead mortality (effluent from this facility was also the most toxic to larval fish) suggests an effective correspondence between toxicity test results and instream biological responses. That this relationship was lost in the canonical matrices (Marcus and McDonald, 1992) of toxicity and bioassessment metrics is not surprising.
In a study done at the same time as USEPA's, Burton and Lanza (1987) reported that microbial assays revealed toxicity in ambient waters below the two discharge points and that these toxicity results were inversely correlated with instream biological community data.

5.5.4 Ohio River
In this study, a 12 km segment of the Ohio River was investigated. Within the study area there was a steel mill with multiple outfalls and a WWTP. This study included eight sampling sites. One site was located above the steel mill and WWTP outfalls. Other sites were situated immediately upstream and downstream of the outfalls. The last river site was approximately 2.5 km downstream of the last steel mill outfall.

Planktonic and benthic macroinvertebrate data were collected at each site only once. Ambient water samples from these sites were tested only once with the 7-d larval fathead minnow and *Ceriodaphnia* toxicity tests; effluents were not tested.

None of the surface water samples yielded significant toxicity to *Ceriodaphnia* in a 7-d test. The larval fathead minnow toxicity test results were variable and inconsistent. Examination of the plankton data revealed little correspondence to points of discharge. The benthic macroinvertebrate data indicated possible impacts only at sites immediately below steel mill outfalls. These potential impacts were not predicted by the *Ceriodaphnia* or larval minnow toxicity tests. If the instream biological responses were ecologically meaningful, they were underestimated (i.e, yielded false negatives) by the USEPA toxicity tests.

Given the above observations, it is not particularly surprising that the canonical correlations (Marcus and McDonald, 1992) did not identify a statistically significant relationship between toxicity test results and instream biological measurements, because neither varied greatly. Therefore, failure to find a significant canonical correlation in this study should not be used to discredit the USEPA toxicity tests, since there was little gradient in either the toxicity or biological community variables.

5.5.5 General Comments Regarding the Four CETTP Studies Summarized
After reviewing the four CETTP studies in which the Marcus and McDonald (1992) canonical correlation did not find a statistically significant correlation between a matrix of toxicity test results and a matrix of bioassessment metrics, it is not surprising that a statistically significant relationship was not identified. Moreover, in three of the studies, consolidation of the data obscured clear relationships between effluent/ambient water toxicity and instream measurements. Furthermore, Marcus and McDonald (1992) argued against placing too much value on the use of statistical significance and emphasized the high correlation coefficient values identified in their analysis of the CETTP and associated studies data. In the study, significant toxicity was infrequent and differences in instream parameters were minimal--not ideal for demonstrating statistically significant correlations. It would be incorrect to suggest that these four CETTP studies described above constitute evidence that the USEPA toxicity tests are unreliable qualitative predictors of instream biological community responses.

6.0 Criticisms of CETTP and Associated Studies
A group of authors (Parkhurst et al., 1990; Marcus and McDonald, 1992; Parkhurst, 1995, 1996) has criticized the CETTP and associated studies. The criticisms generally relate to design and analysis considerations, most of which are stated below. These publications consist of criticisms of the CETTP and associated studies and do not provide additional data regarding the predictiveness of the USEPA toxicity tests results. Moreover, empirical evidence which suggests that the USEPA toxicity tests are not reliable qualitative predictors of instream impairments has not been provided. Criticisms of the CETTP studies are stated and discussed below.

6.1 CETTP Studies Compared Ambient Water Test Results with Bioassessment Variables
A major criticism of the CETTP studies was that comparisons were made between ambient water rather than effluent toxicity test results and biological community responses. The implication appears to be that abiotic and biotic factors other than dilution can mitigate effluent toxicity. Parkhurst (1995) suggests that a missing link in these studies was to connect surface water with effluent toxicity.

Discussion: Effluent toxicity was measured in seven of the eight CETTP studies and, although statistical correlations were not performed, effluent toxicity corresponded with ambient water toxicity and ecological responses. This criticism fails to recognize that the most probable cause (critics point this out, see Section 6.6) of toxicity in the streams/rivers investigated was discharged effluents.

In the seven CETTP studies where effluent toxicity was measured, ambient water was not significantly toxic at sites above discharge points (or it was less than below discharge points). Where effluent toxicity was noted, ambient water toxicity was generally seen at sites below the discharge point when dilution was taken into consideration. Furthermore, in most of the seven CETTP studies when effluent toxicity was identified there tended to be gradients (i.e., greatest toxicity immediately below discharge points, with progressively lower levels of toxicity at sites downstream) of ambient water toxicity below points of discharge. Also, where there was effluent toxicity there was generally evidence of instream impairments below the discharge points when dilution was taken into consideration. Although statistical correlations were not performed between effluent and ambient water toxicity (or instream biological measurements), it seems that effluent was responsible for ambient water toxicity and ambient water toxicity was the major cause of instream impairments.

6.2 Nonrandom Selection of Study Areas and Sites

Another major criticism of the CETTP studies is that study areas and sampling sites were not selected randomly. Because of this, the contention is that findings cannot be extrapolated using statistical-based induction to other aquatic ecosystems and, secondly, there was not a strong statistically based experimental design. A corollary to this criticism is that USEPA intentionally selected rivers and streams where there were likely to be water quality problems caused by discharged effluents.

Discussion: This criticism has merit and should be considered when evaluating the CETTP data. Design of the CETTP studies was not perfect from a statistical analysis standpoint. More upstream (control) sites would have been desirable. Some argue that all sites below a discharge point represent pseudoreplicates. However, as a practical matter limited funds and other resources require regulatory agencies to focus on areas where there are likely to be environmental problems so there can be remediation and restoration. Indeed, the idea was to study streams potentially impacted by effluent toxicity. It has not been the focus of regulatory agencies to study areas which are pristine or which have a low probability of water quality problems. Moreover, the intent of the CETTP studies was to examine the relationship of probable effluent toxicity and potential instream toxicity, as well as biological community responses.

Random selection of study areas would have resulted in investigations of rivers and streams where there were no discharges and possibly waterbodies known to receive effluents free of toxicity. A recommendation has not been advanced as to the number and types of aquatic ecosystems which should be studied before a consensus can be achieved on the effectiveness, or lack thereof, of single species toxicity test results in predicting qualitative ecosystem responses. USEPA (1991) suggests that it is reasonable to assume that in the *absence of data showing otherwise* the relationship between ambient water toxicity and aquatic ecosystem impacts is independent of waterbody type.

Random selection of sites on a stream would result in confounding factors. For comparison among sites or to a reference site, all sites should be equivalent, including physical/chemical habitat and substrate; with this control the major variable would be the potential of chemical toxicity from point or nonpoint sources. Random selection of sites could also introduce the confounding factor of non-chemical, anthropogenic effects on biotic communities.

The criticisms that more sites (controls) upstream of discharge points were necessary, that more non-impacted sites were necessary for an acceptable statistical design, and that all sites below discharge points were pseudoreplicates have some merit, but also disregard some facts and observations. In design of the CETTP and associated studies, it was unknown whether or not sites downstream of discharge points would show ambient water toxicity; whether or the ecological surveys would indicate whether these sites were impacted or not also was unknown. While, from a purely statistical standpoint, the sites downstream of discharge points could be considered pseudoreplicates, this criticism fails to recognize that in a majority of the CETTP and associated studies there were progressive gradients of decreasing ambient water toxicity below discharge points which corresponded with progressive gradients of "improvements" in biological community indices.

The criticism that there was limited statistical correlative analysis in the original CETTP publications is valid. However, as indicated above, this was a consequence of relatively small sample sizes (i.e., number of sites in each study). This statistical analysis criticism has been addressed in part by the Dickson et al. (1992) and Marcus and McDonald (1992) analyses. As indicated in Section 5.4 above, there are other ways that the CETTP data could be grouped and statistically analyzed.

6.3 Use of the Most Sensitive Toxicity Test Results

Marcus and McDonald (1992) called attention to the use in two CETTP studies (Norberg-King and Mount, 1986; Mount et al., 1986) of data from the most sensitive of two toxicity tests to relate with the most sensitive bioassessment measurements.

Discussion: The USEPA procedure has biological and statistical limitations, however, it also has some logic from an ecological perspective. Because sensitivities of different test organisms vary with the toxic chemical or combination of chemicals, the occurrence and combinations of toxic chemicals can vary along a stream, and assemblages of organisms change along a stream, it seems ideal to test with a suite of species and then relate these data to instream biological community variations. Likewise, different components of the instream communities are likely to respond to different chemicals or combinations of chemicals.

The limited responses (only two USEPA toxicity tests) tested in the laboratory toxicity tests compared to the multiple responses in aquatic ecosystems necessitates that all possible relationships be explored. Therefore, while recognizing the limitations of using maximum responses, they may provide insights into interactions of toxicity and community responses. Because of the extremely limited number of species and biological endpoints represented in the USEPA toxicity tests there has been a tendency for regulatory conservatism (use of results of the sensitive species). Whether or not this conservatism is completely justified remains to be determined; however, the results of this review show that laboratory single species tests more frequently yield reliable predictions, or underestimates, of biological community responses than overestimates of impacts.

6.4 Relationship Between Toxicity Test Results and Instream Biological Measurements Relied Heavily on High Magnitude Toxicity.

Another criticism of the CETTP conclusions is that the correspondence between ambient water toxicity and ecosystem community impairments relied extensively on areas and sites where toxicity was relatively high.

Discussion: There is merit to this criticism, but the overall significance is uncertain since toxicity theory is based on a concentration-response relationship (i.e., a greater response with higher toxicity). There should be no surprise that higher levels of toxicity (enough to cause lethality) in ambient or effluent water samples can yield measurable responses in ecosystem parameters. Furthermore, biological responses, as all measurements, are less reliable near detection limits. "False positives" are of greater concern in situations where surface water of effluent toxicity is relatively low and near detection limits. The ability to reliably detect biological community impairments when the concentrations of toxic chemicals are near the effect thresholds is difficult; detection of such impairments also will be obscured by the complexity and natural variability in aquatic ecosystems. It should be emphasized that, in the CETTP studies, toxicity test "predictions" were based on effects (including sublethal) in the 7-d early life stage tests.

6.5 Temporal Repeatability of the Ambient Water Toxicity/Biological Response Was Not Demonstrated

The CETTP studies did not confirm through time the correspondence of surface water toxicity with instream biological variables.

Discussion: There is some validity in this criticism, yet there can be wide temporal variations in effluent and ambient toxicity. Temporal variations in the relationships between toxicity and biological community parameters were considered in some of the CETTP and associated studies (Dickson et al., 1989; Mount et al., 1984; Mount et al., 1985). Defining the magnitude, duration, and frequency of effluent/ambient water toxicity is important. Understanding natural seasonal variations in aquatic biological communities is essential when attempting to relate these variables to potential controlling factors. Significant variations in stream flow and physico/chemical factors can also influence the relationship between effluent toxicity and biological community responses and must be considered in describing a temporal relationship. Failure to demonstrate a statistically significant correlation between effluent/ambient water toxicity throughout the year does not discount the possibility of ecosystem impairments from toxic chemicals (from point or nonpoint sources) during portions of the year. The issue of temporal repeatability of the relationship between effluent or ambient water toxicity and biological community responses has been addressed by Dickson et al. (1989, 1996).

6.6 Confounding Factors Were Not Considered

Parkhurst and associates (Parkhurst, 1995, 1996; Parkhurst et al., 1990) suggested that several factors other than ambient water toxicity could have affected biological community, but were not considered in the CETTP studies. They contend that both natural (e.g., poor habitat, low oxygen, nutrient enrichment, organic enrichment, natural seasonal variations) and non-effluent, anthropogenic factors could have been responsible for biological community changes in the CETTP studies.

Discussion: While contending that confounding factors were not considered, these authors also point out that discharged effluents were the most probable cause of water quality problems. If their confounding factors theory is correct one would expect a high percentage of "false negatives" (toxicity test results predict no instream impact, but impact measured) in the CETTP and associated studies. However, "false negatives" were noted in only 6.3% of the 160 sites in the CETTP and associated studies.

Irrespective of potential confounding factors, statistically significant canonical correlations were seen between ambient water toxicity test results and biological community

responses (Dickson et al., 1992; Marcus and McDonald, 1992). The criticism of confounding factors appears to disregard the CETTP and associated studies observations which revealed impairments on a progressive gradient below effluent discharge points (i.e., the greatest impairments were at sites nearest the discharge point, decreasing with distance from the discharge point). The argument regarding confounding factors makes little biological sense given that CETTP sites upstream of discharge sites generally indicated "healthy" communities, whereas sites below discharge points (which showed toxic effluents) tended to suggest impairments.

The high frequency of accurate predictions in the Dickson et al. (1992) classification system of instream biological responses based on toxicity test results in the CETTP and associated studies is rather surprising given that these relationships were based on the results of single, or few, toxicity tests with a single bioassessment indices (which tends to be temporally integrative, but which does not incorporate natural variations).

6.7 Was the CETTP Classification System Mathematically Biased?

Marcus and McDonald (1992) criticized the procedure used in some of the CETTP studies for identifying correct predictions of biological impairments based on toxicity testing data.

Discussion: This criticism appears accurate. No consistent method was used throughout the CETTP studies to select correct and incorrect predictions. Based on these CETTP comparisons, some studies concluded that the degree of toxicity was related to the degree of instream taxa reduction. The analysis of the data using various analyses appears to have been an attempt to convert a qualitative relationship between toxicity test results and instream biological responses to a quantitative one.

6.8 High Rate of False Positives

Parkhurst (1992) suggested that the rate of "false positives" (toxicity test results predict instream impact, but no impact observed) in the CETTP, South Elkhorn Creek (Birge et al., 1989), and Trinity River (Dickson et al., 1989) studies was 68% and 23% in the North Carolina (Eagleson et al., 1990) study.

Discussion: Using all available data the actual rates of "false positives" were 9.4% and 7%, respectively in the CETTP/Associated studies and the North Carolina study. Parkhurst values are based on only a portion of the data collected in all of the studies, the sites identified as not impacted. While there may be some value in the approach presented by Parkhurst (1992), it certainly ignores a very large portion of the data collected.

6.9 Miscellaneous Criticisms

Some criticisms of the CETTP studies do not relate directly to those investigations. These criticisms include:

- The size and assimilative capacity of the receiving waterbody is not considered when evaluating WET test results,
- the duration of exposure in aquatic ecosystems, relative to test duration, is not considered in the evaluation of WET test results, and
- actual effluent dilution and flow conditions are not usually considered in the evaluation of USEPA toxicity tests results.

Discussion: Some of these criticisms have merit, yet these criticisms are less concerned with the reliability with which USEPA toxicity tests results predict ecosystem responses than with concern that the results of single (or few) toxicity test results could be used as evidence of an effluent permit violation (i.e., they represent potential implementation problems). Certainly, such factors must be considered and incorporated into risk assessments.

6.10 Conclusions

The CETTP studies suffered from some design and interpretive problems. However, even critics of the CETTP and associated studies tend to agree that there is a good qualitative relationship between USEPA toxicity test results and aquatic ecosystem community responses. These critics correctly assert that a quantitative relationship has not been established. Although critical of the CETTP and associated studies, Parkhurst et al. (1992) accept that these studies demonstrate that, if adequate consideration is given for effluent dilution, USEPA toxicity tests results should be reliable predictors of ecological impairments. What appears to be lacking in the criticisms of the CETTP studies are: 1) experimental data which indicate that single species (EPA toxicity tests) test results are more frequently unreliable rather than reliable predictors of ecosystem impacts. and 2) suggestions for effective alternatives to the single species tests.

Recognizing that ecosystems are complex and multivariate, with many interacting factors and that sample sizes were rather small, it is not surprising that the CETTP and associated studies did not establish a quantitative relationship between USEPA toxicity tests results and biological community responses. However, the qualitative association established was convincing enough to accept the results as predictive of probable biological impacts. If a series of ambient water or effluent water tests produce statistically significant toxicity in the USEPA toxicity tests, some degree of ecosystem impairment is likely. Since the USEPA toxicity tests provide an early warning and are predictive of probable aquatic ecosystem impairments, it is not essential that they be highly quantitative predictors of biological community impacts.

7.0 Single Species Tests with Effluent

Investigations in which effluents were tested with single species toxicity tests and in which some ecological survey data were collected from the receiving stream for comparative purposes were reviewed. A summary of these reviews is presented in Appendix A. Studies reviewed in this Appendix, as well as in Appendices B and C were located through literature searches. All studies related to the topic were reviewed, none were screened out. These studies represent a special concern because of the criticism related to the correspondence between single species toxicity test results and ecosystem responses.

Appendix A summarizes 13 publications and the tabulations presented below are by study (i.e., by the outcome of the entire study, not by subcomponents within studies). In nine (69%) of the 13 studies early life stage test NOEC/LOEC s from effluent tests provided reliable qualitative predictions of instream impairments. In three (23%) studies early life stage effluent test NOEC/LOECs underestimated instream responses. Results from one study was inconclusive, consequent to study design and interpretive inconsistencies. Based on effluent toxicity test results no overestimations of instream impacts were noted in these 13 studies.

The 13 studies summarized in Appendix A, as well as the Eagleson et al. (1990) study discussed above, demonstrate that single species toxicity test results on effluents can provide reliable qualitative predictions of biological community responses or tend to underestimate ecosystem impairments.

8.0 Single Species Tests with Individual Chemicals or Small Groups of Chemicals

Studies in which single species toxicity tests were used to assess the toxicity of a single chemical or a small combination of chemicals and predict aquatic ecosystem biological responses were evaluated and summarized in Appendix B, which is subdivided into sections on pesticides, other organic chemicals, metals and miscellaneous substances.

8.1 Organic Chemicals: Pesticides

Eighteen studies dealing with pesticides are summarized in Appendix B. The most studied pesticide in this group of investigations is the organophosphorus insecticide chlorpyrifos (seven studies). In 14 (78%) of the 18 studies, single species laboratory toxicity test results reliably predicted direct field adverse effect concentrations. In many of the studies the single species laboratory tests failed to predict the secondary (indirect) effects seen the field experiments, such that biological community effects were underestimated by the laboratory single species toxicity test results. In four of the studies reviewed in Appendix B the laboratory single species toxicity test effect concentrations overestimated the field effect concentration (i.e., the laboratory single species data underestimated the biological community responses). Although use of daphnids in laboratory tests has been criticized by some because they are indicator, rather than resident species, data in 12 of the 18 studies suggest that daphnids are reliable (or tend to underestimate aquatic ecosystem impacts) predictors of a biological community response.

8.2 Organic Chemicals: Nonpesticides

Eleven investigations of organic chemicals were reviewed and summarized in Appendix B. Laboratory single species toxicity tests results were reliable predictors of biological community effect concentrations in seven (64%) of the eleven studies. In most of these six studies in which laboratory effect concentrations were considered reliable predictors, single species test results were somewhat higher than the field effect concentrations (i.e., biological communities were somewhat more sensitive to chemicals than predicted by the laboratory tests). Laboratory toxicity tests overestimated field effect concentrations in two (18%) studies. Results of two studies were inconclusive or mixed.

8.3 Metals

Ten studies dealing with metal toxicity are reviewed in Appendix B. Results of five (50%) of the ten studies suggest that laboratory single species test effect concentrations are reliable qualitative predictors of biological community effect concentrations and responses. In four (40%) of the studies laboratory single species effect concentrations were notably higher than effect concentrations (i.e., laboratory single species tests underestimated aquatic ecosystem impacts). One of the ten studies was inconclusive.

8.4 Other Data and Views of Predictiveness of Single Species Test Results

Persoone and Janssen (1994) submitted that environmental factors may notably modulate toxicity (e.g., alter bioavailability) as measured in laboratory tests. A majority of the studies, with the exception of investigations on metals, summarized in Appendix B do not support that claim. Speculations that laboratory toxicity test results estimate effect concentrations (e.g., LOECs, NOECs) that are considerably below instream effect concentrations have been voiced, but most of the data reviewed herein fail to support those conjectures.

La Point (1994) concludes that direct, but not secondary, responses of fish in ecosystems can be predicted from laboratory single species test results. Luoma (1995) suggests that accurate predictions of metal impacts based on single species test results are rare. Luoma (1995) also wrote, "As toxicity tests are increasingly used in contaminant management, reliance on insensitive

Table 2. Equations showing relationships between laboratory (single species) and ecosystem determined endpoints (data from Slooff et al., 1986)
Using acute toxicity data the following equation was derived: log NOEC(ecosystem) = -0.55+0.81 log LC50(single species tests). In this case, n = 54, r = 0.77, and the uncertainty factor was 85.7.
Using chronic toxicity data the following equation was derived: log NOEC(ecosystem) = 0.63+0.85 log NOEC(single species tests) In this case, n = 51, r = 0.85, and the uncertainty factor was 33.5.

procedures dominated by type II error (false negative) will lead to regulations that underprotect nature." Luoma (1995) listed the uncertainties in single species tests which result in underestimation of impacts due to metals on biological communities. These sources of uncertainty include:

- choice of species (sensitive and ecological keystone species unrepresented),
- exposure time (underestimated),
- exposure route (rarely considered),
- multigenerational life cycle (unrepresented),
- higher-order secondary effects (rarely considered), and
- interaction with natural disturbances (rarely considered).

Margins of uncertainty in predicting toxicity from laboratory single species tests to higher levels of biological organization were determined by regression and correlation analyses (Slooff et al., 1986). Analyses were performed on log-transformed data. The 95% uncertainty factors were determined as the minimum ratio of the estimated toxicity value and its upper and lower 95% confidence (prediction) limits.

