EVALUATION OF METHODS FOR MEASURING SEDIMENT TOXICITY IN CALIFORNIA BAYS AND ESTUARIES

Technical Report 503
March 2007

Steven Bay
Darrin Greenstein
Diana Young

Southern California Coastal Water Research Project

Evaluation of Methods for Measuring Sediment Toxicity in California Bays and Estuaries

Steven Bay, Darrin Greenstein and Diana Young

Southern California Coastal Water Research Project
3535 Harbor Blvd., Suite 110
Costa Mesa, CA 92626

www.sccwrp.org

March 2007

Technical Report 503

FOREWARD

The State Water Resources Control Board (SWRCB) initiated a program in 2003 to develop sediment quality objectives (SQOs) for chemical contaminants in California bays and estuaries. The SQOs will include narrative descriptions of the condition to be protected and the associated analytical methods needed to determine whether the condition has been attained. The Southern California Coastal Water Research Project, in partnership with other state and federal agencies, conducted a series of technical studies in order to provide a sound scientific foundation for the selection of methods and the development of a data interpretation framework for use in the SQO program. This report presents the results of an evaluation of sediment toxicity test methods for use in the assessment of the direct effects of sediment contamination. Other reports will describe studies related to the assessment of benthic macrofaunal community condition, sediment contamination, assessment of indirect effects from consumption of contaminated seafood by humans and wildlife, and the integration of all of these data to assess overall sediment quality. Copies of this and related reports are available for download at www.sccwrp.org and www.waterboards.ca.gov.

This study was funded in part by agreement 01-274-250-0 with the State Water Resources Control Board.

ACKNOWLEDGMENTS

The authors would like to thank Art Barnett and Stephen Weisberg for their assistance in the preparation of this document. We would like to thank J. Daniel Farrar, David Moore, Bryn Phillips, and Michele Redmond for providing data that was used in the evaluation and calculation of thresholds.

Many thanks are due to the members of the Scientific Steering Committee (Todd Bridges, Rob Burgess, Tom Gries, Peter Landrum, Ed Long, and Bob Van Dolah) for their advice during all phases of this project. Finally, the authors wish to thank members of the Sediment Quality Advisory Committee and Agency Coordination Committee for their input during this study.

EXECUTIVE SUMMARY

Toxicity tests have been widely used to assess sediment quality in a variety of research, monitoring, and regulatory programs. While many programs use a combination of test methods and follow standardized protocols, there is variation between programs in the selection of test methods or in the way that the data are interpreted. This is problematic when incorporating sediment toxicity into a regulatory program with broad applicability, such as the sediment quality objectives program under development in California. Multiple factors such as test feasibility, relevance to program/policy objectives, data comparability, cost, and sensitivity must be considered, yet this information is frequently not available. In addition, a consistent method of toxicity data interpretation is needed so that station assessments conducted in one region are comparable to the results from other locations or times.

The current study had two objectives: to evaluate a variety of acute and sublethal toxicity tests in order to identify methods that were best suited for use in a statewide regulatory program, and to develop a system to classify the toxicity test results into a series of categories of effect. A list of candidate test methods was developed based on a literature review and consultation with other scientists. The candidate test methods list included acute test methods with four amphipod species. Six sublethal methods were also evaluated: a copepod life cycle test, amphipod growth, polychaete growth, clam growth, oyster cell stress, and mussel or sea urchin embryo development.

Data on the feasibility, sensitivity, variability, and cost of each candidate method were compiled from the literature and from two sets of laboratory experiments. The first set of experiments compared the relative sensitivity of each of the candidate test methods for detecting toxicity in a set of 15 sediment samples from various California embayments. A wide range of responsiveness to the samples was observed. The copepod life cycle and polychaete growth tests showed the greatest responses to the sediment samples. Some of the sublethal tests identified a smaller total number of stations as toxic than the standard amphipod survival test(s), yet each of the sublethal tests detected toxicity in some samples that were classified as nontoxic by the amphipod survival test. This suggests that sublethal tests and acute tests are complementary rather than redundant and can provide different sensitivity responses.