The regression analysis consisted of regressing ecosystem-determined effect concentrations on laboratory single species toxicity test effect concentrations. The uncertainty factor was defined as the minimum ratio of the estimated effect concentration and its 95% prediction limit. So, the smaller the value of the uncertainty factor, the more reliably would single species toxicity test results predict biological community effect concentrations.

Using acute toxicity data for 34 chemicals, the following relationships in Table 2 were determined. Slooff et al., (1986) concluded that data from laboratory single species toxicity tests are reliable enough for ecological risk assessments.

The studies summarized in Appendix B suggest that laboratory single species test results afford a reliable qualitative prediction (are reliable for extrapolations) of aquatic biological community responses or of environmental effect concentrations. Tabulation of the 47 studies (tabulation is by outcome of the entire study) reviewed in Appendix B yields the results presented in Table 3.

Single species toxicity test results usually provide enough information to take action. These tests can be used to determine concentrations of chemicals in a water sample are sufficient to affect biological functions. Subsequent action can be taken to determine the chemicals causing toxicity and/or the persistence and magnitude of the toxicity in the effluent or the water body. Clearly, the results of a single toxicity test should not be equated with ecosystem impairment; a test result is not *de facto*, definitive proof of biological impairment.

Table 3. Summary of studies examining the relationship between laboratory single species test results and aquatic ecosystem responses (Appendix B).	
Laboratory single species effect concentration provides reliable prediction of biological community effect concentration and/or responses	68%
Laboratory single species effect concentration > field effect concentration (single species test underestimates biological community responses)	23%
Mixed or inconclusive results.	9%

9.0 Comparison of Single Species and Multiple Species (Microcosm, Mesocosm) Toxicity Test Results

Intuitively one might suspect that single species toxicity test results would not predict biological community responses as reliably as multiple species (this term is used to include both micro- and mesocosm studies) test results.

Direct comparisons have not been frequent, but five groups of authors (Slooff, 1985; Emans et al., 1993; Okkerman et al., 1993; Persoone and Janssen, 1994; Dorn, 1996) have published literature reviews which address this issue.

9.1 Okkerman et al. (1993)

Results from NOECs from single species and multiple species tests were compared by Okkerman et al. (1993) in an endeavor to gain insight into whether aquatic ecosystems can be protected by setting a "safe" concentration derived from single species toxicity test results compared To achieve this, Okkerman et al. (1993) performed an extensive literature search to locate all available multiple species studies. These studies were then put through rigorous criteria to identify the multiple species studies considered to be reliable. Some important criteria were that a study had to include several taxonomic groups in fairly realistic ecosystems, the concentration of the chemical had to be analytically verified, and a concentration-response relationship had to occur. NOECs from the multiple species studies were for direct effects only.

For those compounds where a multiple species NOEC was considered reliable, the authors searched for single species tests with an NOEC they considered reliable. Data were sufficient and reliable enough to make the multiple species and single species comparison for only ten organic compounds, most of them were pesticides. When more than one single species NOEC was available, the comparison was made using the value for the most sensitive species. The comparison was the ratio of the multiple species and single species NOECs. The closer the ratio was to one, the less divergent the single species and multiple species NOECs.

For all ten chemicals, the ratio was five or less; for six of the compounds, the ratio was approximately one or less than one; for the remaining four chemicals the ratio ranged from 2.5 to 5. These investigators concluded that despite the general concept that effects assessments should be conducted in actual aquatic ecosystems or multiple species tests, in only a few cases did NOECs differ greatly between single species and multiple species tests. They also surmised that with some caution, due primarily to a paucity of data, single species toxicity test data are a good starting point for establishing "safe" concentrations for aquatic ecosystems.

9.2 Emans et al. (1993)

The accuracy of extrapolating from single species toxicity test results to aquatic ecosystem communities also was examined by Emans et al. (1993). Their approach was to compare NOECs derived from multiple species field studies with those from single species toxicity tests. If field multiple species toxicity test results were more reliable predictors of how biological communities would respond to a chemical(s) than single species test results, then one would suspect that "safe" concentrations generated from these two different procedures would differ appreciably.

After an extensive literature search, acceptable data for the comparison of single species and multiple species tests were identified for 29 chemicals. Based on statistical analyses, the authors concluded that "there seems to be no reason to believe that organisms differ in sensitivity under field and laboratory conditions." Moreover, when species tested in the multiple species experiments were compared with similar or related species in single species studies (given corresponding response parameters and equivalent exposure concentrations) their response/sensitivity to a given chemical appeared essentially equivalent. Results of this inquiry suggest that single species toxicity test results are reliable predictors of biological community responses. With the caution that there are limited data, these authors conclude that it is acceptable to derive "safe" concentrations from single species toxicity test data.

9.3 Slooff (1985)

Slooff (1985) reached the same conclusion as Okkerman et al. (1993) and Emans et al. (1993) regarding the equivalency of effect concentrations from single species and multiple species toxicity tests after reviewing the literature studies.

9.4 Persoone and Janssen (1994)

The potential of laboratory single species test results to reliably predict biological community responses was examined in an extensive four year interlaboratory study with four chemicals (copper, atrazine, lindane, and dichloroaniline) by Persoone and Janssen (1994). NOECs from outdoor stream and pond microcosms were compared with those from single species laboratory tests. The NOECs from the field studies were within one order of magnitude of the NOECs of the most sensitive laboratory test, suggesting that the single species tests are effective qualitative predictors of ecosystem effect concentrations.

In their review of the literature on field "validation" of predictions based on single species toxicity test data Persoone and Janssen wrote, "One of the most striking conclusions of this literature study is that, in general, NOECs derived from (a selected battery of) single species laboratory tests relate relatively well to single species and multiple species NOECs obtained in field studies." Such studies are not truly "validation", but they do argue that abiotic and biotic factors in aquatic ecosystems do not greatly modify effect concentrations or bioavailability of chemicals compared to laboratory tests.

9.5 Phluger (1994)

NOECs from multiple species field and single species laboratory tests for ten pesticides were compared by Phluger

(cited in Persoone and Janssen, 1994). For all ten pesticides there was less than one order of magnitude difference between the single species laboratory and multiple species field test NOECs, suggesting that the single species test results were reliable qualitative predictors of biological community responses.

9.6 Dorn (1996)

In three separate stream mesocosm experiments, testing a homologous series of nonionic alcohol ethoxylate surfactants, Dorn (1996) found that laboratory single species toxicity test effect concentrations were within a factor of three of mesocosm effect concentrations. In summarizing a review of the literature Dorn (1995) concluded that effects observed in numerous mesocosm studies are consistent with laboratory single species toxicity test results when exposures are reconciled correctly.

9.7 Crane (1995)

In a Society of Toxicology and Chemistry (SETAC) News article, Crane (1995) states, "Such tests (mesocosms) cost several million dollars to perform, but the results obtained from them have shown no greater sensitivity or predictive power and certainly no greater interpretability, than considerably cheaper laboratory tests with single species". Crane refers to the reviews of Okkerman et al. (1993) and Emans et al.(1993) as support for his position.

10.0 Alternatives to Single Indicator Species Tests

If the desire is to continue with testing which can provide an early warning of probable aquatic biological community impairments while having good qualitative reliability in predicting ecosystem responses, possible options to the existing USEPA toxicity tests and other single species procedures include: 1) single indigenous species tests and 2) multiple indigenous species tests. Desirable test and endpoint characteristics for reliably predicting instream biological community responses have been listed (16 items) by Cairns and Niederlehner (1995). No current single species test can meet these criteria.

10.1 Tests with Single Indigenous Species

A common criticism of the indicator species tests is that the species does not occur in a particular waterbody. The argument is that the indicator species test should be replaced with an indigenous species test. From a biological perspective, the use of an indigenous species is sound, but care must be taken in the selection of a replacement species from the same phyletic group. Selecting an indigenous species from an impaired or partially impaired waterbody could be a mistake since that species would represent a species likely to have developed tolerance to chemical stressors. In the case of impaired systems selecting a species that could or one that previously did (from historical data) live in such a habitat may be necessary. While single indigenous species tests may decrease the uncertainty associated with extrapolating from single species test results to biological community responses, we need evidence that such tests will significantly increase the accuracy of predicting instream impairments. Such indigenous species tests may not significantly improve predictive accuracy enough to justify the time, effort, and cost of developing standard (with the essential QA/QC) protocols with indigenous species for multiple watersheds. Is it desirable, feasible, and cost effective to develop protocols for indigenous species in each watershed or subregions of watersheds? Furthermore, there is little evidence that indigenous species test results are more reliable predictors of biological community responses in complex and multivariate ecosystems.

Currently available single species tests can effectively reveal when there are significant levels of toxic chemicals in an effluent or ambient water sample. The statistical probability that any one test species represents the most or the least sensitive species, life stage, or endpoint in a given ecosystem is very low. Persoone and Gillett (1990) conclude that single indicator species toxicity tests do not represent the most sensitive species or endpoints, and especially key components, in aquatic ecosystems. In fact, one could argue that the single species tests (especially the USEPA toxicity tests) have been effective predictors of ecosystem responses because they manifest relatively average sensitivities compared to most aquatic ecosystem species and endpoints. Probability theory also advises us that there can never be enough predictive potential in the results of a single species toxicity test to encompass all possible effects on ecosystem structure and function.

Luoma and Carter (1993) conclude that single species toxicity tests results, when combined with chemical measurements and benthic community surveys have shown reliable qualitative relationships between toxicity tests responses, chemical concentrations, and changes in biological community structure. Slooff and Canton (1983) asserted that the sensitivities in three indicator species testing (alga, daphnid, fish) effectively represented aquatic organism sensitivity ranges for approximately 75% of the chemicals they tested. Many of the pitfalls of developing new single species tests are discussed by Luoma (1995). Luoma is not convinced that increasing the number of standardized single species toxicity tests will improve the accuracy of predicting ecosystem impacts. In reviewing the validity of using indicator, rather than resident species toxicity tests, Dorn (1996) suggested that use of "new" tests with resident species may not give us better resolution of aquatic ecosystem responses than do the well-developed indicator species tests. Rather than develop a host of new single species tests, Dorn (1996) advises that a better use of resources would be to assure

that laboratory and field exposure regimes are comparable (i.e., improve exposure assessments).

Chapman (1995a) concludes that the current standardized single species test protocols do not represent the more sensitive ecosystem endpoints; he also notes that daphnids and fathead minnows are usually not the most sensitive components in aquatic ecosystems. Several other authors (Persoone et al., 1990; Baird, 1992; Forbes and Depledge, 1992; Clements and Kiffney, 1996; LaPoint et al., 1996) also proposed that the USEPA toxicity tests and other single species tests most frequently underestimate effects in aquatic ecosystems. Examination of USEPA's chemical-specific water quality criteria documents illustrates that the USEPA toxicity test species (USEPA, 1994a,b) are not consistently among the most sensitive species tested.

In combination with Toxicity Identifications Evaluations (TIEs) and chemical analyses, the current set of single species toxicity tests appear to be effective in the identification of toxicity, as well as its sources and causes. Therefore, available funds and efforts could be focused on improving these procedures rather than on developing a host of new indigenous species testing procedures. Persoone et al. (1990) assert that despite limitations of indicator single species tests, they have been extremely useful and reliable predictors of ecosystem responses. Cairns and Mount (1990) conclude that developing toxicity test methodologies with "new" aquatic organisms is probably not productive unless the response of this new species has a high correspondence with responses of many other aquatic species. Cairns and Mount state, "For regulatory purposes, it is unquestionably sound to use test organisms that have been widely used for toxicity testing and whose strengths and weaknesses for this purpose are well known." There is little or no evidence that use of indigenous species in single species laboratory toxicity tests will improve our ability to predict responses in the field.

10.2 Tests With Multiple Indigenous Species

Several researchers have argued for the use of multiple species (micro/mesocosm) tests rather than single species tests in regulatory settings. The literature on multiple species toxicity test strengths and limitations will not be summarized here. Suffice it to say that the limitations of these tests seem to be equivalent or greater than for single species toxicity tests (Dickson et al., 1985; Mount, 1985; Slooff, 1985; Cairns et al., 1993; Cairns and Smith, 1994; Dickson, 1995; LaPoint, 1995; Smith, 1995). Generally, the designs of multiple species tests are highly variable with no standardized protocols or endpoints. Multiple species tests are predisposed to be ecosystem specific, which is a strength as well as a weakness. Results of multiple species tests tend to be more variable than those from single species toxicity tests. Factors that increase complexity of toxicity tests may boost ecological relevance, but result in greater variability, as well as less reliability and repeatability.

Although there is controversy on this issue, multiple species tests have not been found to be more sensitive than single species tests. Information is increasing, but "validation" of these multiple species test results with ecosystem bioassessments has not been frequent. Neuhold (1986) postulates that microcosm tests present interpretation problems and are not likely to offer reliable predictions of natural ecosystem impacts. Responses in control systems are difficult to replicate and responses in treatment groups tend to diverge greatly (Gearing, 1989). Several groups of investigators (Cairns, 1983; Giesy and Allred, 1985; Slooff et al., 1986; Luoma and Ho, 1993) conclude that mesocosms are better suited to testing process questions than to replicating nature. In regard to multiple species tests Bailey (1995) concluded that "I am not convinced that complex (i.e., multiple species) tests accompanied by simple models offer a reduction in (predictive) uncertainty over simple tests accompanied by complex models." Slooff (1985) argued that there is no evidence that multiple species tests are more reliable in predicting instream impacts than are results of single species toxicity tests.

Although multiple species tests may have greater predictive capacity than single species test results, they have limitations which include:

- There are no standardized protocols and developing multiple species testing procedures will be costly. Endpoints have not been agreed upon. Designs, endpoints measured, and statistical analyses of multiple species tests vary widely, resulting in considerable debate regarding the interpretation of every study.
- Multiple species protocols may have to be altered or developed for each aquatic ecosystem.
- Most multiple species tests are not designed to be early warning signals.
- Multiple species tests tend to have high within and between test variability, and especially high between laboratory variability (more than for the single species tests).
- Predictions of ecosystem responses based on multiple species test results will likely be qualitative rather than quantitative.

According to Giesy and Allred (1985), variability increases with the size and complexity of multiple species study design. These authors contend that replicability (ability to establish more than one experimental unit within a particular experimental treatment) of multiple species tests is generally sufficient, but that the realism and accuracy in these tests is largely unresolved. Further, Giesy and Allred

(1995) claim that repeatability (duplicating results of a test at a later time) of multiple species tests has seldom been examined.

The intent is not to discredit the importance of multiple species tests. The point is that multiple species tests may not be more reliable screening or predictive tools than are single species Tests. Multiple species tests are important for providing fundamental information on structure and function of aquatic ecosystems and for potential following up on single species toxicity test data.

11.0 Studies in Ocean or Estuarine Settings

Although the relationship between ocean water toxicity and water column biological community health has not been examined to any great extent, the link between laboratory marine sediment toxicity and biological community response has been studied. This review does not include an exhaustive examination of the literature dealing with marine sediment toxicity. However, several studies (see Appendix C) suggest that the results of sediment toxicity tests are fairly reliable qualitative predictors of benthic community responses. In a review of the literature on laboratory sediment toxicity testing with single species Lamberson et al. (1992) concluded that, despite realized and potential problems, the test results have proven "enormously successful as both research and management tools." As Luoma and Ho (1993) conclude in a review of the literature on sediment toxicity tests, it is inappropriate to use data from single species tests alone to quantitatively predict specific aquatic ecosystem impacts.

Appendix C includes summaries of ten studies in which laboratory single species toxicity test results on marine sediment samples were evaluated in terms of predicting benthic biological community responses. In all ten of these studies, the laboratory sediment tests were reliable qualitative predictors of benthic community effects; the laboratory tests tended to underestimate the extent of benthic community impacts.

Richardson and Martin (1994) critiqued the strengths and shortfalls of using ocean and estuarine toxicity testing as a procedure for evaluating potential water quality impacts. While constraints, including the laboratory-to-field verification, of single species toxicity testing are thoroughly discussed, these authors strongly advocate toxicity water quality standards and toxicity monitoring, similar to that outlined in the California Ocean Plan (State Water Resources Control Board, 1990), on a world-wide basis.

Section 2

1.0 Conclusions

Criticisms of single species tests, including the USEPA toxicity tests, have included excessive between test and between laboratory variability, as well as questions regarding the ecological relevance of test results. While some accept that contaminants are responsible for biological impacts based on an inverse correlation between chemical concentrations and biological indices, there has been some reluctance to make similar, parallel interpretations when ambient water or wastewater toxicity and biological indices are inversely correlated. Determination of contaminant concentrations *per se* provides no information on the bioavailability of these compounds to resident biota.

The information presented in this review offers compelling evidence that the USEPA toxicity tests and other single species toxicity test results are, in a majority of cases, reliable qualitative predictors of aquatic ecosystem community responses. However, this qualitative relationship must be based on a series of test results (persistent toxicity) not on a single test result. Participants at a 1996 Pellston conference on toxicity testing (Grothe et al., 1996; Waller et al., 1996) concluded that the USEPA toxicity tests provide "an effective tool for predicting receiving system impacts when appropriate considerations of exposure are considered. Further, laboratory to field validation is not essential for the continued use of these toxicity tests." According to that reference, the participants felt "It is unmistakable and clear that WET procedures, when used properly and for the intended purpose, are reliable predictors of environmental impacts."

Ideally, laboratory toxicity tests should provide ecologically relevant, reliable, and repeatable data. In practice, however, incorporating these desirable characteristics into laboratory toxicity tests has been difficult. The single species toxicity tests are successful (compared to multiple species tests) in providing reliable and repeatable data, but at some expense to ecological relevance (e.g., Calow, 1992). To assess effects of contaminants on aquatic biological communities there is a need for integrated, weight-of-evidence approaches. Especially at moderately polluted sites, a multiplicity of testing methods is helpful in estimating and evaluating biological community responses.

The optimal approach may be to integrate ecological surveys, toxicity tests, and chemical analyses to better understand contaminant effects on the health of aquatic ecosystems. The principal hinderance for this approach has been the complexity and costs of such combined, extensive efforts.

While there is some merit to the criticisms of the CETTP studies, they are not persuasive enough to doubt the effectiveness of the USEPA toxicity tests as qualitative forecasters of biological community responses. There are no empirical data which demonstrate that these tests fail to render reliable extrapolations to instream biological responses. Thus, when used appropriately as early warning screening procedures, these tests provide a powerful monitoring tool. When test results fail as reliable qualitative predictors of instream impairments, they have more frequently underestimated aquatic ecosystem impacts. The idea that biotic and abiotic factors in the environment significantly decrease bioavailability and toxicity was not supported by a majority of the studies reviewed.

Data from the 7-d *Ceriodaphnia* test, in particular, have been very reliable predictors of instream biological responses. This may be due, in part, to the large database for this species. On the other hand, Slooff and Canton (1983) contend that single species tests with daphnids have been extremely efficient in the identification of chemical concentrations harmful to aquatic ecosystems. Summarized in this document are some 49 studies in which a cladoceran (*Ceriodaphnia* or *Daphnia*) was utilized as the laboratory test species. These tests were performed at many locations across the country with a wide variety of ambient water types (in most of which these cladocerans were not resident), effluents, and chemicals (including pesticides, other organic chemicals, metals, and inorganic chemicals). Results from these laboratory cladoceran tests were reliable predictors of aquatic ecosystem biological community responses or adverse effect concentrations in 33 (67%) of the 49 studies (cf., Figure 3). The laboratory cladoceran tests underestimated biological community responses (or overestimated ecosystem adverse effect concentrations of a chemical) in 16 (33%) of the investigations. There were no studies in which the cladoceran tests overestimated impairments to biological communities.

Defending single species toxicity tests beyond their capabilities is to no one's advantage. Single species test results alone are not reliable quantitative forecasters of

toxic chemical impacts on complex ecosystems. The simplicity of the single species tests comes at a cost of interpretation and predictive depth. The test protocols can always be improved so that we are more confident of their meaning and so that their results are more reliable predictors of instream impacts. A better understanding of the limitations of extrapolations is needed so that modifications can be made which increase reliability and decrease uncertainty, as well as establish a stronger theoretical basis for extrapolations. A list of limitations and strengths of single species tests is provided in Appendix D.

Establishing a quantitative correlation between a biological response from a single grab or composite water sample and the biological community responses is not only impossible, but unnecessary. However, with a good temporal representation of ambient or effluent toxicity and with carefully designed/performed seasonal bioassessment data from streams, statistically significant correlations between data sets have been possible. However, thorough ecological surveys in large rivers and bays will be difficult and expensive; there is likely to be considerable controversy regarding what sites, if any, should serve as reference ("clean") sites and, if there are not reference sites, what represents a "healthy" aquatic ecosystem. Many rivers and bays in the U.S. are significantly altered by human activities; attempting to attribute degradation in these systems to toxic chemicals with the use of bioassessments will be very difficult. Even though there cannot be direct proof that toxic chemicals are a cause of declining aquatic organism populations, it is not advisable to forsake the qualitatively predictive early warning tools.