Experiments were also conducted to evaluate the interlaboratory variability of the clam growth and embryo development test methods when applied to both field and laboratory-spiked sediments. The interlaboratory variability of these tests was greater than reported for some amphipod survival tests, but was within the range of variability for other sublethal test methods.

The data were compiled into a matrix of test characteristics and scored based on relative performance of each test. The acute and sublethal methods were evaluated separately.

The following five tests were identified as best suited for use in a California statewide sediment quality assessment program.

Species	Taxonomic Group	Matrix	Duration (days)	Endpoint(s)
Acute				
Eohaustorius estuarius *Leptocheirus plumulosus* *Rhepoxynius abronius*	Amphipod	Whole sediment	10	Survival
Sublethal				
Neanthes arenaceodentata	Polychaete	Whole sediment	28	Growth
Mytilus galloprovincialis	Bivalve	Sediment-water interface	2	Embryo development

The use of multiple toxicity tests to assess sediment quality is suggested, as none of the test methods ranked consistently highest with respect to sensitivity or reliability. The use of a diversity of test methods provides two key advantages: it reduces the influence of spurious results from a test and it also increases the overall sensitivity of the testing program by using species with different patterns of contaminant sensitivity.

A data analysis framework was developed for the highest rated test methods. This framework was based on an ordinal scoring system consisting of four categories of effect.

- **Nontoxic:** Response not substantially different from that expected in sediments that are uncontaminated and have optimum characteristics for the test species
- **Low toxicity**: A response that is of relatively low magnitude; the response may not be greater than test variability
- **Moderate toxicity**: High confidence that a statistically significant effect is present
- **High toxicity**: Highest confidence that a toxic effect is present and the magnitude of response is among the strongest effects observed for the test

Three response thresholds (low, moderate, and high) were developed for use in assigning one of the above response categories to each test result.

Species	Low (%)	Moderate (% Control)	High (% Control)
Eohaustorius estuarius	90	82	59
Rhepoxynius abronius	90	83	70
Leptocheirus plumulosus	90	78	56
Neanthes arenaceodentata	90[1]	68	46
Mytilus galloprovincialis	80	77	42

[1] % of control growth.

Several data limitations were encountered in the course of this study that either reduced the ability of a test method to meet the minimum evaluation criteria or complicated the calculation of the classification thresholds. Research is needed to improve the feasibility

of some of the candidate test methods. Additional data are also needed to refine the thresholds for the *Leptocheirus plumulosus* and *Neanthes arenaceodentata* tests.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Toxicity tests are an integral part of the sediment quality triad used in many monitoring and assessment programs (Long and Chapman 1985). These tests provide information on the potential for adverse biological effects from contaminants and are recognized as a key component of the ecological risk assessment process (USEPA 1998) and programs to evaluate the suitability of dredged material for ocean disposal (USEPA 1991, PSWQA 1995). Sediment toxicity tests have also been widely used in monitoring and assessment programs to evaluate sediment quality within coastal bays and estuaries (Fairey *et al.* 1998) and at regional and national scales (Long 2000, USEPA 2004).

A wide variety of methods have been used to measure sediment toxicity (Lamberson *et al.* 1992). Many studies use a suite of tests that includes both acute (short-term survival) and sublethal methods. Much of the acute testing has employed amphipod survival methods using standard protocols established by the U.S. Environmental Protection Agency (USEPA 1994). The use of these standard protocols provides a measure of biological effects that can be compared among regions statewide and nationwide; such comparisons are not always possible using other measures of biological effects. The types of sublethal toxicity tests used in assessment studies is more variable, with methods including growth and reproduction tests of whole sediment, pore water, water or solvent extracts of the sediment (Ringwood *et al.* 1996, Bay *et al.* 1998, Long *et al.* 1999, Long *et al.* 2005). There is little consistency among programs in the types of the sublethal tests used; selection is often performed on a site-specific basis and is based on factors such as availability of test organisms, expected sensitivity, cost, local interests, and availability of collaborators. Consequently, only a few sublethal methods have been used commonly; they include the amphipod *Leptocheirus plumulosus* 28-day growth and reproduction test (USEPA 2001), a 20-day polychaete growth test using *Neanthes arenaceodentata* (PSWQA 1995), pore water or elutriate tests using echinoderm or bivalve gametes or embryos (PSWQA 1995, ASTM 2002a, Carr and Nipper 2003), and a sediment-water interface (SWI) test using sea urchin or mussel embryos (Anderson *et al.* 1996).

Information on the comparative sensitivity of sediment toxicity tests is an important factor to consider in test selection, yet only limited data are available. Most comparative studies include just a few species and sometimes provide conflicting results (Table 1). The differences in species, test methods, sample type, and relative sensitivity of the test methods complicate the integration of the results of these studies for use in selecting methods for use in other studies. Additional comparative studies that use a consistent study design applied to each test are needed to help evaluate the relative sensitivity of the toxicity tests of interest.

The selection of sediment toxicity test methods requires a consideration of many factors in addition to sensitivity, depending upon the study's objectives and design. Much variability in method selection is found among research studies conducted on a small scale, as the emphasis is often on selecting methods to address site-specific scientific questions, method development, or building upon previous work by an investigator. Additional factors must be considered when selecting test methods for use in large-scale monitoring or regulatory programs. For example, the methods must be feasible for use by many different laboratories and at different times of the year, and have a wide tolerance of habitat variables such as sediment grain size and salinity.

Toxicity test method selection for these types of programs must consider factors such as test feasibility, relevance to program/policy objectives, data comparability, and cost, in addition to sensitivity. The sediment quality objectives (SQO) program under development by the State of California provides an example of the many factors to be considered when sediment toxicity tests are used in a regulatory context. The California SQO program is based on the sediment quality triad and will be applied to bays and estuaries throughout the state (SWRCB 2006). The selection of toxicity test methods for a statewide regulatory program must be sensitive to environmental contamination at levels that are ecologically relevant, standardized to ensure consistent application, and feasible for application in a variety of situations. The test methods should also be ecologically relevant, meaning that the choice of species and test conditions results in a test that responds to environmental contamination on a scale that is useful for describing potential impacts on California species. In addition, a consistent and relatively simple method of toxicity data interpretation is needed so that station assessments conducted in one region are comparable to the results from other locations or times. Past comparisons of sediment toxicity test methods have not addressed many of these issues or were limited to a small subset of test methods that do not fully address the needs of a statewide regulatory program.

The current study had two principal objectives. The first objective was to evaluate a variety of acute and sublethal toxicity tests in order to identify methods that were best suited for use in a statewide regulatory program. To address this objective, a candidate list of potential tests was identified and evaluated with respect to feasibility, performance, and cost. The second objective was to develop a consistent and comparable system to classify the toxicity test results into a series of categories of effect. The approach to address this second objective included developing a conceptual data analysis framework and identifying a series of test-specific response thresholds that incorporated the magnitude and uncertainty in the test response.

Table 1. Summary of studies comparing the sensitivity of acute survival (A) and sublethal (S) toxicity tests.