The reliability with which single species toxicity test results predict biological community responses relates to several factors. One major factor was addressed by Dickson et al. (1992); they observed that when effluent or ambient water toxicity is relatively low or when impacts on aquatic ecosystems are moderate it will be difficult to establish a relationship between toxicity and instream ecological responses. The strength of the predictive capacity of single species test results is substantially enhanced when the test is performed with ambient water (e.g., as compared to effluent) and with higher magnitude toxicity in the sample. Chapman et al. (1987) came to a similar conclusion regarding magnitude of toxicity in relation to sediment tests. We appear to be approaching consensus that when significant lethality (and in the case of effluents, assuming accurate dilution has been considered) is seen in toxicity tests there is a very high potential of aquatic ecosystem impairment. As this connection is accepted, we continue to struggle with the idea that sublethal effects on indicator species can result in detectable adverse ecosystem responses.

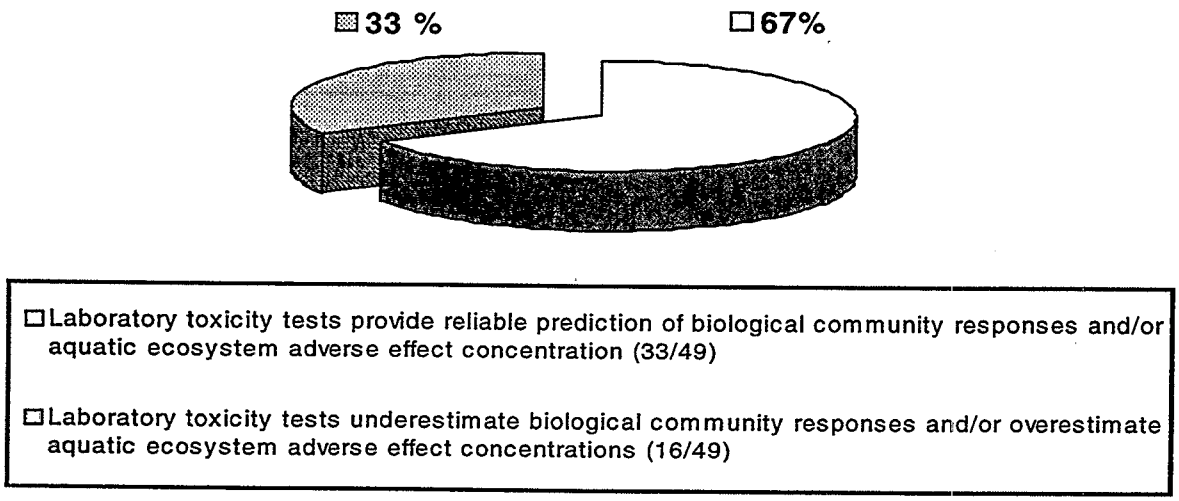


Figure 3. Summary of studies in which a cladoceran was used as a laboratory test organism when comparing toxicity test results to ecological survey data and/or field effect concentrations. Percentages represent the outcomes of the studies. Total number of studies is 49.

Possibilities for decreasing the extrapolation uncertainties and improving the predictiveness of single species test results include: 1) more thorough characterization (persistence, frequency, magnitude) of ambient water or effluent toxicity, 2) more effective matching (or accounting for) of exposure patterns in natural ecosystems as compared to laboratory tests, 3) develop a more thorough comprehension of what constitutes critical aquatic ecosystem endpoints, 4) improved simulation, or consideration of, ecosystem characteristics and processes in laboratory tests (a corollary to this point would be to avoid defaulting to worst case scenarios in all cases), 5) more thorough knowledge of environmental fate and bioavailability of chemicals, 6) develop models which map the quantitative and qualitative relation between single species test endpoints and important ecosystem endpoints (this would include focusing on the relative sensitivities of surrogate species compared to key ecosystem endpoints), 7) focus on or develop tests with endpoints which have a clear connection to important ecosystem structures/functions, 8) enhance the intertest repeatability of single species tests, 9) improve understanding of how toxicity is manifested in complex ecosystems, and 10) develop field and laboratory approaches which are complementary.

A convincing relationship has been established between ambient water toxicity (as manifested by single species tests) and biological community responses, but has such connection been authenticated between effluent toxicity and instream impairments? The effluent-biological community link has not been as thoroughly investigated. Nonetheless, in several recent studies (see Section 1.0, subsection 7.0), as well as the CETTP and associated studies, where effluent toxicity was assessed, a reliable qualitative estimate of instream biological effects was obtained. This relationship was most evident when flow and dilution of the receiving water were effectively estimated and when environmental exposure duration was matched (or account for) by laboratory toxicity test duration.

Recently, Sprague (1995) observed that single species toxicity test results give us answers that support action, and the important response is to take action rather than wait for a 98% certainty. Because these single species toxicity test results are, in a large majority of cases, reliable qualitative predictors of biological community responses, the controversy surrounding these tests can diminish if the data from these tests are used appropriately. Moreover, if the results of a single test are not characterized as a violation of an effluent limit or a water quality standard, but rather as a gauge of relative toxicity and, therefore, a signal to initiate repeated or more frequent sampling/testing (or TIEs) to better characterize potential effluent or ambient water toxicity, regulated entities may be less critical of the single species tests. Prior to making predictions regarding biological community impacts, ambient water or effluent toxicity must be characterized so there can be more certainty regarding the nature of the toxicity. Furthermore, it is difficult to control toxicity until its nature, cause, and source are known. If the results of single species tests are used to signal the potential for instream impairments, then the toxicity test (USEPA, 1994a,b) results do not have to be quantitative predictors, but rather effective qualitative predictors. These tests do provide a reliable and repeatable qualitative predictive capability. Mount (1995) stated, "It is the application of the toxicity data, not its inherent validity, that is questionable." In harmony with Mount, Luoma and Carter (1993) conclude that it is the "interpretation and application of results" from single species tests that are controversial.

Critics of the single species tests fail to recognize that these tests, even with nonindigenous species, reveal that water samples contain biologically significant concentrations of toxic chemicals. Results of laboratory single species tests are based on the toxicological principle of concentration-response. This principle is fundamental and well established. The effectiveness of the single species tests in predicting biological community responses is centered in this principle of concentration-response.

If the single species tests continue to be used as early warning, screening tools (identification of a potential problem which is further investigated), there is less necessity for developing new standardized indigenous species testing protocols. The probability that any particular test species, whether or not it is indigenous to the particular aquatic ecosystem, represents the most or the least sensitive group of species or endpoints in a specific biological assemblage is very low. More likely, the sensitivity of the test species would fall somewhere within the sensitivity distribution of organisms from an aquatic ecosystem. This is perhaps one of the reasons why the short-term toxicity tests (USEPA, 1994a,b) and other single species test results have been effective qualitative predictors of ecosystem biological responses. Developing testing protocols with indigenous species in many different aquatic ecosystems may improve accuracy of predictions, however, such efforts will be expensive and difficult undertakings (e.g., Persoone and Gillett, 1990;Luoma, 1995). Furthermore, there is little evidence, and no guarantee, that the reliability of environmental impact predictions will be significantly enhanced with indigenous species tests.

Slooff (1985), as well as Persoone and Janssen (1994) discuss the wide range of sensitivities of aquatic organisms, life stages, and endpoints to toxic chemicals. These authors also observe that enlarging the suite of test species, life stages, and endpoints almost always results

in lowering environmental effect concentrations (i.e., LOECs/NOECs decrease with increasing number of test species, life stages, and endpoints--i.e., with increasing amounts of toxicity data). Several authors (e.g., Slooff and Canton, 1983; Persoone and Gillett, 1990; Persoone et al., 1990; Chapman, 1995b; Luoma, 1995; Underwood, 1995; Dorn, 1996) maintain that indicator species are not likely to represent the most sensitive aquatic ecosystem response, but rather have been selected for robustness, ease of culture and availability. Bartell et al. (1992) propose that area and sensitive species, by definition, are seldom selected for routine toxicity testing. Several other authors (Persoone et al., 1990; Baird, 1992; Forbes and Depledge, 1992; Clements and Kiffney, 1996; LaPoint et al., 1996) also proposed that the USEPA's toxicity tests for effluents and receiving waters (USEPA, 1994a,b) and other single species toxicity tests most frequently underestimate effects in aquatic ecosystems.

Because single species test results are reliable qualitative predictors of biological community responses, the burden of proof in demonstrating that persistent toxicity is not impacting biological communities perhaps should rest with the entity(ies) responsible for the contaminants. Moreover, at some stage it should become incumbent on the entity responsible for the probable environmental impacts to demonstrate the absence of ecosystem impairments. As stated by Luoma (1995), "The toxicity tests tool may never achieve the high probability prediction capabilities 'required' by ardent critics; however, this does not prevent the approach from being a useful tool in the developing arsenal available to study the effects of contaminants".

2.0 Summary

Regulatory agencies have tended to rely on single species toxicity tests, particularly USEPA's toxicity tests, on surface or effluent water samples to identify potential chemical toxicity threats to aquatic biological communities. Questions regarding the reliability of these laboratory test results in predicting impairments to biological communities have been advanced. Of particular concern are uncertainties of extrapolating from the outcomes of these highly controlled laboratory tests to complex and multivariate ecosystems. This document is an interpretive review of the literature on this question of ecological relevance of single species toxicity test results; it includes, but is not restricted to USEPA's Complex Effluent Toxicity Testing Program (CETTP--conducted for the purpose of examining the predictive correspondence between the short-term toxicity tests (USEPA, 1994a,b) and instream impacts).

Aquatic ecosystem surveys typically have been used to assess the reliability of single species toxicity test results extrapolations. Potential limitations to the use of these bioassessments for "validating" predictions extrapolated from single species tests are discussed; caution is urged in the interpretation of these surveys. Strengths and limitations of the CETTP and associated studies, as well as of two recent statistical analyses of those studies, are evaluated.

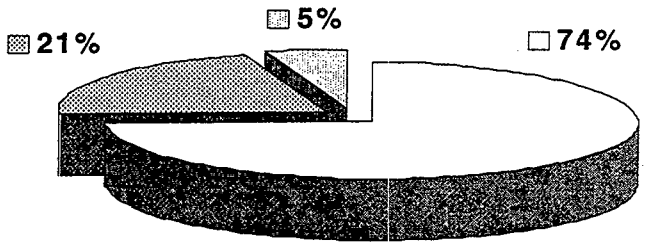
Approximately 80 studies in which single species tests were used to assess ambient water or effluent toxicity and in which some ecological survey data were gathered, for the purpose of exploring the correspondence between toxicity data and biological community responses, are critically evaluated. A preponderance of evidence reveals that USEPA's toxicity tests (USEPA, 1994a,b) and other single species test results are, in a majority of cases, reliable qualitative (some level of response seen) predictors of aquatic ecosystem community effects. In this document 77 independent studies in which the results of laboratory indicator single species toxicity tests are assessed with regard to reliability in predicting aquatic ecosystem biological community responses (and/or adverse effect concentrations) are summarized. In 57 (74%) of the studies the indicator single species tests provided reliable qualitative predictions of biological community impacts or adverse effect concentrations (cf., Figure 4). The laboratory single species tests underestimated aquatic ecosystem effects (and/or overestimated the biological community adverse effect concentration of a chemical) in 16 (21%) of the 77 studies. Results of four (5%) of the studies were inconclusive or mixed. There are no experimental data which demonstrate that the single species tests generally fail to render reliable qualitative extrapolations to biological community responses.

While criticisms of the USEPA toxicity tests (USEPA, 1994a,b) and other single species tests have some merit, they are not persuasive enough to cast doubt on the effectiveness of these tests in predicting ecosystem impacts. When used appropriately as early warning signals and with dependable temporal representation of ambient water or effluent toxicity, these tests provide a powerful monitoring tool. When single species tests fail as reliable qualitative predictors, they most frequently underestimate impacts to the ecosystem community. Single species test results alone are not reliable quantitative forecasters of toxic chemical impacts on complex ecosystems.

The predictive power of single species tests is substantially enhanced when ambient water, as compared to discharge, is tested and when higher magnitude toxicity exists; reliability is also improved when exposure patterns in natural ecosystems are matched or accounted for and, in the case of effluents, when realistic estimates of dilution are taken into account.

Alternatives to indicator species tests are explored. There is a paucity of evidence that the current standardized toxic-

ity testing protocol (including the USEPA toxicity tests; USEPA, 1994a,b) test species are more sensitive to toxic chemicals than resident species. If the single species tests continue to be used as early warning signals there is less necessity for developing new standardized indigenous species testing protocols. The wisdom of developing a host of new standardized tests with indigenous species, unless they will substantially improve accuracy of predicting ecosystem impacts, is questionable.



□ Laboratory single species toxicity tests provide reliable prediction of biological community responses and/or aquatic ecosystem adverse effect concentration [57/77]

▨ Laboratory single species toxicity tests underestimate biological community responses and/or overestimate aquatic ecosystem adverse effect concentration (i.e., adverse effects occur at lower concentration than predicted in laboratory test [57/77]

▨ Laboratory single species toxicity test yielded mixed or inconclusive results [4/77]

Figure 4. Summary of studies reviewed in this report in which the results of laboratory single species toxicity tests were compared to biological community surveys and/or field effect concentrations. Tabulation is by overall outcome of the study. Total number of studies summarized is 77.

Section 3

1.0 References

*Review articles or publications relating to the ecological relevance of single species toxicity tests are noted by ***. References are inclusive for appendices.*

Adams, W.J., R.A. Kimerle, B.B. Heidolph, and P.R. Michael. 1983. Field Comparison of Laboratory-derived Acute and Chronic Toxicity Data. pp. 367-385. In: W.E. Bishop, R.D. Cardwell, and B.B. Heidolph, eds., *Aquatic Toxicology and Hazard Assessment*, ASTM STP 802. American Society for Testing and Materials, Philadelphia, PA.

Bailey, H.C. 1995. Letter to The Editor. Human Ecol. Risk Assess. 1. 459-463.

Bailey, H.C., C. Alexander, C. DiGiorgio, M. Miller, S.I. Doroshov, D.E. Hinton. 1994. The Effects of Agricultural Discharges on Striped Bass (*Morone Saxatilis*) in California's Sacramento-San Joaquin Drainage. *Ecotoxicology* 3:123-142.

Baird, D.J. 1992. Predicting Population Response to Pollutants: in Praise of Clones. A Comment on Forbes & Depledge. *Funct. Ecol.* 6:616-617.

Barbour, M.T., J.M. Diamond, and C.O. Yoder. 1996. Biological Assessment Strategies: Applications and Limitations. pp. 245-270. In: D.R. Grothe, K.L. Dickson, and K. Reed-Judkins (eds), Whole Effluent Toxicity Testing: *An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Bartell, S.M., R.H. Gardner, and R.V. O'Neill. 1992. *Ecological Risk Estimations.* Lewis Publishers, Boca Raton, FL.

Baughman, D.S., D.W. Moore, and G.I. Scott. 1989. A Comparison and Evaluation of Field and Laboratory Toxicity Tests with Fenvalerate on an Estuarine Crustacean. *Environ. Toxicol. Chem.* 8:417-429.

Becker, D.S., G.R. Bilyard, and T.C. Ginn. 1990. Comparisons Between Sediment Toxicity Tests and Alterations of Benthic Macroinvertebrate Assemblages at a Marine Superfund Site: Commencement Bay, Washington. *Environ. Toxicol. Chem.* 9:669-685.

Birge, W.J. and J.A. Black. 1990. In Situ Toxicological Monitoring: Use in Quantifying Ecological Effects of Toxic Wastes. pp. 215-231. In: S.S. Sandhu, W.R. Lower, F.J. de Serres, W.A. Suk, and R.R. Tice, eds., *In Situ Evaluation of Biological Hazards of Environmental Pollutants.* Plenum Press, New York, NY.

Birge, W.J., J.A. Black, T.M. Short, and A.G. Westerman. 1989. A Comparative Ecological and Toxicological Investigation of Secondary Wastewater Treatment Plant Effluent and Its Receiving Stream. *Environ. Toxicol. Chem.* 8:437-450.

Birge, W.J., D.J. Price, D.P. Keogh, J.A. Zuiderveen, and M.D. Kercher. 1992. *Biological Monitoring Program for the Paducah Gaseous Diffusion Plant. Annual Report for study period Oct.* 1990 through March 1992. Submitted to Oak Ridge National Laboratory, Oak Ridge, TN.

Boelter, A.M., F.N. Lamming, A.M. Farag, and H.L. Bergman. 1992. Environmental Effects of Saline Oil-field Discharges on Surface Waters. *Environ. Toxicol. Chem.* 11:1187-1195.

Boyle, T.P., S.E. Finger, R.L. Paulson, and C.F. Rabeni. 1985. Comparison of Laboratory and Field Assessment of Fluorene. Part Ii: Effects on the Ecological Structure and Function of Experimental Pond Ecosystems. pp. 134-151. In: T.P. Boyle, ed., *Validation and Predictability of Laboratory Methods for Assessing the Fate and Effects of Contaminants in Aquatic Ecosystems.* ASTM STP 865. American Society for Testing and Materials, Philadelphia, PA.

Brock, T.C.M., S.J.H. Crum, R. van Wijngaarden, B.J. Budde, J. Tijink, A. Zuppelli, and p. Leeuwangh. 1992. Fate and Effects of the Insecticide Dursban 4E in Indoor *Elodea*-dominated and Macrophyte-free Freshwater Model Ecosystems: I. Fate and Primary Effects of the Active Ingredient Chlorpyrifos. *Arch. Environ. Contam. Toxicol.* 23:69-84.

Burton, G.A. Jr. and G.R. Lanza. 1987. Aquatic Microbial Activity and Macrofaunal Profiles of an Oklahoma Stream. *Wat. Res.* 21:1173-1182.

Burton, G.A. Jr., A. Drotar, J. M. Lazorchak, and L.L. Baals. 1987. Relationship of Microbial Activity and *Ceriodaphnia* Responses to Mining Impacts on the Clark Fork River, Montana. *Arch. Environ. Contam. Toxicol.* 16:523-530

***Cairns, J. Jr. 1983. Are Single Species Toxicity Tests Alone Adequate for Estimating Environmental Hazard? *Hydrobiologia* 100:47-57.

***Cairns, J. Jr. 1986. What Is Meant by Validation of Predictions Based on Laboratory Toxicity Tests? *Hydrobiologia* 137:271-278.

***Cairns, J. Jr. 1988a. What Constitutes Field Validation of Predictions Based on Laboratory Evidence? pp.361-368. In: Adams, G.A. Chapman, and W.G. Landis, eds., *Aquatic Toxicology and Hazard Assessment*: Tenth Volume, ASTM STP 971, American Society for Testing and Materials, Philadelphia, PA.

***Cairns, J. Jr., 1988b. Should Regulatory Criteria And Standards Be Based on Multispecies Evidence? *Environ. Profess.* 10:157-165.

***Cairns, J. Jr. 1988c. Putting the Eco in Ecotoxicology. *Regulatory Toxicol. Pharm.* 8:226-238.

Cairns, J. Jr., and D.S. Cherry. 1983. A Site-specific Field and Laboratory Evaluation of Fish and Asiatic Clam Population Responses to Coal Fired Power Plant Discharges. *Wat. Sci. Tech.* 15:31-58.

***Cairns, J., Jr. and D.I. Mount. 1990. Aquatic Toxicology, Part 2. *Environ. Sci. Technol.* 24: 154-161.

Cairns, J. Jr., and B.R. Niederlehner. 1995. Predictive Ecotoxicology: Methods for Making Estimates and Predictability in Ecotoxicology. pp.667-680. In: D.J. Hoffman, B.A. Rattner, G.A. Burton, and J. Cairns Jr., eds., *Handbook of Ecotoxicology*, CRC Press, Inc., Boca Raton, FL.

Cairns, J. Jr., and E.P. Smith. 1994. The Statistical Validity of Biomonitoring Data. pp. 49-68. In: S.L. Loeb, and A. Spacie, eds., *Biological Monitoring of Aquatic Systems.* Lewis Publishers, Boca Raton, FL.

Cairns, J. Jr., D.S. Cherry, and J.D. Grattina 1982. Correspondence Between Behavioral Responses of Fish in Laboratory and Field Heated Chlorinated Effluents. pp. 207-215. In: W.J. Mitsch, R.W. Bosserman, and J.M. Klopatek, eds., *Energy and Ecological Modelling.* Elsevier Scientific Publishers Co., Amsterdam.

Cairns, J. Jr., P.V. McCormick, and S.E. Belanger. 1993. Prospects for the Continued Development of Environmentally-realistic Toxicity Tests Using Microorganisms. *J. Environ. Sci.* 5:253-268.

Calow, P. 1992. The Three R's of Ecotoxicology. *Funct. Ecol.* 6:617-619.

Canfield, T.G., N.E. Kemble, W.G. Brumbaugh, F.G. Dwyer, C.G. Ingersoll and J.F. Fairchild. 1994. Use of Benthic Invertebrate Structure and the Sediment Quality Triad to Evaluate Metal-contaminated Sediment in the Upper Clark Fork River, Montana. *Environ. Toxicol. Chem.* 13: 1999-2012.

Carlson, A.R., H. Nelson, and D. Hammermeister. 1986. Development and Validation of Site-specific Water Quality Criteria for Copper. *Environ. Toxicol. Chem.* 5:997-1012.

***Chapman, P.M. 1995a. Extrapolating Laboratory Toxicity Results to the Field. *Environ. Toxicol. Chem.* 14: 927-930.

Chapman, P.M. 1995b. Do Sediment Toxicity Tests Require Validation? *Environ. Toxicol. Chem.*:14:1451-1453.