Species and Methods	Sample Type	Relative Sensitivity	Reference
Ampelisca abdita (A) *Eohaustorius estuarius* (A) *Leptocheirus plumulosus* (A)	Field Sediment and Cadmium	Sediment: *A. abdita>L. plumulosus>E. estuarius* Cd: *A. abdit*a =*L. plumulosus>E. estuarius*	(Schlekat *et al.* 1995)
A. abdita (A) *Ampelisca verrilli* (A) *Mercenaria mercenaria* (A) *Palaemonetes pugio* (A) *Brachionus plicatilis* (A) *Amphiascus tenuiremis* (A) Microtox (S)	DDT, Fluoranthene, Cadmium	DDT: *P. pugio* most sensitive Fluoranthene and Cd: *M. mercenaria* most sensitive	(Fulton *et al.* 1999)
Polydora cornuta (S) *Boccardia proboscidea* (S) *Neanthes arenaceodentata* (S) *L. plumulosus* (S) *Schizopera knabeni* (S)	Copper	*S. knabeni* most sensitive *L. plumulosus* and *B. proboscidea* least sensitive	(Farrar *et al.* 1998)
L. plumulosus (S) *E. estuarius* (A) *N. arenaceodentata* (S)	Field Sediment	*E. estuarius>N. arenaceodentata>L. plumulosus*	(Pinza *et al.* 2002)
L. plumulosus (S) *A. abdita* (A) *N. arenaceodentata* (S)	Field Sediment	*L. plumulosus>A. abdita>N. arenaceodentata*	(Kennedy *et al.* 2004)
L. plumulosus (A) *L. plumulosus* (S) *N. arenaceodentata* (S)	Field Sediment	*L. plumulosus> N. arenaceodentata*	(Moore *et al.* 2003)
A. abdita (A) *Rhepoxynius abronius* (A) *Mytilus galloprovincialis* (S) *Strongylocentrotus purpuratus* (S) *Dinophilus gyrociliatus* (S)	Field Sediment	*M. galloprovincialis* and *R. abronius* most sensitive *A. abdita* least sensitive	(Long *et al.* 1990)

EVALUATION OF ACUTE AND SUBLETHAL TESTS

Approach

A set of candidate acute and sublethal test methods was selected for evaluation. Methods were selected that had a direct sediment exposure, appeared to be technically feasible and had data available that indicated sensitivity to contaminated sediments. The test methods and species included those that have been recommended for use in other regulatory programs in California (USEPA and Engineers 1998) or were documented in standard procedures developed by government or scientific agencies (e.g., EPA or ASTM). Priority was given to methods using species resident in California and species representative of important infaunal groups. In order to increase the diversity of life histories and biological endpoints evaluated, additional candidate methods were selected based on a review of the scientific literature and from recommendations by other scientists familiar with sediment toxicity testing. This process led to the identification of six candidate sublethal methods for evaluation (Table 2). Four amphipod species recommended by the USEPA for testing acute sediment toxicity were also included in the list (USEPA 2001).

Each test was evaluated based on a set of characteristics relating to test feasibility, performance and cost. The list of characteristics was established to include parameters used in previous test comparisons (Long *et al.* 1990, Lamberson *et al.* 1992) and was refined using input from an external scientific review committee. The following characteristics were evaluated:

- **Organism availability**. This category relates to both abundance of suppliers of the animals and any seasonal aspect of either their availability or sensitivity. Ideally, test organisms should be available from multiple suppliers on a year-round basis with no seasonal variation in test sensitivity. Information for this parameter came from contacting suppliers or from experience in using the organisms.

- **Method description.** This category describes whether a standardized protocol for a given test has been established. Methods that are termed as "standard" have a protocol that has received the rigorous testing necessary to be published as an EPA or ASTM method and is the preferred level of method description. These methods have control acceptability criteria and quality assurance standards for parameters such as water quality associated with them.

- **Technical difficulty**. An important consideration is the ease for laboratories to successfully conduct the test. If a method is difficult to perform, laboratories may have to perform multiple tests just to obtain acceptable results. The difficulty was rated based on ability to obtain acceptable controls (i.e., relative number of test failures), the necessity of special techniques or equipment, and complexity of the exposure system. The information for this parameter was based on a combination of personal experience of the authors and comments from others who routinely perform the tests.

- **Concordance of results**. For evaluating the degree of concordance, the effects on the sublethal methods were compared to those of the acute methods tested simultaneously. For the sublethal methods, there was an expectation that if a site were strongly, acutely

toxic to a test organism, then an effect would also be seen for the sublethal test. Conversely, if a site were considered to be in "reference condition" then there would be an expectation that no toxicity would be found for any of the test methods. The information for this parameter was taken from published reports in which both an amphipod species and at least one of the sublethal tests had been applied on the same samples. To evaluate concordance, the acute amphipod test was used as the ground truth, so no acute amphipod data appear in Table 3.