***Chapman, P.M. 1995c. Ecotoxicology and Pollution–Key Issues. *Marine Poll. Bull.* 16:405-415.

Chapman, P.M., R.N. Dexter, and E.R. Long. 1987. Synoptic Measures of Sediment Contamination, Toxicity and Infaunal Community Composition (The Sediment Quality Triad) in San Francisco Bay. *Mar. Ecol. Prog. Ser.* 37:75-96.

Clark, J.R., P.W. Borthwick, L.R. Goodman, J.M. Patrick, Jr., E.M. Lores, and J.C. Moore. 1987. Comparison of Laboratory Toxicity Test Results with Responses of Estuarine Animals Exposed to Fenthion in the Field. *Environ. Toxicol. Chem.* 6: 151-160.

Clements, W.H. and P.M. Kiffney. 1994. Integrated Laboratory and Field Approach for Assessing Impacts of Heavy Metals at the Arkansas River, Colorado. *Environ. Toxicol. Chem.* 13:397-404.

Clements, W.H. and P.M. Kiffney. 1996. Validation of Whole Effluent Toxicity Tests: Integrated Studies Using Field Assessments, Microcosms, and Mesocosms. pp. 229-244. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Cooper, W.E., and R.J. Stout. 1985. The Monticello Experiment: A Case Study. pp. 96-116. In: *Multispecies*

Toxicity Testing, J. Cairns, Jr., ed., Pergamon Press, New York, NY.

Crane, M. 1995. Society of Environmental Toxicology and Chemistry (SETAC) News Article. Vol. 15, No. 2., March.

Crossland, N.O. 1984. Fate and Biological Effects of Methyl Parathion in Outdoor Ponds and Laboratory Aquaria. *Ecotox. Environ. Safe.* 8:482-495.

Crossland, N.O. and J.M. Hillaby. 1985. Fate and Effects of 3,4-dichloroaniline in the Laboratory and in Outdoor Ponds: II. Chronic Toxicity to *Daphnia* Sp. And Other Invertebrates. *Environ. Toxicol. Chem.* 4:489-499.

Crossland, N.O. and C.J.M. Wolff. 1985. Fate and Biological Effects of Pentachlorophenol in Outdoor Ponds. *Environ. Toxicol. Chem.* 4:73-86.

Crossland, N.O., G.C. Mitchell, and P.B. Dorn. 1992. Use of Outdoor Artificial Streams to Determine Threshold Toxicity Concentrations for a Petrochemical Effluent. *Environ. Toxicol. Chem.* 11:49-59.

Davis, W.S. and T.P. Simon. 1995. *Biological Assessment and Criteria.* Lewis Publishers, Boca Raton, FL.

deNoyelles, F. Jr., and W.D. Kettle. 1985. Experimental Ponds for Evaluating Toxicity Tests Predictions. pp. 91-103. In: T.P. Boyle, ed., *Validation and Predictability of Laboratory Methods for Assessing the Fate and Effects of Contaminants in Aquatic Ecosystems,* ASTM STP 865. American Society for Testing and Materials, Philadelphia, PA.

Diamond, J.M., J.C. Hall, D.M. Pattie, and D. Gruber. 1994. Use of an Integrated Approach to Determine Site-specific Effluent Metal Limits. *Water Environ. Res.* 66:733-743.

Dickson, K.L. 1995. Progress in Toxicity Testing–An Academic's Viewpoint. pp. 209-216. In: J Cairns, Jr. and B.R. Niederlehner, eds., Ecological Toxicity Testing, Lewis Publishers, Boca Raton, FL.

Dickson, K.L., T. Duke, and G. Loewengart. 1985. A Synopsis: Workshop on Multispecies Toxicity Tests. pp. 76-88. In: J. Cairns, Jr., ed., *Multispecies Toxicity Testing.* Pergamon Press, New York, NY.

Dickson, K.L., W.T. Waller, J.H. Kennedy, W.R. Arnold, W.P. Desmond, S.D. Dyer, J.F. Hall, J.T. Knight, Jr., D. Malas, M.L. Martinez and S.L. Mutzner, 1989. A Water Quality and Ecological Survey of the Trinity River, Vols. I and II. Final Report. City of Dallas Water Utilities, Dallas, TX.

***Dickson, K.L., W.T. Waller, J.H. Kennedy, and L.P. Ammann. 1992. Assessing the Relationship Between Ambient Toxicity and Instream Biological Response. *Environ. Toxicol. Chem.* 11:1307-1322.

Dickson, K.L., W.T. Waller, J.H. Kennedy, L.P. Ammann, R. Guinn, and T.J. Norberg-King. 1996. Relationships Between Effluent Toxicity, Ambient Toxicity, and Receiving System Impacts: Trinity River Dechlorination Case Study. pp 287-308. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Dorn, P. 1996. An Industrial Perspective on Whole Effluent Toxicity Testing. pp. 16-37. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Dorn, P.B., R. van Compernolle, C.L. Meyer, and N.O. Crossland. 1991. Aquatic Hazard Assessment of the Toxic Fraction from the Effluent of a Petrochemical Plant. *Environ. Toxicol. Chem.* 10:691-703.

Eagleson, K.W., D.L. Lenat, L.W. Ausley, and F.B. Winborne. 1990. Comparison of Measured Instream Biological Responses with Responses Predicted Using the *Ceriodaphnia dubia* Chronic Toxicity Test. *Environ. Toxicol. Chem.* 9:1019-1028.

Eaton, J., J. Arthur, R. Hermanutz, R. Kiefer, L. Mueller, R. Anderson, R. Erickson, B. Nordling, J. Rogers, and H. Pritchard. 1985. Biological Effects of Continuous and Intermittent Dosing of Outdoor Experimental Streams with Chlorpyrifos. pp. 85-118. In: R.C. Bahner and D.J. Hansen, eds., *Aquatic Toxicology and Hazard Assessment: Eighth Symposium, ASTM STP 891*, American Society for Testing and Materials, Philadelphia, PA.

Eisle, P.J. and R. Hartung. 1976. The Effects of Methoxychlor on Riffle Invertebrate Populations and Communities. *Trans. Am. Fish. Soc.* 105:628-633.

***Emans, H.J.B., E.J. v.d.Plassche, J.H. Canton, P.C. Okkerman, and P.M. Sparenburg. 1993. Validation of Some Extrapolation Methods Used for Effect Assessment. *Environ. Toxicol. Chem.* 4:155-166.

Fairchild, J.F., F.J. Dwyer, T.W. La Point, S.A. Burch, and C.G. Ingersoll. 1993. Evaluation of a Laboratory-generated NOEC for Linear Alkylbenzene Sulfonate in Outdoor Experimental Streams. *Environ. Toxicol. Chem.* 12:1763-1775.

Fairchild, J.F., T.W. La Point, J.L. Zajicek, M.K. Nelson, F. J. Dwyer, and P.A. Lovely. 1992. Population-, Community- and Ecosystem-level Responses of Aquatic Mesocosms to Pulsed Doses of a Pyrethroid Insecticide. *Environ. Toxicol. Chem.* 11:115-129.

Ferraro, S.P., R.C. Swartz, F.A. Cole, and D.W. Schults. 1991. Temporal Changes in the Benthos Along a Pollution Gradient: Discriminating the Effects of Natural Phenomena from Sewage-industrial Wastewater Effects. *Estuarine, Coastal and Shelf Science* 33:383-407.

***Forbes, V.E. and M.H. Depledge. 1992. Predicting Population Response to Pollutants: Significance of Sex. *Funct. Ecol.* 6:376-381.

Franco, P.J., J.M. Giddings, S.E. Herbes, L.A. Hook, J.D. Newbold, W.K. Roy, G.R. Southworth, and A.J. Stewart. 1984. Effects of Chronic Exposure to Coal-derived Oil on Freshwater Ecosystems: I. Microcosms. *Environ. Toxicol. Chem.* 3:447-463.

Frithsen, J.B., D. Nacci, C. Oviatt, C.J. Strobel, and R. Walsh. 1989. Using Single-species and Whole Ecosystem Tests to Characterize the Toxicity of a Sewage Treatment Plant Effluent. pp. 231-250. In: G.W. Suter II and M.A. Lewis, eds., *Aquatic Toxicology and Environmental Fate: Eleventh Volume, ASTM STP 1007*, American Society for Testing and Materials, Philadelphia, PA.

Gearing, J.N. 1989. The Role of Aquatic Microcosms in Ecological Research as Illustrated by Large Marine Systems. pp. 411-448. In: *Ecotoxicology: Problems and Approaches* (S.A. Levin, M.A. Harwell, J.R. Kelly and K.D. Kimball. Springer Verlag, New York, NY.

Geckler, J.R., W.B. Horning, T.M. Nieheisel, Q.H. Pickering, E.L. Robinson, and C.E. Stephan. 1976. Validity of Laboratory Tests for Predicting Copper Toxicity in Streams. EPA 600/3-76-116. Cincinnati, OH.

Giddings, J.M., and P.J. Franco. 1985. Calibration of Laboratory Toxicity Tests with Results from Microcosms and Ponds. pp. 104-119. In: T.P. Boyle, ed., *Validation and Predictability of Laboratory Methods for Assessing the Fate and Effects of Contaminants in Aquatic Ecosystems*, STP 865. American Society for Testing and Materials, Philadelphia, PA.

Giddings, J.M., P.J. Franco, S.M. Bartell, R.M. Cushman, S.E. Herbes, L.A. Hook, J.D. Newbold, G.R. Southworth, and A.J. Stewart. 1984. *Effects of Contaminants on Aquatic Ecosystems: Experiments with Microcosms and Outdoor Ponds.* Oak Ridge National Laboratory, Oak Ridge, TN.

Giesy, J.P. and P.M. Allred. 1985. Replicability of Aquatic Multispecies Test Systems. pp. 187-247. In: J. Cairns Jr., ed. *Multispecies Toxicity Testing.* Pergamon Press, New York, NY.

Giesy, J.P., Jr., H.J. Kania, J.W. Bowling, R.L. Knight, S. Mashburn, and S. Clarkin. 1979. *Fate and Biological Effects of Cadmium Introduced into Channel Microcosms.* EPA 600/3-79-039. Duluth, MN.

Gonzalez, M.J. and T.M. Frost. 1994. Comparisons of Laboratory Toxicity Tests and a Whole-lake Experiment: Rotifer Responses to Experimental Acidification. *Ecological Applications* 4(1):69-80.

***Grothe, D.R., K.L. Dickson, and D.K. Reed-Judkins (eds.). 1996. *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Hansen, S.R. and R.R. Garton. 1982. Ability of Standard Toxicity Tests to Predict the Effects of the Insecticide Diflubenzuron on Laboratory Stream Communities. *Can. J. Fish. Aquat. Sci.* 39:1273-1288.

Havas, M. and T.C. Hutchinson. 1982. Aquatic Invertebrates from the Smoking Hills, N. W. T.: Effect of pH and Metals on Mortality. *Can. J. Fish. Aquat. Sci.* 39:890-903.

Herbold, B., A.D. Jassbay, and P.B. Moyle. 1992. *Status and Trends Report on Aquatic Resources in the San Francisco Estuary.* San Francisco Estuary Report, CA.

Hitchock, S.W. 1965. Field and Laboratory Studies of DDT on Aquatic Insects. *Conn. Ag. Exp. Bull.* (New Haven) 668:1-32.

Kersting, K. and R. van Wijngaarden. 1982. Effects of Chlorpyrifos on a Microecosystem. *Environ. Toxicol. Chem.* 11:365-372.

Larnberson, J.O., T.H. DeWitt, and R.C. Swartz. 1992. Assessment of Sediment Toxicity to Marine Benthos. pp. 183-211. In: G.A. Burton, Jr., ed. *Sediment Toxicity Assessment,* Lewis Publishers, Boca Raton, FL.

LaPoint, T.W. 1994. Interpreting the Results of Agricultural Microcosm Tests: Linking Laboratory and Experimental Field Results to Predictions of Effect in Natural Ecosystems. pp. 83-94. In: I.R. Hill, F. Heimbach, P. Leeuwangh, and P. Matthiesen, eds., *Freshwater Field Tests for Hazard Assessment of Chemicals.* Lewis Publishers, Boca Raton, FL.

LaPoint, T.W. 1995. Signs and Measurements of Ecotoxicology in the Aquatic Environment. pp. 13-24.

In: D.J. Hoffman, B.A. Rattner, G.A. Burton, Jr., and J. Cairns, Jr., eds., Handbook of Ecotoxicology. Lewis Publishers, Boca Raton, FL.

LaPoint, T.W., M.T. Barbour, D.L. Burton, D.S. Cherry, W.H. Clements, J.M. Diamond, D.R. Grothe, M.A. Lewis, D.K. Reed-Judkins, and G.W. Saalfeld. 1996. Field assessments. pp. 191-228. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

LaPoint, T.W., J.F. Fairchild, E.E. Little, and S.E. Finger. 1989. Laboratory and Field Techniques in Ecotoxicological Research: Strengths and Limitations. pp. 239-255. In: *A. Boudou and F. Ribeyre, eds, Aquatic Ecotoxicology: Fundamental Concepts and Methodologies, II.* CRC Press, Inc. Boca Raton, FL.

Larsen, D.P., F. deNoyelles Jr., F. Stay, and T. Shiroyama. 1986. Comparisons of Single-species, Microcosm and Experimental Pond Responses to Atrazine Exposure. *Environ. Toxicol. Chem.* 5:179-190.

Leeuwangh, P., T.C.M. Brock, and K. Kersting. 1994. An Evaluation of Four Types of Freshwater Model Ecosystem for Assessing the Hazard of Pesticides. *Human and Experimental Toxicology* 13:888-899.

Little, E.E., F. J. Dwyer, J.F. Fairchild, A.J. DeLonay, and J.L. Zajicek. 1993. Survival of Bluegill and Their Behavioral Responses During Continuous and Pulsed Exposures to Esfenvalerate, a Pyrethroid Insecticide. *Environ. Toxicol. Chem.* 12:871-878.

***Livingston, R.J. and D.A. Meeter. 1985. Correspondence of Laboratory and Field Results: What Are the Criteria for Verification? pp. 76-88. In: J. Cairns, Jr., ed., *Multispecies Toxicity Testing.* Pergamon Press, New York, NY.

Long, E.R. and P.M. Chapman. 1985. A Sediment Quality Triad: Measures of Sediment Contamination, Toxicity and Infaunal Community Composition in Puget Sound. *Marine Pollution Bulletin* 10:405-415.

***Luoma, S.N. 1995. Prediction of Metal Toxicity in Nature from Toxicity Tests: Limitations and Research Needs. pp. 610-659. In: A. Tessier and D. Turner, eds., *Metal Speciation and Bioavailability in Aquatic Systems.* John Wiley & Sons, Ltd., New York, NY.

Luoma, S.N. and J.L. Carter. 1993. Understanding the Toxicity of Contaminants in Sediments: Beyond the Toxicity Tests-based Paradigm. *Environ. Toxicol. Chem.* 12:793-796.

Luoma, S.N. and K.T. Ho. 1993. Appropriate Uses of Marine and Estuarine Sediment Toxicity Tests. pp. 193-225. In: P. Calow, ed., *Handbook of Ecotoxicology.* Blackwell Scientific, Oxford, U.K.

***Marcus, M.D. and L.L. McDonald. 1992. Evaluating the Statistical Basis for Relating Receiving Impacts to Effluent and Ambient Toxicities. *Environ. Toxicol. Chem.* 11: 1389-1402.

Marshall, J.S. 1978. Field Verification of Cadmium Toxicity to Laboratory *Daphnia* Populations. *Bull. Environ. Contam. Toxicol.* 20:387-393.

Mayer, F.L. and M.R. Ellersieck. 1986. *Manual of Acute Toxicity: Interpretation and Database for 410 Chemicals and 66 Species of Freshwater Animals.* Reference Source Publication 160. U.S. Fish and Wildlife Service, Dept. of Interior, Washington D.C.

McBride, G.B., J.C. Loftis, and N.C. Adkins. 1993. What Do Significance Tests Really Tell Us about the Environment? *Environ. Manage.* 17:423-432.

Moore, M.V., and R.W. Winner. 1989. Relative Sensitivity of *Ceriodaphnia dubia* Laboratory Tests and Pond Communities of Zooplankton and Benthos to Chronic Copper Stress. *Aquat. Toxicol.* 15:311-330.

***Mount, D.I. 1985. Scientific Problems in Using Multispecies Toxicity Tests for Regulatory Purposes. pp. 13-18. In: J. Cairns, Jr., ed., *Multispecies Toxicity Testing.* Pergamon Press, New York, NY.

Mount, D.I. 1995. Development and Current Use of Single Species Aquatic Toxicity Tests. pp. 97-104. In: J. Cairns, Jr. and B.R. Niederlehner, eds., *Ecological ToxicityTesting,* Lewis Publishers, Boca Raton, FL.

Mount, D.I. and T.J. Norberg-King, eds. 1985. *Validity of Effluent and Ambient Toxicity Tests for Predicting Biological Impact, Scippo Creek, Circleville, Ohio.* EPA 600/3-85-044. Duluth, MN.

Mount, D.I. and T.J. Norberg-King, eds. 1986. *Validity of Effluent and Ambient Toxicity Tests for Predicting Biological Impact, Kanawha River, Charleston, West Virginia.* EPA 600/3-86-006. Duluth, MN.

Mount, D.I., T.J. Norberg-King and A.E. Steen. 1986a. *Validity of Effluent and Ambient Toxicity Tests for Predicting Biological Impact, Naugatuck River, Waterbury,Connecticut.* EPA 600/8-86-005. Duluth, MN.

Mount, D.I., N.A. Thomas, T.J. Norberg, M.T. Barbour, T.H. Roush and W.F. Brandes. 1984. *Effluent and Ambient Toxicity Testing and Instream Community Response on*

the Ottawa River, Lima, Ohio. EPA 600/3-84-080. Duluth, MN.

Mount, D.I., A.E. Steen and T.J. Norberg-King, eds. 1985. *Validity of Effluent and Ambient Toxicity Testing for Predicting Biological Impact on Five Mile Creek, Birmingham, Alabama.* EPA 600/8-85-015. Duluth, MN.

Mount, D.I., A.E. Steen and T.J. Norberg-King, eds. 1986b. *The Validity of Effluent and Ambient Toxicity Tests for Predicting Biological Impact, Back River, Baltimore Harbor, Maryland.* EPA 600/8-86-001. Duluth, MN.

Mount, D.I., A.E. Steen and T.J. Norberg-King, eds. 1986c. *Validity of Ambient Toxicity Tests for Predicting Biological Impact, Ohio River, near Wheeling, West Virginia.* EPA 600/3-85-071. Duluth, MN.

Mount, D.R., K.R. Drottar, D.D. Gulley, J.P. Fillo, and P.E. O'Neil. 1992. Use of Laboratory Toxicity Data for Evaluating the Environmental Acceptability of Produced Water Discharge to Surface Waters. pp. 175-185. In: J.P. Ray and F.R. Engelhardt, eds., *Produced Water.* Plenum Press, New York, NY.

Neuhold, J.M. 1986. Toward a Meaningful Interaction Between Ecology and Aquatic Toxicology. pp. 11-21. In: T.M. Poston and R. Purdy, eds., *Aquatic Toxicology and Environmental Fate,* ASTM STP 921. American Society for Testing and Materials.

Niederlehner, B.R., K.W. Pontash, J.R. Pratt, and J. Cairns, Jr. 1990. Field Evaluation of Predictions of Environmental Effects from a Multispecies-Microcosm Toxicity Test. *Arch. Environ. Contam. Toxicol.* 19:62-71.

Niederlehner, B.R., J.R. Pratt, A.L. Buikema, Jr., and J. Cairns, Jr. 1985. Laboratory Tests Evaluating the Effects of Cadmium on Freshwater Protozoan Communities. *Environ. Toxicol. Chem.* 4:155-165.

Nimmo, D.R., D. Link, L.P. Parrish, G.J. Rodriguez, and W. Wuerthele. 1989. Comparison of On-site and Laboratory Toxicity Tests: Derivation of Site-specific Criteria for Unionized Ammonia in a Colorado Transitional Stream. *Environ. Toxicol. Chem.* 8:1177-1189.

Nimmo, D.R., M.H. Dodson, P.H. Davies, J.C. Greene, and M.A. Kerr. 1990. Three Studies Using *Ceriodaphnia* to Detect Nonpoint Sources of Metals from Mine Drainage. *J. Water. Poll. Contr. Fed.* 62:7-14.

Norberg-King, T.J. and D.I. Mount, eds. 1986. *Validity of Effluent and Ambient Toxicity Tests for Predicting Biological Impact, Skeleton Creek, Enid, Oklahoma.* EPA/600/8-86-002. Duluth, MN.

Obrebski, S., J.J. Orsi, and W. Kimmerer. 1992. Long-term Trends in Zooplankton Distributions and Abundance in the Sacramento-San Joaquin Estuary. Interagency Ecological Studies Program for the Sacramento-San Joaquin Delta Estuary. Technical Report No. 32.

*** Okkerman, P.C., E.J. V.D.Plassche, H.J.B. Emans, and J.H. Canton. 1993. Validation of Some Extrapolation Methods with Toxicity Data Derived from Multiple Species Experiments. *Ecotox. Environ. Safe.* 25:341-359.

***Parkhurst, B.R. 1995. Are Single Species Toxicity Test Results Valid Indicators of Effects to Aquatic Communities? pp. 105-121. In: J. Cairns, Jr. and B.R. Niederlehner, eds., Ecological Toxicity Testing, Lewis Publishers, Boca Raton, LA.

***Parkhurst, B.R. 1996. Predicting Receiving System Impacts from Effluent Toxicity. pp. 309-321. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

***Parkhurst, B.R., M.D. Marcus, and C.E. Noel. 1990. Review of the Results of EPA's Complex Effluent Toxicity Testing Program. Utility Water Act Group, Washington, D.C.

***Persoone, G. and J. Gillett. 1990. Toxicological Versus Ecotoxicological Testing. pp. 287-289. In: P. Bourdeau, E. Somers, G.M. Richardson, and J.R. Hickman, eds., *Short-term Toxicity Tests for Non-Genotoxic Effects,* John Wiley and Sons Ltd., New York, NY.

***Persoone, G. and C.R. Janssen. 1994. Field Validation of Predictions Based on Laboratory Toxicity Tests. pp. 379-397. In: I.R. Hill, F. Heimbach, P.I. Leeuwangh, and P. Matthiessen, eds., *Freshwater Field Tests for Hazard Assessment of Chemicals,* Lewis Publishers, Boca Raton, FL.

***Persoone, G., D. Calamari, and D. Wells. 1990. Possibilities and Limitations of Predictions from Short-term Tests in the Aquatic Environment. pp. 301-312. In: P. Bourdeau, E. Somers, G.M. Richardson, and J.R. Hickman, eds., *Short-term Toxicity Tests for Non-Genotoxic Effects,* John Wiley and Sons Ltd, New York, NY.

Pontash, K.W. and J. Cairns Jr. 1991. Multispecies Toxicity Tests Using Indigenous Organisms: Predicting the Effects of Complex Effluents in Streams. *Arch. Environ. Contam. Toxicol.* 20:103-112.

Pontash, K.W., B.R. Niederlehner, and J. Cairns, Jr. 1989. Comparisons of Single-species, Microcosm and Field

Responses to a Complex Effluent. *Environ. Toxicol. Chem.* 8:521-532.

Pratt, J.R., J. Mitchell, R. Ayers, and J. Cairns, Jr. 1989. Comparison of Estimates of Effects of a Complex Effluent at Differing Levels of Biological Organization. pp. 174-188. In: G.W. Suter and M.A. Lewis, eds., *Aquatic Toxicology and Environmental Fate,* ASTM STP 1007. American Society for Testing and Materials, Philadelphia, PA.

Richardson, B.J. and M. Martin. 1994. Marine and Estuarine Toxicity Testing: a Way to Go? Additional sitings from Northern and Southern hemisphere perspectives. *Marine Poll. Bull.* 28:138-142.

Roberts, J.R., D.W. Rodgers, J.R. Bailey, and M.A. Rorke. 1978. Polychlorinated Biphenyls: Biological Criteria for an Assessment of Their Effects on Environmental Quality. National Research Council of Canada, Ottawa.

Robinson, R.D., J.H. Carey, K.R. Solomon, I. R. Smith, M.R. Servos, and K.R. Munkittrick. 1994. Survey of Receiving-water Environmental Impacts Associated with Discharges from Pulp Mills. 1. Mill Characteristics, Receiving-water Profiles and Laboratory Toxicity Tests. *Environ. Toxicol. Chem.* 13:1075-1088.

Sasson-Brickson, G. and G.A. Burton, Jr. 1991. In Situ and Laboratory Sediment Toxicity Testing with *Ceriodaphnia dubia. Environ. Toxicol. Chem.* 10:201-207.

Schimmel, S.C., G.E. Morrison, and M.A. Heber. 1989a. Marine Complex Effluent Toxicity Program: Test Sensitivity, Repeatability and Relevance to Receiving Water Toxicity. *Environ. Toxicol. Chem.* 8:739-746.

Schimmel, S.C., G.B. Thursby, M.A. Heber, and M.J. Chammas. 1989b. Case Study of a Marine Discharge: Comparison of Effluent and Receiving Water Toxicity. pp. 159-173. In: G.W. Suter, II and M.A. Lewis, eds., *Aquatic Toxicology and Environmental Fate*: Eleventh Volume, ASTM STP 1007, American Society for Testing and Materials, Philadelphia, PA.

Sherman, R.E., S.P. Gloss, and L.W. Lion. 1987. A Comparison of Toxicity Tests Conducted in the Laboratory and in Experimental Ponds Using Cadmium and the Fathead Minnow (*Pimephales promelas*). *Water Res.* 1:317-323.

Siefert, R.E., S.J. Lozano, J.C. Brazner, and M.L. Knuth. 1989. Littoral Enclosures for Aquatic Field Testing of Pesticides: Effects of Chlorpyrifos on a Natural System. *Entomological Soc. Amer., Misc. Publ.* 75:57-73.

Slooff, W. 1985. The Role of Multispecies Testing in Aquatic Toxicology. pp 45-60. In: J. Cairns, Jr., ed., *Multispecies Toxicity Testing*, Pergamon Press, New York, NY.

Slooff, W. and J.H. Canton. 1983. Comparison of the Susceptibility of 11 Freshwater Species to 8 Chemical Compounds. II. (Semi) Chronic Toxicity Tests. *Aquati. Toxicol.* 4:271-282.

Slooff, W., J.A.M. van Oers and D. de Zwart. 1986. Margins of Uncertainty in Ecotoxicological Hazard Assessment. *Environ. Toxicol. Chem.* 5:841-852.

Smith, E.P. 1995. Design and Analysis of Multispecies Experiments. pp. 73-95. In: J. Cairns, Jr., and B.R. Niederlehner, eds., Ecological Toxicity Testing, Lewis Publishers, Boca Raton, FL.

Smith, R. 1994. Contract Report by EcoAnalysis Inc., Ojai, CA. Submitted to the State Water Resources Control Board, Sacramento, CA .

Sprague, J. 1995. A Brief Critique of Today's Use of Aquatic Toxicity Tests. *Human Ecol. Risk Assess.* 1: 167-170.

State Water Resources Control Board. 1990. Water Quality Control Plan for Ocean Waters of California (California Ocean Plan). SWRCB Resolution No. 90-27.

Stay, F.S., D.P. Larsen, A. Katko, and C.M. Rohm. 1985. Effects of Atrazine on Community Level Responses in Taub Microcosms. pp. 75-90. In: T.P. Boyle, ed., *Validation and Predictability of Laboratory Methods for Assessing the Fate and Effects of Contaminants in Aquatic Ecosystems, ASTM STP 865*, American Society for Testing and Materials, Philadelphia, PA.

Stephenson, R.R., and D.F. Kane. 1984. Persistence and Effects of Chemicals in Small Enclosures in Ponds. *Arch. Environ. Toxicol.* 13:313-326.

Swartz, R.C., F.A. Cole, J.O. Lamberson, S.P. Ferraro, D.W. Schults, W.A. DeBen, H. Lee II, and R. J. Ozretich. 1994. Sediment Toxicity, Contamination and Amphipod Abundance at a DDT-and Dieldrin-contaminated Site in San Francisco. *Environ. Toxicol. Chem.* 13:949-962.

Swartz, R.C., W.A. Deben, K.A. Sercu, and J.O. Lamberson. 1982. Sediment Toxicity and the Distribution of Amphipods in Commencement Bay, Washington, USA. *Marine Pollution Bulletin* 13:359-364.

Swartz, R.C., D. W. Schults, G. R. Ditsworth, W.A. DeBen, and F.A. Cole. 1985. Sediment Toxicity, Contamination,

and Macrobenthic Communities near a Large Sewage Outfall. pp. 152-175. In: T.P. Boyle, ed., *Validation and Predictability of Laboratory Methods for Assessing the Fate and Effects of Contaminants in Aquatic Ecosystems,* ASTM STP 865, American Society for Testing and Materials, Philadelphia, PA.

Swartz, R.C., D.W. Schults, J.O. Lamberson, R.J. Ozretich, and J.K. Stull. 1991. Vertical Profiles of Toxicity, Organic Carbon, and Chemical Contaminants in Sediment Cores from the Palos Verdes Shelf and Santa Monica, California. *Marine Environ. Res.* 31:215-225.

Underwood, A.J. 1995. Toxicological Testing in Laboratories Is Not Ecological Testing of Toxicology. *Human Ecol. Risk Assess.* 1: 178-182.

USEPA. 1984. *Ambient Water Quality Criteria for Cadmium 1984.* EPA 440/5-84-032. Washington, D.C.

USEPA. 1991. *Technical Support Document for Water Quality-based Toxics Control.* EPA/505/2-90-001. Washington, D.C.

USEPA. 1994a. Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms. 3rd ed. EPA 600/4-91/002. Cincinnati, OH.

USEPA. 1994b. Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Marine and Estuarine Organisms. 2nd ed. EPA 600/4-91/003. Cincinnati, OH.

Van den Brink, P.J., R.P.A. Van Wijngaarden, W.G.H. Lucassen, T.C.M. Brock, and P. Leeuwangh. 1996. Effects of the Insecticide Dursban 4E (Active Ingredient Chlorpyrifos) in Outdoor Experimental Ditches: II. Invertebrate Community Responses and Recovery. *Environ. Toxicol. Chem.* 15:1143-1153.

Van Wijngaarden, R.P.A., P.J. van den Brink, S.J.H. Crum, J.H. Oude Voshaar, T.C.M. Brock, and P. Leeuwangh. 1996. Effects of the insecticide Dursban 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: I. Comparison of short-term toxicity between the laboratory and the field. *Environ. Toxicol. Chem.* 15:1133-1142.

***Waller, W.T., L.P. Ammann, W.J. Birge, K.L. Dickson, P.B. Dorn, N.E. LeBlanc, D.I. Mount, B.R. Parkhurst, H.R. Preston, S.C. Schimmel, A. Spacie, and G.B. Thursby. 1996. Predicting Instream Effects from Wet Tests. pp. 271-286. In: D.R. Grothe, K.L. Dickson, and D.K. Reed-Judkins, eds., *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts.* SETAC Press, Pensacola, FL.

Weiss, C.M. 1976. Field Evaluation of the Algal Assay Procedure on Surface Waters of North Carolina. pp. 29-76. In: E.J. Middlebrooks, D.H. Falkenburg and T.E. Maloney, eds, *Biostimulation and Nutrient Assessment,* Ann Arbor Science, MI.

Yoder, C.O. 1991. Answering Some Concerns about Biological Criteria Based on Experiences in Ohio. pp. 95-104. In: *EPA Water Quality Standards for the 21st Century, Proceedings of a Conference.* USEPA, Washington, D.C.

2.0 Bibliography

***Boyle, T.P. 1985. Research Needs in Validating and Determining the Predictability of Laboratory Data to the Field. pp. 61-66. In: R.C. Bahner and D.J.Hansen, eds., *Aquatic Toxicology and Hazard Assessment,* ASTM STP 891. American Society for Testing and Materials, Philadelphia, PA.

***Cairns, J. Jr. 1993. Environmental Science and Resource Management in the 21st Century: Scientific Perspective. *Environ. Toxicol. Chem.* 12:1321-1329.

***Cairns, J. Jr., and J.R. Pratt. 1989. The Scientific Basis for Toxicity Tests. *Hydrobiologia* 188/189:5-20.

***Kimball, K.D. and S.A. Levin. 1985. Limitations of Laboratory Toxicity Tests: The Need for Ecosystem-level Testing. *Bioscience* 35:165-171.

***Maltby, L. and P. Calow. 1989. The Application of Toxicity Tests in the Resolution of Environmental Problems; Past, Present and Future. *Hydrobiologia* 188/189:65-76.

***Mount, D.I. 1994. *A Comparison of Strengths and Limitations of Limitations of Chemical Specific Criteria, Whole Effluent Toxicity Testing, and Biosurveys.* Contract report submitted to USEPA Office of Wastewater Enforcement and Compliance, Washington, DC.

***Parkhurst, B.R. and D.I. Mount. 1991. The Water Quality-based Approach to Toxics Control: Narrowing the Gap Between Science and Regulation. *Water Environ. Tech.* 3:45-47.

Appendix A
Single Species Tests with Effluent

The following consists of an interpretive summary of studies in which effluents were tested with single species toxicity tests and in which some ecological survey data were collected for comparative purposes.

A.1 Dickson et al. (1996)

A study (thesis project of R. Guinn, as summarized by Dickson et al., 1996) was conducted by the Institute of Applied Sciences at the University of North Texas to examine the effects of dechlorinating the effluent from a wastewater treatment facility (WWTP) on aquatic biological communities in the West Fork of the Trinity River, Texas. The WWTP effluent, at its discharge point, constitutes up to 96% of the river's flow during low flow periods. An objective of the study was to evaluate the relationships among effluent toxicity, river water toxicity, and biological community responses.

Field assessments were performed to determine resident biota and abiotic factors in the river both upstream and downstream of the WWTP. Effluent and ambient water toxicity were assessed with USEPA's 7-d *Ceriodaphnia* survival/reproduction and larval fathead minnow survival/growth tests. In addition, ambient water toxicity in the river was assessed *in situ* with caged organisms--fathead minnows and Asiatic clams (*Corbicula fluminea*). Two sampling sites (controls) were located upstream and five sites were downstream of the WWTP outfall. The first two sites below the outfall were within 1.25 miles of the discharge point and the remaining three sites were at various locations 17 miles or less downstream. Ecological surveys included fish and benthic macorinvertebrate collections. Ambient water toxicity testing was conducted with samples collected at all seven sites.

When this study was initiated, the WWTP was chlorinating its effluent. Effluent and ambient water toxicity testing, as well as biological sampling, was conducted during this period to establish a baseline for comparison with data collected after the implementation of dechlorination. During this pre-dechlorination period data were collected during two months (August and October).

With both the larval fathead minnow survival and growth endpoints, statistically significant toxicity (compared to the two upstream sites) was observed in the effluent and in ambient water from the first two sites downstream of the WWTP outfall; results were the same in August and October. Dechlorination of the water samples from the two sites below the outfall removed the toxicity. Statistically significant toxicity in the larval fathead tests was not observed in ambient water samples from sites 5, 6, and 7 downstream of the outfall.

Statistically significant toxicity (compared to the upstream sites) was recorded in the effluent and water samples from all sites downstream of the WWTP with the *Ceriodaphnia* tests (both survival and reproduction). Dechlorination of the toxic ambient water samples failed to remove the toxicity, suggesting that other contaminants were causing the water flea responses. In October, statistically significant toxicity (compared to upstream sites) was noted in the *Ceriodaphnia* tests with effluent and in water samples from the two downstream sites nearest the outfall, but not at sites 5, 6, and 7.

In the biological surveys, no fish were collected at the two sites below the WWTP outfall. Between 200 and 4,500 fish were collected at other sites on the river. Fish species richness, evenness, and diversity were fairly equivalent at all sites except the two below the outfall. Densities of benthic macroinvertebrates were lower at the two sites below the outfall than at the two upstream reference sites as well as at sites 5, 6, and 7, below the outfall.

Based on the data collected during the pre-dechlorination period, the authors predicted that effluent dechlorination would remove toxicity to larval fathead minnows and possibly restore the environment below the WWTP outfall so that those areas could be colonized by fish. Because the toxicity to the water flea could not be totally attributed to chlorine, the authors suggested that dechlorination might not alter *Ceriodaphnia* responses. Potential for impacts to instream biota was possible due to non-chlorine contaminants.

Following activation of the dechlorination system WWTP effluent and river water samples at all seven sites were collected and tested on a monthly basis for a total of 17 test periods. Dechlorination appeared to remove effluent and ambient water toxicity when larval fathead minnows were used to screen samples. Dechlorination did not remove all of the effluent or ambient water toxicity detected with *Ceriodaphnia*. The TIE identified disunion as a major cause of the effluent and ambient water daphnid toxicity.

During the pre-dechlorination period, caged fathead minnows did not survive at river stations 3 and 4, immediately below the WWTP outfall and approximately one mile downstream, respectively. With the exception of one of four testing periods after implementation of dechlorination, survival of caged fathead minnows at stations 3 and 4 was equivalent to all other stations. Juvenile *Corbicula* were exposed *in situ* for one month periods on five different test dates--one pre-dechlorination and four after initiation of dechlorination. Prior to initiation of dechlorination clam mortality was 100% at stations 3 and 4, while there was 100% survival at all five of the other stations. Post-implementation of dechlorination, clam survival at stations 3 and 4 was 100%. However, shell growth was significantly lower at stations 3 and 4 (compared to all other stations), suggesting the presence of an effluent contaminant other than chlorine. The *in situ* tests support the results observed in the laboratory toxicity tests with effluent and ambient water samples.

Following dechlorination, fish were present at all river stations, supporting the author's prediction of the possibility of recolonization at sites 3 and 4 with the implementation of dechlorination. However, in three of four surveys after dechlorination was initiated, the river station nearest the outfall was found to have fish assemblages dissimilar to those of the other stations. Macroinvertebrate surveys revealed significant improvement in diversity and evenness at stations 3 and 4 following initiation of dechlorination, although the total number of organisms was lower compared to the other stations.

In concluding, Dickson et al., (1996) state 1) "The results of this case study add to the growing weight-of-evidence to document a relationship between effluent toxicity (even chronic toxicity) and receiving system impacts for effluent-dominated systems" and 2) "We believe that establishing a quantitative relationship between WET test results, ambient toxicity, and receiving systems effects, as a means for validating WET test results, is not possible given the methods, approaches, and resources currently available. However, we believe the weight-of-evidence strongly supports that such a qualitative relationship exists."

A.2 Pontash et al. (1989)

These researchers compared microcosm (multiple species tests consisting of indigenous benthic macroinvertebrates and protozoans) responses to a complex effluent with responses observed in short-term estimates of chronic toxicity (*Ceriodaphnia* survival and reproduction). The predictive utility of these tests was evaluated in relation to observed effects in the stream receiving the complex effluent.

The results of this study demonstrated that the *Ceriodaphnia* reproduction response successfully estimated no effect concentrations for the assessment of aquatic community biological responses. Information from the multiple species tests provided more specific predictions than did the single species test.

The cladoceran reproduction results slightly underestimated the effects of the complex effluent on the receiving stream. *Ceriodaphnia* survival results in the laboratory toxicity tests underestimated instream impairments of the effluent. Similar findings had been made by Pontasch and Cairns (1991) in which laboratory toxicity tests with *D. magna* underestimated biological community impairments in the stream receiving the discharge. Underestimation refers to the situation in which the laboratory toxicity test indicates a higher effect concentration than that which actually causes instream impairments. On the other hand, Cairns and Cherry (1983) demonstrated, in tests with a power plant effluent, that single species test results can effectively predict ecosystem biological responses.

A.3 Niederlehner et al. (1990)

The predictive validity of a microcosm (multiple species) toxicity test was evaluated by Niederlehner et al. (1990). The study was conducted on a stream which receives a complex industrial discharge. A control site was located immediately upstream of the outfall, site 1 was approximately five meters downstream of the outfall; sites 2, 3, and 4 were 0.25, 1.4, and 6.4 km downstream of the outfall, respectively. Care was taken to select sites with similar characteristics, especially substrate type. The concept was to assure that the effluent was the major variable among the sites. Effluent dilutions at each of these sites was estimated using electrical conductivity. In addition to the microbial microcosm test, dilutions of the effluent also were tested with the 7-d *Ceriodaphnia* test. The instream measurements taken at each of the sites as indicators of biological community health, included species richness of protozoans and a semi-quantitative survey of benthic macroinvertebrates.

In both the microcosm and the water flea reproductive response tests the LOEC and NOEC were 3% and 1% effluent, respectively. In the field survey, significant effects on protozoan and macroinvertebrate species richness were seen at site 1, just below the outfall; estimated effluent concentration at this site was 14.1%. High percentages of chironomid species and low percentages of mayfly species were seen at sites 1, 2, and 3, but at site 4 the composition of the two groups was similar to the control site. Generally, chironomids are considered tolerant and mayfly species intolerant of water pollution.

If the species composition of these two groups is used as a sensitive indicator of ecosystem responses, then effluent effects were seen all the way down to site 3. Estimated effluent concentration at this site was 3.5%. The *Ceriodaphnia* reproduction and microcosm tests estimated the LOEC to be 3% effluent. Therefore, both tests reliably predicted instream biological community responses.

A.4 Diamond et al. (1994)

A rather detailed analysis of this publication is provided because the ecological survey, as well as other components, of this study represent the type of design and analysis that is to be avoided when attempting to assess the reliability of extrapolations from laboratory toxicity test data to instream responses. Our evaluation of the analysis of the data presented resulted in a conclusion that the effluent under study was adversely impacting the stream and river into which it was discharged. Diamond et al. (1994) concluded that the effluent was not impacting the stream.

Diamond and associates conducted a study on a wastewater treatment facility (WWTP) effluent, the stream (X-trib) into which the effluent was discharged, and the South Anna River (in Virginia) into which X-trib discharged. Chemical specific analyses and USEPA toxicity tests (USEPA, 1994a) were performed on effluent samples; stream bioassessments were implemented on X-trib, the South Anna River, and on two reference streams (other tributaries to Santa Anna River).

X-trib, the receiving water, was described as heavily channelized with concrete structures. WWTP effluent comprised approximately 98% of X-trib during low flow periods. The Santa Anna River was described as being forested over much of its watershed and apparently unimpaired by anthropogenic influences above its confluence with X-trib. Two sites on X-trib were selected, one in an open, sunny area above the WWTP point of discharge and the other in a shaded area below the point of discharge. The selection of these two sites appears unfortunate in that the two sites fail to match in habitat type; therefore, the primary variable between sites is more than effluent constituents.

Two reference sites were chosen; one on an open stream which discharged into Santa Anna River and was to serve as a matched or control site for the upstream site on X-trib. The second reference site was on a shaded stream which also discharged into Santa Anna River; this site was intended as a control for the lower site on X-trib. The authors indicated that the reference sites provided information on fauna capable of inhibiting X-trib. However, the authors concluded that the two reference sites on the Santa Anna River tributaries had better habitats for fish and macroinvertebrates than the X-trib sites. Therefore, these sites should be disqualified as reference sites because habitat differences rather than water quality could account for biological community differences. Selection of such sites reveals questionable study design and represents a serious flaw in this study. Interpretation of results are clearly confounded. Four sites were selected on the Santa Anna River. One site was above the confluence with the X-trib and the three other sites were downstream of the confluence.

Bioassessments focused on benthic macroinvertebrates and fish populations. Two types of bioassessments were performed. The first type involved introduced substrate at the sites. This substrate consisted of rocks collected at the upstream Santa Anna River site. The authors rationale for this procedure related to previous impact to X-trib and the Santa Anna River from toxic substances discharged from the WWTP. The authors fail to address the question of why fauna would not have naturally recolonized sites on X-trib and Santa Anna River if water quality had improved. From a biological perspective, the existing macroinvertebrate communities at a given site better represent water quality over time than introduced fauna. If the introduced substrate procedure is to be used, information on response time (to toxic substances and particularly metals, which tends to be slow as bioaccumulation occurs) of the introduced macroinvertebrates should have been provided, but was not. Furthermore, the introduced substrates were placed at each site for only four weeks; macroinvertebrate communities tend to respond slowly to metals and other toxicants which exert effects after bioaccumulation.

The authors placed much less emphasis on the second type of bioassessment procedure which was grab samples at each site. Clearly, however, these resident communities would be much more representative of bioaccumulative substances. Sampling was conducted during fall (October) and spring (April).

Toxicity tests were performed using the 7-d larval fathead minnow and *Ceriodaphnia* protocols. *Ceriodaphnia* tests were completed on two effluent samples taken in May and two collected in October. No sample revealed toxicity. Larval minnow tests were conducted with two effluent samples collected in October (neither indicated toxicity) and one sample taken in May. This May sample indicated significant toxicity. Unfortunately, two other effluent samples collected May (after the first May sample indicated toxicity) were not tested with larval fathead minnows. The failure to follow up on the first indication of toxicity was an experimental error. Furthermore, the very few effluent samples which were tested do not allow characterization of the WWTP effluent (WWTP effluents tend to show considerable temporal variability). Therefore, the authors' conclusion that the toxicity data indicated the effluent should not impact the receiving water biota is not supported by data presented. The seven day tests are not good measures of bioaccumulative impacts of metals.

Although replicates were included in the ecological survey, variability among the replicates was not reported and further complicates data interpretation.

Grab sample bioassessment data collected in the fall explicitly revealed that the X-trib sites did not correspond with the reference sites. According to several of the macroinvertebrate indices, the lower X-trib site was impaired compared to the upstream site and to the reference site; fish data also suggested that the lower site was impacted. As, indicated above, the introduced substrate (IS) data should be interpreted with caution; nonetheless, even these data imply that the lower X-trib site was impaired compared to the site above the discharge point. Perhaps more importantly, the fall grab sample at Santa Anna River sites downstream of the confluence with X-trib indicate that they were impacted. No fish data were reported for the Santa Anna River.

IS data were presented for only two X-trib sites and two Santa Anna River sites (those above and below the X-trib confluence). Failure to include other downstream sites, as well as the short exposure time (see above) limits the value of these data. Nevertheless, examination of the dominant taxa on the IS suggests that water quality at the upstream Santa Anna River site was better than at the site below the confluence. Although differences between means of several of the bioassessment metrics when comparing the upstream and downstream river sites are large, they were reported as not being statistically different. This is likely due to the fact that an analysis of variance was applied to data from both X-trib and Santa Anna River. This application does not seem justified given that the tributary and the river are such different habitats. Moreover, the two X-trib site macroinvertebrate indices means were frequently so large that variation in the data sets masked differences between Santa Anna River sites.

In the spring collections, the macroinvertebrate grab samples analyzed from the lower X-trib site indicated that it was impacted compared to both the upstream and reference sites. IS data from X-trib for the spring sampling period were not presented. During the spring grab samples for macroinvertebrate analysis were taken at only two Santa Anna River sites, the upstream and the downstream site nearest the confluence. The absence of data from the other two Santa Anna River sites further limit this data set. Although there were few apparent statistical differences between macroinvertebrate indices from the two sites, biological community composition indicated that the site below the X-trib confluence was impacted; the same trend was noted in the IS data.

Data presented in this publication do not support the authors' contention that neither X-trib nor the Santa Anna River are impacted by WWTP effluent constituents. They attribute the impacts indicated in X-trib by the grab sample macroinvertebrate data to habitat limitations. If this is actually the case, one must conclude that their study design was flawed from the outset. However, their conclusion is not supported by the differences between the upstream and downstream sites (as shown in all types of bioassessment data).

A.5 Birge et al. (1992)

Birge et al. were involved in a relatively long-term study of the effluents produced by the Paducah Gaseous Diffusion Plant (PGDP) and the streams into which these effluents are discharged, Big Bayou and Little Bayou Creeks. Toxicity, chemical, and bioassessment monitoring were performed. Specifically investigating the relationship between effluent/ambient toxicity test results and instream biological responses was not a stated goal of this study, but some interesting information can be gleaned from their results.

The PGDP has 16 potential discharge points into the two creeks. The focus was on eight of these effluents because they constitute continuous discharge to the streams. Seven-day *Ceriodaphnia* and larval fathead minnow tests were conducted with 51 undiluted effluent samples and with 37 stream samples collected on four different occasions. Instream biological assessments (primarily number of taxa and density) of benthic macroinvertebrates were performed at three separate times (1987-91) at eight sites. One of these sites was above discharge points, three sites were at increasing distances from the last discharge point, and the other four sites were a gradient within the spatial range of the several discharge points.

Bioassessment data were collected in 1987 through mid-1988. Four separate sampling events indicated instream biological impairment at sites within the range of discharge points (as compared to the upstream reference site). At sites below the last discharge point, there appeared to be progressive recovery as measured by number of taxa and density of macroinvertebrates. Ecological survey data collected in 1990 and 1991, but not 1989, were similar to those collected in earlier years. The toxicity testing data are summarized below:

Larval Fathead Minnows

Effluent:
Significant mortality in 31/51 samples (61%)

Ambient water downstream:
Significant mortality in 18/37 samples (49%)

Ceriodaphnia

Effluent:
Significant toxicity in 11/51 samples (22%)

Ambient water downstream:
Significant toxicity in 4/37 samples (11%)

The difference in undiluted effluent and ambient water toxicity appeared to be primarily a dilution phenomenon. Generally, effluent toxicity predicted instream toxicity when dilution was taken into consideration. On a qualitative basis, instream toxicity reliably predicted instream biological responses.

A.6 Pratt et al. (1989)

The potential impact of a municipal sewage effluent on Smith River (Virginia) was evaluated using acute and chronic single species toxicity tests and a microcosm test consisting of indigenous microbiota. Effect levels obtained in the single species and microcosm studies on effluent were compared with the estimated instream waste concentration (IWC) and with results of an ecological survey.

The study consisted of two sites upstream of the wastewater treatment facility (WWTP) and three sites below the outfall of the facility. A survey of benthic macroinvertebrates and protozoan communities was conducted at each of these sites. Effluent from the WWTP was tested in the 7-d larval fathead minnow and *Ceriodaphnia* tests, as well as in the indigenous species microcosm test. The microcosm test consisted of microorganisms. River water samples collected at one site above the WWTP outfall and at all sites below the outfall were tested in the 7-d *Ceriodaphnia* test, but not the larval fathead minnow test.

The macroinvertebrate data suggested impairments (compared to the upstream control) at the first two sites below the WWTP outfall, with recovery at the third site. The *Ceriodaphnia* tests did not show significant toxicity in water samples collected at the two impacted sites. LOECs were 30% effluent in the microcosm and *Ceriodaphnia* tests and 15% in the larval minnow test. Maximum IWC was estimated to be 9.5% effluent; NOECs in all laboratory tests were 10% effluent. Therefore, both the effluent and ambient water single species tests underestimated instream impacts.

A.7 Crossland et al. (1992)

The toxic fraction (chlorinated ethers) of a petrochemical manufacturing plant effluent was studied in simulated outdoor streams. Four different concentrations of the effluent extract were tested in the streams; exposure was for 21 to 28 days. Two untreated streams served as controls.

The LOEC and NOEC (*Gammarus pulex*) in the simulated streams were 0.86 µg/L and 0.44 µg/L, respectively. These values were compared to the NOEC from a 7-d *Daphnia magna* laboratory test; the reproduction NOEC in this test was 1.0 µg/L. Although a 21-day *Daphnia* test would have been more appropriate, the result from the single species test was an effective qualitative predictor of effect concentration in the mesocosm; the *Daphnia* data slightly overestimated the artificial stream effect concentration.

A.8 Robinson et al. (1994)

These investigators conducted an examination of the relationship between environmental responses at 11 pulp mills, their pulping processes, degree of effluent treatment, and bleaching technologies. Water samples from upstream and downstream of the pulp mill discharge points were screened in the 7-d larval fathead minnow and *Ceriodaphnia* tests. These data were compared to physiological data collected from fish and benthic macroinvertebrate data from above and below the discharge points.

At four of 11 pulp mills the benthic macroinvertebrate communities were characterized as highly impacted below the discharge point compared to upstream sites. Statistically significant toxicity was detected in water samples downstream of all four of these mills in the larval fathead test, but at only one site in the *Ceriodaphnia* test. These four mills only had primary effluent treatment. Although the larval minnow test reliably indicated instream impacts at the four sites, the *Ceriodaphnia* test was less effective.

Neither of the single species tests predicted the physiological impairments seen in fish collected below the pulp mill outfalls. Physiological responses associated with reproductive dysfunction (decreased sex steroid levels and gonad size) and other disturbances (increased liver size and enzyme abnormalities) were observed in fish collected below pulp mill discharge points regardless of mill process, bleaching technology, or effluent treatment. This study represents a case in which the single species tests failed to predict (i.e., underestimated) instream impacts of effluents.

A.9 Sasson-Brickson and Burton (1991)

In situ exposures of *C. dubia* were conducted in a stream know to be impacted (based on benthic macroinvertebrate and fish community data) by several effluents. The *C. dubia* were in sediment exposure chambers placed in the stream for 48 h at an impacted site and at a reference site. Sediments from the impacted and reference sites also were tested in the laboratory with *C. dubia* using sediment solid phase, interstitial water, and elutriate tests.

Both the *in situ* and laboratory tests indicated statistically significant sediment toxicity; the responses in the laboratory tests were greater than in the *in situ* exposures. The authors concluded that the *in situ* exposures proved to be sensitive indicators of both degraded and nondegraded stream conditions. They also implied that the *in situ* responses were more reliable than the laboratory responses. This may not be a valid conclusion since neither the

laboratory nor the *in situ* responses were quantitatively correlated with instream biological community impacts.

A.10 Barbour et al. (1996)

Barbour et al.(1996) summarized studies conducted by Ohio EPA (see also Yoder, 1991) in which agreement between data from 48-h *C. dubia* and fathead minnow toxicity tests and from biosurveys were analyzed. Toxicity tests were performed on effluent and in some cases on mixing zone water samples. These authors surmised that the Ohio EPA analysis indicates that "the observance of acute toxicity, or lack thereof, in an effluent and to a lesser degree in mixing zones is not necessary, reflected by the instream communities." According to these authors other impacts often pre-empted or masked effects of toxicity. The authors concluded, "These results should not be misconstrued to claim toxicity testing is an invalid assessment and regulatory tool."

Indeed, caution should be used in making conclusions from the Ohio EPA data for several reasons. Toxicity tests were not performed on water samples collected at biosurvey sites. No information was provided on the degree of effluent dilution at each of the biosurvey sites. It appears that toxicity tests were performed on whole effluent, without dilution series to assess effect concentrations. Predicting biological community impacts based on the results of one toxicity test on effluent (or mixing zone sample), as the authors indicate, is unsound.

Barbour et al. (1996) also summarized similar studies performed by the Florida Department of Environmental Protection (DEP). In this project, 48 h toxicity tests with *Ceriodaphnia* and *Notropis leedsi* (a marine minnow) were performed on effluents from 107 facilities classified into several industrial categories. Macroinvertebrate surveys were conducted on streams into which the facilities discharged. Comparisons were made between effluent toxicity and biosurvey data.

Effluent toxic, stream site impaired = 24.0%;

Effluent toxic, stream site not impaired = 10.7%;

Effluent not toxic, stream site impaired = 41.3%;

Effluent not toxic, stream site not impaired = 24.0%.

Combining data from all facilities the following relationships between effluent toxicity and instream survey data were obtained. Toxicity tests reliably "predicted" instream conditions in 48% of the 107 situations. "False positives" were relatively rare (10.7 %). "False negatives" were much more common (41.3 %). Florida DEP attributed biological impairment at a large portion of the "false negative" sites to non-effluent related factors.

Barbour et al. (1996) concluded that lack of agreement in this study was not necessarily due to contradiction between the toxicity testing and biosurveys. This conclusion seems valid since in many cases biological impairment was due to causes other than effluent toxicity. Also, toxicity tests were performed on only one sample from each facility (as indicated above, one sample is not likely to characterize the effluent of a facility). Furthermore, the same cautions as mentioned above in regard to the Ohio EPA data apply here. That is, toxicity tests were not performed on water samples collected at biosurvey sites. No information was provided on the degree of effluent dilution at each of the biosurvey sites. It appears that toxicity tests were performed on whole effluent, without dilution series to assess effect concentrations.

A.11 Mount et al. (1992)

During fossil fuel production water pumped from the formation is separated and discarded, frequently into marine or freshwater environments. This fraction, commonly termed "produced water" can contain a diverse array of contaminants including brine, hydrocarbons, heavy metals, surfactants, and corrosion inhibitors.

Mount and colleagues reported on a series of laboratory and field studies which were conducted on produced water from a coal bed methane operation in the Cedar Cove Degasification Field of Alabama. The produced waters were discharged into Little Hurricane Creek. The primary goal of the studies was to determine the environmental acceptability of discharging produced water into this creek.

Toxicity tests were performed on the produced water using USEPA's fathead minnow and *Ceriodaphnia* tests; the cladocerans proved to be the more useful monitoring tool in these studies. Concurrent with the laboratory toxicity tests, a series of instream surveys were performed on Little Hurricane Creek. Based on these data the authors concluded, "Research conducted at Cedar Cove suggests that laboratory toxicity tests can be used to predict instream effects of produced water discharge.

Appendix B
Single Species Tests with Individual Chemicals

The following consists of an interpretive summary of studies in which single species tests were used to assess the toxicity of a single chemical or combination of a small number of chemicals and predict effect concentrations on aquatic ecosystem biological responses.

B.1 Organic Chemicals: Pesticides

B.1.1 Hansen and Garton (1982)

These investigators assessed the ability of single species toxicity test results to reliably predict the effects of the insecticide diflubenzuron on complex laboratory stream communities. The single species tests included five "chronic tests" with five different species, including a 21-d *Daphnia magna* test. The laboratory stream communities were stocked from a natural source and then exposed to the pesticide for five months. Effects on the stream communities were appraised at the functional group level using biomass and diversity.

For *Daphnia*, the 21-day LC50 for this pesticide was 0.1 µg/L. Statistically significant effects on invertebrate shredder, scraper, and collector/gather/filterer functional groups were evident in the mesocosm after 5 to 7 months exposure at a nominal concentration of 0.1 µg diflubenzuron/L. The *Daphnia* toxicity test results appeared to reliably predict the responses of aquatic invertebrate communities. However, there is uncertainty in these data. For example, duration of exposure in the laboratory and field setting were very different and mesocosm exposures were not analytically confirmed (dissipation and degradation usually results in lower than predicted exposure concentrations). LC50s are not necessarily the optimal predictor tool; however, if environmental concentrations of a chemical approach the LC50 level, biological community impairments are probable. Another confounding factor in this study was that the control populations declined during the five month course of the study.

Although there were uncertainties and confounding factors in this study, the correspondence between laboratory and field effect concentration supports the hypothesis that laboratory test results are predictive indicators of direct effects in the environment. Concentrations below the laboratory-determined LC50 were not tested in the mesocosms, so it is not possible to know whether field effect levels were overestimated.

B.1.2 Baughman et al. (1989)

To evaluate the usefulness of laboratory toxicity tests in predicting fenvalerate (a pyrethroid insecticide) impacts, Baughman et al. conducted laboratory and field tests with the grass shrimp (*Palaemonetes pugio*). Two types of laboratory tests were conducted: 96-h static-renewal tests and 6-h pulse dose exposures. The response (endpoint) compared between laboratory and field was the LC50.

Response of grass shrimp in the field was similar to the laboratory toxicity tests (i.e., concentrations which were shown to produce lethality in the laboratory also caused mortality in field settings). These results indicated that physical and chemical factors in natural ecosystems did not appreciably modify the toxicity of fenvalerate.

In this study, laboratory test data were not extrapolated across species, but rather to the same species in natural stream conditions. Although not a powerful support of the reliability of laboratory single species test results as predictors of instream biological impacts, this study does show a correspondence between laboratory and natural ecosystem effect concentrations.

B.1.3 Clark et al. (1987)

Clark and colleagues scrutinized laboratory toxicity test results as predictors of effects of fenthion, an organophosphorus insecticide, on caged animals in field settings. The laboratory tests were 96-h mortality determinations on a mysid (*M. bahia*), the pink shrimp (*Penaeus duoraum*), the grass shrimp (*Palaemonetes pugio*), the sheepshead minnow (*C.variegatus*). The responses used for comparisons were 24-h and 48-h LC50s. Caged animal tests and environmental chemical studies (measurements of fenthion) were executed in a bay and a pond connected to Santa Rosa Sound, as well as in an estuarine bay.

Results of this study reveal that the laboratory-derived LC50s were reasonable predictors of mortality to the same species in the field, but only when laboratory and field exposure regimes were similar. The laboratory LC50s were not effective predictors of sublethal effects. As in many of the studies summarized above, the findings of this study disclose that physical and chemical factors in aquatic ecosystems did not appreciably alter the toxicity of this pesticide. Caging of animals in the field did not allow for

possible avoidance behavior. The advantages and limitations of using acute exposure LC50s as predictors of instream biological responses were mentioned above.

B.1.4 Fairchild et al. (1992)

Population, community and ecosystem level responses to pulse doses of esfenvalerate, a pyrethroid insecticide, were studied in experimental aquatic mesocosms. Different mesocosms were dosed at nominal concentrations of 0, 0.25, 0.67, and 1.71 µg/L esfenvalerate (each concentration had triplicated mesocosms). The pulse dosings were 15 minute applications to achieve the nominal concentrations every two weeks for a total of three months.

Static acute (48-h) toxicity tests with the insecticide were conducted with *D. magna* and provided an LC50 of 0.27 µg/L esfenvalerate; neither a LOEC or NOEC were reported. This laboratory effect level was compared to effect levels seen in the mesocosm portion of the study.

In the mesocosm component of the study, zooplankton and benthic macroinvertebrate populations were significantly decreased at the pulse dose treatment of 0.25 µg/L. There were also shifts in community composition and dominance at this treatment level. This was the lowest pulse dose tested, so an NOEC was not established. The effect concentration in the mesocosm was compatible to the laboratory generated 48-h LC50 for this pesticide. With the difference in exposure patterns and durations in the laboratory (single species) test and multispecies mesocosm studies, it is remarkable that a mesocosm effect concentration corresponded so well with the laboratory test results.

In the mesocosm, 0.67 µg/L esfenvalerate reduced survival, biomass, and reproductive success of bluegill sunfish. The laboratory LC50 for juvenile bluegills exposed to esfenvalerate for 96 h ranged from 0.42 to 1.35 µg/L (Mayer and Ellersieck, 1986). This finding indicates that a laboratory effect concentration translates reliably into a field effect level. Also inherent in this observation is that the complex, multivariate conditions in the mesocosm did not appreciably modify the toxicity (i.e., bioavailability) of chemicals seen in highly controlled laboratory studies.

Overall, the results of this study imply that single species toxicity test results can qualitatively predict effect concentrations in more complex, multivariate systems. Another study (Little et al., 1993) with fenvalerate also indicated that laboratory determined effect concentrations were reliable predictors of effect concentrations in natural systems. For fenvalerate, and possibly other pesticides which have relatively short half-lives and may not exist in aquatic ecosystems for extended periods, acute toxicity test endpoints may be reliable predictors of biological community responses.

B.1.5 Slooff (1985)

A multiple species microcosm toxicity test was conducted in the Netherlands to determine the NOEC for the herbicide, dichlorbenil. An NOEC was also determined for *Daphnia magna* in the 21-d short-term estimate of chronic toxicity.

The microcosm NOEC from a 400 day exposure was 0.3 µg/L. The NOEC, 0.1 µg/L, determined in the *Daphnia* test was a qualitatively accurate predictor of the mesocosm no effect level. In Slooff's review of the literature on dichlorbenil, the NOEC determined from 167 field exposures of various species was also 0.1 µg/L.

After reviewing other data in the literature and in relation to these data, Slooff (1985) submits that multiple species (micro- or mesocosm) toxicity test results are not better predictors of aquatic ecosystem responses than are single species toxicity test results. He concludes that the multiple species test results have many uses, but that, at their current stage of development, they do not improve predictions of ecosystem impairments.

B.1.6 Larsen et al. (1986)

Microcosm test data have been proposed as having more ecological relevance than laboratory indicator species test results. Larsen and co-workers compared the predictive reliability of "surrogate" species and mesocosm toxicity test results with responses in experimental ponds to the herbicide, atrazine. This study compared the responses of algal tests, a algal microcosm, and experimental ponds exposed to similar concentrations of the herbicide. Eight different algal species were included in the indicator species tests. The endpoints used for comparisons in all three systems were EC50s (the chemical concentration at which 50% of the test population exhibits a response).

According to these investigators, the basic similarity among the EC50 values across test systems suggests that results from a combination of single species tests or from the mesocosm provided a reasonable estimate of the concentration of atrazine that produced similar effects on the experimental pond. Both the lowest and highest EC50 came from single species tests. These authors conclude that, "because broad ranges in species sensitivities occur, use of only a few test species might not offer sufficient environmental protection." Improvement in predictive ability occurs when several species are used as test organisms. Although this study provided valuable information, the EC50 endpoint may not be the most realistic response measure to compare tests. One would predict that a concentration of chemical(s) which is high enough to affect 50% of a test population, has a high potential of evoking significant biological community responses.

Single species toxicity tests, microcosm, and outdoor experimental pond exposures have been employed by other investigators (Stay et al., 1985; de Noyelles and Kettle, 1985) to ascertain the effects of atrazine on algal primary production. Both the single species and microcosm tests were predictive of atrazine concentrations which significantly reduced production in the outdoor ponds. However, recovery from atrazine stress was not predicted by the laboratory tests. In the single species and microcosm tests there was only limited recovery whereas the pond communities recovered more quickly because sensitive algal species were replaced by algal species *more resistant* to atrazine. The ecological significance (over time) of this shift to more chemically resistant assemblages was not discussed and is unknown. Composition could change at all trophic levels due to the shift in algal species. All algal species were affected by atrazine in all test regimes, but only the pond study revealed assemblage shifts.

B.1.7 Crossland (1984)

Studies with the insecticide methyl parathion revealed that laboratory single species toxicity test results underestimated the secondary effects (indirect effects that are not represented by direct action of a chemical on an individual species, but rather result from interrelationships among components of a biological community) of this pesticide in outdoor ponds. The concentrations of methyl parathion eliciting toxicity, and thus decreasing populations, in zooplankton and benthic macroinvertebrates in the outdoor ponds were reliably predicted by the laboratory single species toxicity test results.

The decreased populations of mayfly larvae and daphnids, cased by methyl parathion, secondarily resulted in blooms of filamentous algae. Death and decay of the algae in turn decreased dissolved oxygen resulting in death of fish. Loss of invertebrate food items also caused reduced fish populations and smaller sized fish.

B.1.8 Stephensen and Kane (1984)

The fate and biological effects of the insecticides methyl parathion and linuron in outdoor ponds were studied. The relative sensitivities (response per concentration) were similar in both the laboratory and ponds for both pesticides. Furthermore, the response concentrations determined for *Daphnia magna* in the laboratory correlated closely with effect concentrations in the outdoor pond. The authors concluded that biotic and abiotic factors existing in ponds did not alter the toxicity (i.e., bioavailability compared to the laboratory tests) of these two pesticides.

B.1.9 Van Wijngaarden et al. (1996)

Using the insecticide Dursban 4E (active ingredient chlorpyrifos, an organophosphorus pesticide) these investigators compared the results of laboratory indicator species toxicity tests with laboratory tests on indigenous species, as well as with data from outdoor mesocosm tests. Mesocosms were sprayed once with the intent of achieving nominal chlorpyrifos concentrations of 0.1, 0.9, 6, and 44 µg/L. Analytical measurements of chlorpyrifos were used to determine exposure and effect concentrations. Effects in the mesocosms were assessed by sampling zooplankton and macroinvertebrates; in addition, *in situ* cage experiments were performed with several species.

The indicator species, *D. magna*, was almost as sensitive to chlorpyrifos as the indigenous species. The difference between the laboratory EC50s for the daphnid (1.0 µg/L) and that for the most sensitive indigenous species, *Gammarus pulex* (0.8 µg/L) was small, suggesting that the indicator species was not more sensitive to the insecticide. Effect concentrations (for nine invertebrate species) determined in single species laboratory tests were compared to effect concentrations derived in the mesocosm exposures.

The authors concluded that laboratory single species EC values were reliable estimators of mesocosm ECs, differing by less than a factor of three for the seven species studied. Essentially the same conclusions were reached when comparing ECs from the laboratory toxicity tests with those from the cage experiments. These data indicate that chlorpyrifos bioavailability was not significantly reduced under the mesocosm conditions.

Although there was considerable spatial and temporal variation of chlorpyrifos concentrations in the mesocosm exposures, ECs determined under those conditions were similar to the laboratory ECs obtained under constant exposure regimes (i.e., variable and constant exposure regimes led to comparable effects). According to these authors, laboratory single species toxicity test results can be used to estimate direct effects in field populations.

In a subsequent publication, van den Brink et al. (1996) reported that recovery of invertebrate populations after the single application of chlorpyrifos required three to six months. The investigators also suggested that "safe" concentrations determined in short-term single species laboratory toxicity tests are sufficient to protect invertebrate communities.

Several other investigators (Eaton et al., 1985; Brock et al., 1992; Leeuwangh et al., 1994) also concluded that the direct effects of chlorpyrifos on aquatic invertebrate communities can be reliably predicted on the basis of laboratory single species toxicity data; that is, population responses observed in microcosms and mesocosms were consistent with laboratory single species toxicity test results.

B.1.10 Kersting and van Wijngaarden (1982)

The effects of chlorpyrifos were studied in a laboratory microcosm system. *Daphnia magna* was the herbivore component in the microcosm; this species was also the subject of a laboratory 48-h lethality test. The microcosms

received a single application of chlorpyrifos, and the responses were followed for 130 days.

Chlorpyrifos concentration in the *Daphnia* component of the mesocosm was 0.5 µg/L on day 1, decreasing to 0.2 µg/L by day 7. The laboratory 48-hour LC25 for *Daphnia* was 0.4 µg/L. Although pesticide concentration decreased rapidly to below the LC25, *Daphnia* populations in the exposed microcosms decreased 36% and 42% in the two replicates. Populations recovered within two weeks. The ecological significance of the magnitude and duration of population declines is unknown.

Arguably, the laboratory LC25 was an effective predictor of the population decline in the mesocosm. However, chlorpyrifos treatment resulted in other biotic and abiotic changes in the microcosms. The laboratory *Daphnia* tests underestimated these other, "secondary" mesocosm effects.

B.1.11 Siefert et al. (1989)

The effects of chlorpyrifos in natural pond enclosures were investigated. The pesticide was applied once to sets of replicate ponds to achieve three different test concentrations; pond concentrations of the pesticide were analytically monitored. Phytoplankton, periphyton, zooplankton, benthic macroinvertebrates, and fish were sampled periodically up to 30 d post-application. Limitations in this study include the absence of normal exchange between the enclosures with the remainder of the pond; also chlorpyrifos adsorbed to the wall material of the enclosures (this problem relates mostly to environmental fate, rate of loss, but also to a decrease in potential exposure); this difficulty was partially offset by the monitoring of pesticide concentrations in the water column.

The targeted concentrations were 20, 5, and 0.5 µg/L. Chlorpyrifos concentrations decreased rapidly after application (see above) to 10, 1, and 0.2 µg/L by day two post-application. Cladocerans were the most sensitive of the zooplankton species, with all five identified species showing dramatic and statistically significant population declines at the lowest chlorpyrifos concentrations. Chironomids were the most sensitive benthic macroinvertebrates, with 9 of 10 of the identified species responding to the lowest chlorpyrifos treatment with statistically significant population declines.

Laboratory determined acute toxicity LC50s (54 species) from the literature were compared to the pond effect concentrations. In general, the single species LC50 values were higher than the LOEC determined in the pond study (i.e., LC50s underestimated biological community responses); LC50s from *Daphnia* and *Gammarus* were the most accurate forecasters of the pond LOEC. Significant reductions in growth rates in larval fish, not predicted by direct effects of chlorpyrifos, were also noted in this study. The authors attributed these secondary effects to chlorpyrifos-caused declines of invertebrate forage organisms.

B.2 Additional Organic Chemicals

B.2.1 Cooper and Stout (1985)

The effects of p-cresol on the biota in outdoor experimental stream channels (analogs of natural streams) were compared to the results of single species tests with this chemical. Three hypotheses were tested:

1) The transfer of laboratory acute toxicity test results to field situations is possible without serious distortion.
2) Data from single species tests with p-cresol will yield similar results as multiple species, community level, tests.
3) Pulsed exposures with short time intervals between events will produce the same ecological responses as continuous exposure with the same integral of exposure (integral of concentration X time).

In regard to the first hypothesis, data from this study showed that the acute toxicity tests with fathead minnows, large mouth bass, small mouth bass, damsel fly larvae, and amphipods estimated survivorship rates consistent with results of the experimental stream experiments. These investigators also concluded that results of the single species tests were effective predictors of community level responses. The third hypothesis was found to be untrue in that the pulse exposure (with same integral) produced greater impacts than continuous exposure. These pulse-response results are useful in interpreting data collected in agricultural settings where aquatic communities may be exposed to pulses of pesticides.

B.2.2 Dorn et al. (1991)

These researchers undertook a project to estimate the environmental effects of a choloether fraction from a chemical plant effluent. The chemical plant effluent had been shown to be toxic to sheepshead minnows (Cyprinodon varigatus) and a mysid (*Mysidopsis bahia*). Toxicity identification evaluation (TIE) procedures demonstrated that the primary causes of toxicity was a mixture of pentacholroethers.

To gauge effect concentrations of the chloroether fraction, laboratory toxicity tests were executed with *Daphnia magna*, fathead minnow larvae, and *Mysidopsis bahia*. Effect concentrations for the chloroether fraction also were assessed in outdoor artificial streams.

The most sensitive indicator species was the water flea; the NOEC for this species was 1.0 mg/L. NOECs in the outdoor streams were 0.44 and 0.26 mg chloroethers/L for *Gammarus* and rainbow trout, respectively. The laboratory effect concentrations for the chloroethers in single species

tests were reliable qualitative predictors of effect concentrations in the outdoor stream experiments. The outdoor stream communities were somewhat more sensitive to the toxicants than indicated by the laboratory single species tests.

In a follow-up study (Crossland et al., 1992), a range of chloroether fraction concentrations was tested in outdoor artificial streams. Exposure was for 28 days. Four different concentrations were tested; there were no replicate streams except for the control treatment. Three mesh bags of macroinvertebrates were introduced into each stream. However, the number of benthic macroinvertebrates of a given species was not equivalent in the different treatment groups at the time of pretreatment sampling; thus statistical comparisons among the treatments was not possible. Furthermore, there was considerable variation among "replicates" within each stream. Feeding rates of the amphipod, *Gammarus pulex,* also were assessed in the artificial streams. These and other factors render interpretation of macroinvertebrate data difficult.

Gammarus numbers were significantly reduced at a chloroether concentration of 0.86 mg/L, but not at 0.44 mg/L (the NOEC). Invertebrate drift (possibly indicating an unhealthy condition also appeared to be increased at chloroether concentrations of 0.44 m/L and above. In laboratory 21-d *Daphnia magna* tests, the chloroether NOEC was 1.0 mg/L and the LOEC was between 1 and 2.5 mg/L. Although comparable effect concentrations were seen in the laboratory single species test and the artificial stream data, the outdoor populations were somewhat more sensitive to the chloroethers.

B.2.3 Fairchild et al. (1993)

Laboratory and field studies were conducted with linear alkylbenzene sulfonate (LAS, an anionic surfactant), by Fairchild et al., to evaluate the use of laboratory-generated NOECs for protecting aquatic ecosystems. Laboratory toxicity tests included the 7-d fathead minnow test and a 7-d test with the freshwater amphipod, *Hyalella azteca.* A series of these tests with exposures to a range of LAS concentrations resulted in a laboratory estimate of a NOEC. This laboratory test predicted NOEC was then tested in the field with a 45-d exposure in outdoor experimental streams (three replicates).

In these experimental streams, exposure to LAS concentrations equivalent to the laboratory NOECs, no biological community impairments were seen as gauged by surveys of benthic macroinvertebrates, periphyton growth, detrital processing, and fathead minnow populations. The authors concluded that their results indicated that the laboratory-generated NOEC for LAS predicted environmental protective concentrations. Results of this study do not demonstrate that concentrations above the laboratory NOEC would have engendered impacts in the outdoor experimental streams, but do suggest that the single species toxicity test results can be useful tools in predicting environmentally "safe" concentrations.

B.2.4 Boyle et al. (1985)

These investigators compared the responses (survival and growth) of bluegill sunfish and large mouth bass exposed to fluorene (a polynuclear aromatic hydrocarbon) in laboratory 30-d partial life cycle tests to the responses of the same species exposed in outdoor ponds.

The laboratory toxicity test results underestimated the response of these fish species in the outdoor ponds. Moreover, the responses in the experimental ponds were more sensitive to fluorene (e.g., occurred at a lower concentration than in the laboratory tests). To the contrary, laboratory toxicity tests overestimated responses of zooplankton, phytoplankton, and some insect populations to fluorene.

B.2.5 Giddings and Franco (1985)

The effects of a synthetic coal-derived crude oil were assessed in outdoor ponds and indoor microcosms. The results of these tests were compared with data from laboratory single species toxicity tests. Response concentrations were similar in the microcosms and pond studies. A "safe" exposure concentration for this organic compound was derived from the pond study. Without an application factor, the USEPA final acute and chronic values were higher than this "safe" concentration, whereas the LOEC of a 28-d *D. magna* laboratory test provided an effective prediction of the "safe" concentration.

B.2.6 Crossland and Wolff (1985)

In this study [97] the effects of pentachlorophenol (PCP) were examined in outdoor experimental ponds. PCP was repeatedly applied to the subsurface water of three ponds with the aim of maintaining an average concentration of 50 to 100 µg/L for 30 days. There were also replicate control ponds. Actual pond concentrations of PCP averaged 19 to 21 µg/L (days 1 through 14) and 60 to 69 µg/L (days 15 through 43). No statistically significant effects were observed on algal, zooplankton, benthic macroinvertebrate, or fish populations. It should be noted, however, that there was considerable within and between replicate pond variability. The three lowest laboratory determined PCP LC50 values gleaned from the literature were 52 µg/L (96-h rainbow trout), 100 µg/L (8-d snail egg production and viability), and 130 µg/L (16-day snail egg viability). Since most of these effect concentrations from the most sensitive species in the database were greater than PCP concentrations in the ponds, impacts would not be predicted. Based on these observations, the authors contended that a combination of single species toxicity test results can effectively predict environmentally "safe"

concentrations for a chemical. The variability of the treatment concentrations as well as concentration variation within and between replicate ponds, in addition to the fact that a pond effect concentration was not established renders this study inconclusive regarding the accuracy of single species test results in predicting environmental impacts.

B.2.7 Giddings et al. (1984)

These investigators (Giddings et al., 1984; Franco et al., 1984) examined the impacts of phenolic compounds on biological communities in outdoor ponds. The phenolic compounds were administered to replicate ponds daily for 56 days; five different treatment levels were compared to control ponds. A laboratory-generated 28-d test LOEC for *Daphnia magna* was a relatively good forecaster of a phenol effect concentration in the experimental ponds. However, the most sensitive indices of biological community structure/function were affected at phenol concentrations lower than the laboratory chronic LOEC. A 48-h test LC50 for *Daphnia* was not a good predictor of pond effects because much lower concentrations of the phenolic compounds impacted pond communities.

B.2.8 Nimmo et al. (1989)

Acute (96 h) toxicity tests with fathead minnows, Johnny darters (*Etheostoma nigrum*), white suckers (*Catosotomus commersoni*), as well as acute and 7-d *C. dubia* toxicity tests were conducted by Nimmo et al (1989) to evaluate whether river water (Vrain River in Colorado) ameliorated toxicity of ammonia compared to laboratory tests in which well water was used.

For most of the test species, ammonia LC50s were equivalent in the river water compared to the laboratory well water. These data illustrated that there was not an amelioration of ammonia by river water; that is, the laboratory test results did not overestimate toxicity measured instream. Related to the above observation, laboratory single species toxicity tests with polychlorinated biphenyls overestimated the concentration demonstrated in field studies to decrease diversity in invertebrate populations, that is the field populations were more sensitive (Roberts et al., 1978).

B.2.9 Adams et al. (1983)

The toxicity of a commercial phosphate ester product (PEP) determined in outdoor tanks and in the laboratory were compared. The test organisms were *D. magna* and fathead minnows. Five concentrations of the PEP were tested in the outdoor tanks, without replicates; the five concentrations were tested in a series of tanks with and without sediment. PEP concentrations were analytically monitored and exposure concentration maintained for two months. Laboratory toxicity tests consisted of 30-d fathead minnow and 21-d *D. magna* flow-through tests.

LOECs for fathead survival in the lab, outdoor no sediment tank, and outdoor sediment tank were 410, 826, and 545 µg/L, respectively. The laboratory tests overestimated the toxicity of PEP, but estimates were within an order of magnitude of one another. In the *Daphnia* tests, the reproduction LOECs for the lab, outdoor no sediment tank, and outdoor sediment tank were 100, 136, and 226 µg/L, respectively. The laboratory water flea tests gave a fairly reliable qualitative estimate of the PEP LOEC.

A major limitation of this study was that the outdoor tanks were not ecosystem surrogates; they did not contain other biological communities, but only the test species. Small sample sizes and the lack of replication were among the other factors which compromise the reliability of data generated in this study.

B.3 Metals

B.3.1 Canfield et al. (1994)

This group evaluated the potential impacts of past mining activities on the Clark Fork River (Montana) aquatic ecosystem using a benthic invertebrate community assessment, chemical analyses on sediment, and laboratory whole-sediment toxicity tests with an amphipod, *Hyalella azteca*, a midge, *Chironomus riparius*, a cladoceran, *Daphnia magna*, and larval rainbow trout, *Oncorhynchus mykiss*. The study included six sites in the Clark Fork River watershed, one control/reference station on an uncontaminated tributary and five sites downstream of past mining areas.

Sediment concentration of metals (especially copper) were high at Clark Fork sites 1 through 4. A metals concentration gradient from the most upstream sites to the most downstream site was observed. The control site on the tributary had the lowest sediment metal concentration. The authors cautioned that there were many confounding factors (including a possible sampling bias) influencing the benthic invertebrate data which rendered interpretation difficult. The authors pointed out that many chironomids are tolerant of degraded conditions. Furthermore, the percentage of the Chironomidae community comprised of Tanypodinae (considered to be relatively pollution tolerant) was much higher at sites 1, 2, and 3 (upstream sites) than at the control and downstream sites.

The amphipod tests revealed a gradient of toxicity, being highest at the most upstream site and lowest at the control site. Therefore, the results of the laboratory single species toxicity tests were consistent with sediment metal concentrations and the distribution of chironomids. The investigators concluded that chemical analyses, laboratory toxicity tests, and aquatic community evaluations all provided evidence of metal-induced degradation to benthic populations in the river.

B.3.2 Burton et al. (1987)

The Clark Fork River was also the subject of another study. This group evaluated a battery of aquatic toxicity tests including the 7-d *Ceriodaphnia* test and 12 microbial enzyme activity assays. Results of the laboratory toxicity tests were compared to instream parameters including diatom diversity and density, as well as metal concentrations (in both water column and sediment). Data were collected at 13 sites along the river and one control site. As in the previous study (Canfield et al., 1994), sites were on a downstream gradient below an area with past mining activities. Both *Ceriodaphnia* and microbial tests were conducted in the laboratory with water samples from the river sites.

Ceriodaphnia survival ® = 0.94 and 0.93) and neonate production ® = 0.93 and 0.92) showed statistically significant ($p<0.001$) positive correlations with diatom density and diversity, respectively. Survival ® = -0.92 and -0.94) and neonate production ® = -0.92 and -0.94) were negatively correlated with water column copper and zinc, respectively.

These data suggest the *Ceriodaphnia* toxicity test results were effective predictors of instream metal contamination and of diatom population variations. Laboratory microbial enzyme assays for galactosidase, glucosidase, and protease activities also showed statistically significant negative correlations with diatom populations in the river (i.e., low diatom diversity was associated with high enzyme activity).

B.3.3 Clements and Kiffney (1994)

These researchers attempted to assess impacts of metals from a mining site discharging into the Arkansas River (in Colorado). Three sites were selected: one site upstream of the mining operation discharge and two downstream of the discharge. Whether caution was taken in the selection of these sites to assure similar substrates, as well as other physical and chemical conditions is unclear. The intent was that the upstream site would serve as a reference point. The second site was 6 km downstream and the third site was 45 km downstream of the mining operation input; the third was conceived as a site to represent biological community recovery. Unfortunately, two creeks discharged into the Arkansas River below the input from the mining operation, confounding interpretation of data collected at site 2. Furthermore, site 3 was below a town and no attempt was made to account for toxic inputs from the town or other sources (only metal analyses were performed on water samples).

Water samples collected at each site were screened with the 7-d *Ceriodaphnia* test, neonate production being the endpoint. Benthic macroinvertebrates bioaccumulation of metals was assessed. Benthic invertebrate community structure was surveyed by determining the number of taxa and the number of individuals in each taxa. All assessments were conducted in fall and spring, except toxicity testing at site 2 was performed only with a spring water sample.

In the fall water samples, zinc concentrations were highest at site 1, upstream of the mine input. Water sample toxicity was highest at site 3 and lowest at site 2 (just below input from the mining operation) during the fall. The pattern of toxicity in these fall water samples did not correspond to heavy metal concentrations in the same samples. The causes of toxicity in the fall water samples are unknown because Toxicity Identification Evaluations (TIEs) or organic chemical analyses were not performed.

In the spring, all heavy metal concentrations were lowest in water samples from the upstream "control" site and highest at site 2, immediately below the mining operation input. Cadmium, copper, and zinc concentrations at site 2 were 5, 8, and 5.5 fold higher, respectively, than at the reference site. The *Ceriodaphnia* test was conducted only with a water sample from site 2 and the results mirrored the intense metal contamination.

Neither the number of invertebrate taxa nor the number of individuals within taxa showed the site nearest to the mining operation input to be the most impacted. Only the bioaccumulation data and changes in the composition of dominant macroinvertebrate groups suggested that site 2 to be the most impaired compared to the other two sites.

Interpretation of data collected in this study is difficult for several reasons. It is not clear how carefully the three sites were matched in terms of substrate and other physical/chemical factors. In the spring when the other data were more understandable toxicity testing was incomplete and organic chemical analyses were not performed. Stream flow and rainfall conditions were not included in the manuscript, and these factors could influence the measurements and interpretation of results. The authors counsel that the different approaches used in their study provided divergent information regarding metal impacts, so they recommend an integrated approach to assessing impacts on streams. While an integrated approach to assessing impacts on aquatic ecosystems should be supported, design of this study was not optimal and, thus, the results were inconclusive.

B.3.4 Niederlehner et al. (1985)

Several researchers have counseled that single species toxicity tests lack many important interactive characteristics of multivariate, complex ecosystems and, therefore, may not be accurate predictors of biological community responses. Niederlehner et al. (1985)stated that "A multispecies or microcosm test incorporate some of the emergent properties of communities of ecosystems and

serve as an intermediate between the simplicity of the single species toxicity tests and the unreproducible complexity of the environment."

These researchers scrutinized the responses of protozoan communities to cadmium exposures. Effects of cadmium were evaluated by observing colonization of the protozoans in polyurethane foam (PF) islands for 28 d. Exposures were in duplicate tubs using five different cadmium concentrations.

From the experiments, NOECs for protozoan colonization impairment ranged from 0.8 to 9.5 µg Cd/L. In the ambient water quality criteria document for cadmium (USEPA, 1984) chronic values (ChV) adjusted for hardness range from 0.14 (*Daphnia magna*) to 15.04 µg/L (fathead minnow) (selecting values from studies with hardness equivalent to the range seen in the microcosm study). Cladoceran chronic values range from 3.9 µg Cd/L for *Ceriodaphnia reticulata* to 0.14 µg Cd/L for *D. magna* Overall, the data in the Niederlehner et al (1985) report suggest that the laboratory single species toxicity test results underestimate field effects.

It is not evident that the microcosm results were a better predictor of a safe cadmium concentration since the chronic values from 15 of the 16 species listed in USEPA's criterion document were *within* the range of NOECs noted in the mesocosm study. Arguably, this conclusion that the protozoan microcosm "tests were comparable to traditional single species tests in time and expense required, but had the advantages of utilizing indigenous organisms and including processes characteristic of communities, but not single species." is highly questionable. The microcosm tests were 28 d exposures.

B.3.5 Moore and Winner (1989)

These investigators conducted a study in outdoor ponds in Ohio to ascertain the effects of various concentrations of copper on zooplankton and benthic macroinvertebrates. Laboratory 7-d *Ceriodaphnia* toxicity tests were conducted to evaluate the ability to predict effect levels of copper on pond invertebrate communities.

The results of the *Ceriodaphnia* tests predicted the effects of copper on pond populations of *Daphnia ambigua*, but underestimated the impacts of copper on other important species, such as rotifers, copepods, mayfly juveniles, and chironomids.

B.3.6 Geckler et al. (1976)

The results of laboratory chronic toxicity tests in which fathead minnows, green sunfish, and longear sunfish were exposed, in separate tests (i.e., not a multiple species test), to various concentrations of copper were compared to responses of fish in a natural stream. Effects were seen at a somewhat lower copper concentration in the stream than predicted by the laboratory toxicity tests. The authors concluded that, "Agreement between the predictions from laboratory toxicity tests and the observed field effects is surprisingly close considering the measurement errors involved." Similarly, laboratory toxicity test results provided reasonable estimates of the metal concentrations which impacted crustaceans inhabiting tundra ponds (Havas and Hutchinson, 1982).

B.3.7 Giesy et al. (1979)

Giesy et al. (1979) studied the effects of different cadmium concentrations in outdoor experimental stream channels. The results show that single species toxicity test results do not predict secondary effects (cf., Crossland above) in aquatic ecosystems. The primary direct effect of cadmium in these channels was on crayfish; however the direct effect of cadmium on crayfish could have been measured in the lab. In this mesocosm study, the crayfish was a "keystone" species. The decrease in crayfish population greatly influenced community structure including macrophytes, insects, and clams. These secondary effects were not predicted by laboratory toxicity tests, therefore, underestimating biological community impacts.

B.3.8 Marshall (1978)

Marshall (1978) compared the short-term (7 to 9 d) toxicity of cadmium to laboratory and natural populations of *Daphnia galeata* in Lake Michigan. As well as controls, there were four different exposure concentrations for both the laboratory and field populations. Results of this investigation indicated that the characteristics of Lake Michigan water did not appreciably alter the responses of *Daphnia* to cadmium. Furthermore, responses to cadmium were equivalent in the laboratory and in the lake experiments.

This is in contrast to the study by Sherman et al. (1987) which demonstrated that laboratory toxicity test results with cadmium on fathead minnows could not be used to extrapolate to field situations unless hardness and pH in the laboratory tests are equivalent. In general, laboratory toxicity test results underestimated field effects of cadmium.

B.4 Miscellaneous

B.4.1 Boelter et al. (1992)

Ambient water samples from streams receiving discharges of coproduced brine (water that is extracted along with petroleum products from underground deposits) from an oil field in Wyoming were collected and tested for toxicity (Boelter et al., 1992). The 7-d *Ceriodaphnia* test was one of the testing procedures.

Exposure to water samples collected downstream, but not upstream, of the oil field discharges significantly reduced *Ceriodaphnia* survival and neonate production. Application of TIE procedures to toxic samples signified that toxicity could not be attributed to nonpolar organics, heavy metals, or hydrogen sulfide. TIE results along with analytical

chemistry data established that the cause of toxicity was sodium, potassium, bicarbonate, and carbonate ions. Concentrations of these ions were sufficiently high to be toxic to many aquatic organisms.

This study is one of many studies which illustrate that the *Ceriodaphnia* test in combination with TIE and analytical chemistry procedures have effectively identified causes and sources of toxicity in surface waters, storm waters, and effluents.

B.4.2 Gonzalez and Frost (1994)

The responses of two rotifer species, *Keratella cochlea* and *K. taurocephala*, to low pH were compared in laboratory toxicity tests and in a natural lake (Little Rock Lake in Wisconsin). This lake, formed by seepage, consists of two basins which were separated by a vinyl curtain. One of the basins was acidified over time, whereas the other was not modified. Populations of the two rotifer species in the two basins were compared through time. Short term (30-d to 96-h) laboratory toxicity tests were conducted with each of the species using water samples from the two basins of the lake.

The authors concluded that the laboratory tests were not predictive of results obtained in the lake component of the study and recommended caution when extending results from laboratory studies to natural ecosystems. More specifically, the authors suggested that laboratory tests did not explain the population increase of *K. taurocephala* in the acidified basin. However, the laboratory tests did reveal that *K. cochlea* is very sensitive to low pHs, whereas *K. taurocephala* was much less sensitive. *K. cochlea* essentially disappeared from the acidified basin while the population of *K. taurocephala* in that basin increased. Thus, field observations were not necessarily at odds with the laboratory toxicity tests. Furthermore, the population of *K. taurocephala* in the reference basin remained very low throughout study. The *K. taurocephala* population increase in the acidified basin appeared to be due to a reduction of predators, this reduction being caused by the low pH. The laboratory tests with the rotifers would not predict effects on predators.

Other aspects of this study complicate interpretation of the data and acceptance of the author's conclusions. These factors include the absence of replication in the field component of the study and differences in the two basins. The acidified basin underwent thermal stratification and becomes anoxic where as the reference basin did not.

B.4.3. Other Studies

Other investigators (Hitchock, 1965; Eisle and Hartung, 1976; Weiss, 1976; Cairns et al., 1982; Crossland and Hillaby, 1985) have examined the correspondence of laboratory indicator species toxicity test results and biological community responses; the comparisons generally support a good qualitative adequacy of the single species test results as predictors of instream responses.

Some studies (Carlson et al., 1986; Nimmo et al., 1990) were not specifically designed for examining the reliability of the single species test results in predicting aquatic ecosystem responses, but provided qualitative indications of a good correspondence.

Appendix C
Single Species Tests with Ocean Water or Sediment

C.1 Swartz et al. (1994)

Sediment toxicity, as assessed with the amphipod, *Eohaustorius estuarius*, sediment chemical analyses, and the abundance of benthic amphipods were examined along a gradient in the Lauritzen Channel and adjacent areas of Richmond Harbor, California. Dieldrin and DDT were formulated at a facility on Lauritzen Channel from 1945 to 1966.

Objectives included: 1) Examination of the relationship between sediment contamination by DDT and dieldrin, sediment toxicity to *Eohaustorius*, and the field abundance of amphipods at nine sites in the Lauritzen Channel/Richmond Harbor area; 2) Identification of the lowest DDT and dieldrin concentrations associated with effects on amphipod survival in laboratory toxicity tests and effects on abundance of amphipods in the field; and 3) evaluation of the relative contributions of DDT, dieldrin, PAHs, PCBs, and metals to sediment toxicity, and on amphipod abundance in the study area.

Sediment contamination by both dieldrin and the sum of DDT and its metabolites was positively correlated with sediment toxicity and negatively correlated with the abundance of amphipods in the study area; DDT (plus its metabolites) was the dominant toxicological factor. These researchers concluded, "Correlations between toxicity, contamination, and biology indicate that sediment toxicity to *Eohaustorius estuarius*, *Rhepoxynius abronius*, or *Hyallella azteca* in laboratory tests provide reliable evidence of biologically adverse sediment contamination in the field."

In five other studies (Swartz et al., 1985, 1986, 1991; Ferraro et al., 1991; Hake et al., 1994;) statistically significant positive correlations were found between the sum of DDT plus its metabolites in sediment and mortality of amphipods in laboratory sediment toxicity tests. Statistically significant negative correlations were seen between sediment toxicity and amphipod abundance in field sediment samples (i.e., high sediment toxicity related to low abundance). Thus, the weight-of-evidence from these studies suggests that significant toxicity in laboratory sediment toxicity tests provides a reliable qualitative prediction of benthic biological community responses.

C.2 Chapman et al. (1987)

These reseachers conducted an investigation in the San Francisco Bay area which involved measurements of sediment contamination by: chemical analyses; toxicity through sediment toxicity tests (mortality of the amphipod, *Rhepoxynius abronius*, larval development of the mussel, *Mytilis edulis*, behavior of a clam, *Macoma balthica*, and reproduction of the copepod, Tigriopus californicus; and benthic infaunal community structure through taxonomic analyses of macroinfauna).

Sediment samples were collected at three stations at each of three sites in the San Francisco Bay: Islais Waterway, Oakland, and San Pablo Bay. Chemical analyses indicated that the Islais Waterway site was more contaminated by a number of potentially toxic substances than the Oakland site, while the latter site was more contaminated than the San Pablo Bay site.

Benthic community analyses, as well as toxicity test results (especially the mussel larvae, amphipod, and clam behavior tests) suggested that the rank of pollution-induced degradation was: Islais Waterway > Oakland > San Pablo Bay. Moreover, there was concordance among the three synoptic measurements. The authors argue that all three types of assessment are critical for assessing pollution-induced degradation of aquatic biological communities.

C.3 Swartz et al. (1985)

Sediment toxicity, chemical contamination and macrobenthic community structure were examined at seven stations along a gradient northward from Los Angeles County Sanitation Districts' sewage outfalls on the Palos Verdes Shelf and compared to control conditions in Santa Monica Bay. Sediment toxicity was assessed with laboratory toxicity tests utilizing the amphipod, *Rhepoxynius abronius.*

Significant reductions in macrobenthic species richness, density, biomass, and infaunal indices occurred at the three stations which also showed significant toxicity in the laboratory tests. There was a close inverse relationship between sediment toxicity and benthic community measurements. The authors concluded that sediment toxicity tests can be useful in predicting benthic community impacts, but cautioned that the amphipod test is not particu-

larly sensitive. Moreover, absence of statistically significant toxicity in this test should not be interpreted as evidence of a healthy benthic community (i.e., the test yields many false negatives).

C.4 Long and Chapman (1985)

To assess biological community effects of sediment contamination Long and Chapman advocate the use of a Sediment Quality Triad (chemical, toxicity, and benthic infaunal data). The authors contend that too much emphasis is placed on the determination of distribution and concentration of chemicals in the designation of problem areas or "hotspots." They further assert that chemical data alone provide little or no information regarding the possible biological significance of such chemical accumulations. The objective of this publication was to determine the correspondence among measures of the three components of the Triad; data from several studies on Puget Sound, Washington were used.

Toxicity data were derived from six different laboratory sediment tests (amphipod lethality, oligochaete respiration, oyster larval abnormality, fish cell effects, and polychaete life-cycle effects). Data from these tests were combined into a toxicity summary index. Four indices were concluded to be effective indicators of benthic community health. All four indices represent percent contribution of specific taxonomic groups to the total benthic community--contribution of echinoderms (pollution sensitive, so high percentage represents healthy community); contribution of arthropods (many are pollution sensitive, so higher percentage represents healthy community); contribution of phoxocephalid amphipods (pollution sensitive, so higher percentage represents healthy community); and contribution of polychaetes and molluscs (many are relatively pollution tolerant, so high percentage can represent impacted community).

Using the above indicators of benthic community health, the toxicity tests summary index was a reliable predictor of biological community impacts. In fact, good overall correspondence among the three components of the Triad was observed. On a station-by-station basis, the chemical data alone were not always reliable indicators of biological effects.

C.5 Becker et al. (1990)

Laboratory sediment toxicity tests and benthic macroinvertebrate assemblage surveys were conducted at 43 stations in Commencement Bay, Washington; there were four reference sites in Carr Inlet. The toxicity tests included the amphipod (*Rhepoxynius abronius*) mortality test, the oyster (*Crassostrea gigas*) larval development test, and the Microtox™ test. A numerical classification analysis was applied to the benthic assemblages data. Sediment samples were also subjected to chemical analyses for organic compounds and metals.

Toxicity test results and benthic assemblages alterations were inversely related, whereas toxicity was positively correlated with chemical concentrations. This suggests that most biological effects resulted from chemical toxicity. That is, the laboratory toxicity tests were reliable qualitative predictors of biological community responses.

To evaluate the correspondence between toxicity test results and alterations of benthic assemblages, three types of comparisons were made. Concordance was first determined; this is a measure of agreement between results of toxicity tests and macroinvertebrate surveys (i.e., both show statistically significant effects or both show no statistically significant effects). Statistical significance of concordance was evaluated using a binomial test and an expected level of concordance of 0.5 (i.e., that for random agreement). Of the 47 stations the benthic assemblages at 19 were deemed altered. Concordance was 60% (not significant) with the amphipod test, 81% ($p<0.001$) with the oyster larval test, and 68% ($p<0.01$) with the Microtox™ test. Sensitivity of the toxicity tests was represented as the percentage of stations with altered benthic assemblages that also revealed statistically significant toxicity. Sensitivity was 84%, 68%, and 42%, respectively for the Microtox™, oyster larval, and amphipod, tests. Efficiency of the toxicity tests was determined as the percentage of tests which identified only those stations with altered benthic assemblages. Efficiency was 81%, 57%, and 50% for the oyster larval, Microtox™, and amphipod tests, respectively. The authors concluded that the laboratory sediment tests, especially the oyster larval tests, were reasonable predictors of altered benthic assemblages.

C.6 Swartz et al. (1982)

The toxicity of 175 sediment samples from Commencement Bay was measured in the laboratory *Rhepoxynia abronis* survival test. The relationship between these toxicity test results and benthic community data from these sites was explored. Benthic community data exhibited a negative correlation (decreased amphipod density and species richness with higher levels of toxicity) with laboratory sediment toxicity. The authors concluded that the correlation between laboratory and field results indicated that the sediment toxicity tests were reliable predictors of biological community responses.

C.7 Schimmel et al. (1989a,b)

Studies were conducted to assess the relationship between effluent and ocean water toxicity. The estimates of chronic toxicity were made from 1982 to 1984 at seven locations along the Atlantic and Gulf Coasts with effluent and ocean water samples (USEPA, 1994b).

Effluent dilutions at various locations in receiving waters were estimated with dye studies so that effect concentrations could be compared. Data presented by these investigators reveal that effluent toxicity reliably reflected receiving water toxicity (effect concentrations in effluent and ocean water samples with equivalent dilution corresponded). The results of these studies signify that the ocean receiving waters had little affect on the toxicity/bioavailability of chemicals in the effluent. The "missing link" in these investigations was establishing a connection between marine water toxicity and biological community responses.

C.8 Frithsen et al. (1989)

Using indicator species toxicity tests and an ecosystem survey, a four month study was conducted to evaluate the toxicity of a sewage effluent. Effluent discharge was into Narragansett Bay. Effluent toxicity was evaluated with the sea urchin, *Arbacia punctulata* sperm cell test. Ecological effects of the effluent were assessed in mesocosms considered by the authors to be functional analogs of shallow, unstratified coastal systems such as Narragansett Bay.

The sewage effluent consistently tested toxic in the sea urchin test, with the average EC50 being 1.1% effluent. Little information could be gleaned from the mesocosm data due to several problems. There were unexplained effects on phytoplankton and organic carbon loading which lead to hypoxia. Toxicity measured in the mesocosm did not correlate with that in the sewage effluent. There was incomplete mixing of effluent in the mesocosms. Significant toxicity was detected in the control mesocosms. Toxicity in all mesocosms was highly variable and not related to effluent toxicity. Because of the confounding factors, a conclusion that the mesocosm data failed to confirm laboratory effluent toxicity data would be inappropriate. The study was inconclusive.

Appendix D
Strengths and Limitations of Single Species Toxicity Tests

D.1 Strengths of Single Species Tests

There is no instrument that can measure or predict how organisms will respond to a toxic chemical(s). Furthermore, chemical analyses of effluent or ambient water samples do not yield information on toxicological additivity, bioavailability, synergistic, or cumulative effects. Many wastewater and ambient waters are complex, containing constituents that interact and that differ in toxicity; therefore, single chemical standards are important, but of limited value in protecting water quality.

1) Single species tests integrate additivity and cumulative interactions of chemicals.
2) Single species tests provide a direct measure of chemical bioavailability.
3) Single species tests measure responses to toxicants for which there are no chemical-chemical-specific water quality standards.
4) Single species tests have provided reliable estimates of concentrations (for many different types of chemicals) which cause effects in aquatic ecosystems.
5) Because they are highly standardized with specific quality assurance and control requirements, single species tests provide reliable, repeatable, and comparable results with good precision compared to other types of chemical and biological tests.
6) Single species tests provide an early warning signal so that actions can be taken to minimize significant ecosystem impacts (especially with regard to the discharge or release of toxic chemicals).
7) Single species tests can be performed relatively rapidly and inexpensively. This allows for the accumulation of a data set which better characterizes the wastewater or ambient water system.

D.2 Limitations of Single Species Tests

Several definite and potential limitations of single species toxicity tests have been identified.

1) Results of a single test do not characterize the duration, or frequency of toxicity in wastewater or ambient waters. Instream exposure reflects ambient water/effluent characteristics over time (days, weeks), whereas exposure in the laboratory reflects the characteristics of ambient water or wastewater in a grab sample or composite sample of one day.
2) Results of a test or tests with an effluent do not allow for assessment of cumulative effects of toxic substances from different sources in aquatic ecosystems.
3) The range of sensitivities (to toxic substances) of organisms and functions in aquatic ecosystems may not be encompassed by single species tests.
4) Effects due to bioaccumulation/bioconcentration, delayed, or secondary effects are not measured.
5) Results of single species tests may underestimate ecosystem community responses because of the multiple stressors acting on natural populations and communities. Single species tests include limited range of endpoints (responses) compared to aquatic ecosystems.
6) Results of single species tests may not be predictive of trophic interactions and ecosystem operational processes (tests do not incorporate aquatic ecosystem complexity).
7) Physical and chemical, as well as biotic factors, in aquatic ecosystems could modify (increase or decrease) bioavailability or toxicity compared to laboratory tests. The highly controlled exposure regimes in the laboratory may not reflect the multivariant and complex exposure conditions in natural settings.
8) Single species tests tend to use non-indigenous species that may not represent local biota.
9) Single species toxicity test results fail to account for indirect effects of contaminants.
10) Single species tests tend to use genetically homogenous laboratory populations.

☆U.S. GOVERNMENT PRINTING OFFICE: 1999 - 550-101/20006