- **Relative sensitivity**. This category describes the relative response of the acute and sublethal tests by observing the relative frequency that the test identifies a sample as being toxic, compared to a benchmark test. Sensitivity in the context of this study refers to the range in response obtained using a specific test method, not the inherent sensitivity of a species to individual chemicals. Many factors related to the specifics of the test, such as duration, temperature, and life stage can affect the response and apparent sensitivity of a toxicity test. Test sensitivity was evaluated relative to the acute amphipod test species most commonly used in California, *Eohaustorius estuarius*. This species has a substantial history of use in California for both monitoring and assessment studies. The logic behind this assessment was that if a test method was usually less sensitive than the most commonly used test, then its value in providing additional information would be limited. Information for this characteristic was gathered from published reports where the benchmark test was conducted alongside at least one of the sublethal methods. For many of the methods, no data was available, so a study was conducted to help fill this information gap (Appendix A).

- **Reproducibility among laboratories**. This category describes the relative amount of variability in the results that is observed when multiple laboratories test the same sample. The information was mostly obtained from literature reports on round-robin tests. In the case of the *Mercenaria mercenaria* growth test and the SWI test using mussel embryos, round-robin testing was conducted to add information that was missing from the literature (Appendix B).

- **Reproducibility within laboratories.** This category describes the relative amount of variability in the results when an individual lab tests the same sample multiple times. The information was obtained mostly from reference toxicant exposures.

- **Precision.** The relative precision of response describes the between-replicate variability of the methods. Information for this parameter was obtained from published reports and journal articles.

- **Documentation of confounding factors**. Most toxicity tests are sensitive to some type of non-contaminant effect (e.g., grain size) that can have a confounding effect on test results. Knowledge of which factors can affect a test and the range where effects occur is needed for study design and data interpretation. Information for this parameter was gathered from test protocols or from values published in the literature.

- **Cost.** Cost is a limiting factor in many sediment assessment studies. The use of sensitive tests that are also relatively inexpensive will enable a larger number of stations to be evaluated, thus improving spatial resolution and overall confidence in the results. The unit cost of each test was evaluated relative to the standard 10-day amphipod survival test. The first source of information for this parameter was from the costs associated with the tests that were commissioned as part of this study. Secondarily, biological consulting firms in California provided costs for tests that they currently perform. For the tests that were new to California, the firms were asked to estimate what they would charge to conduct them.

The characteristics were summarized into narrative categories that reflected the relative level of attainment for each of the candidate tests (e.g., poor, fair, good). The acute and sublethal test methods were treated separately during this process due to differences in the characteristics evaluated.

A scoring system was then applied to integrate the category level information in order to produce an overall evaluation and ranking of each test. Test selection was based on consideration of both test feasibility and relative performance/cost. The three feasibility characteristics (organism availability, method description, and technical difficulty) were evaluated using a binary (yes/no) scoring system. These characteristics were deemed to be so important that the test was classified as not feasible if minimum criteria were not met. For organism availability, at least one commercial source of animals must currently be available to purchase animals ready to use for testing. For method description, there must be a published document available that has a complete description of the method, including test acceptability criteria. The technical difficulty criterion was that there was a reasonable expectation that a laboratory experienced in performing other toxicity tests could follow the protocol and successfully conduct the method without receiving additional outside training. For each of these characteristics, the method was assigned a "+" if the criterion was met and a "-" if it was not.

The remaining performance and cost characteristics were evaluated using a weighted scoring system based on the narrative categories. A weighting factor was established for each category based on our assessment of the relative importance of each category. The comparative sensitivity category was assigned the highest weight: a factor of 4. The high weight given to this category was based on the assumption that high sensitivity to contaminants was the most desirable trait for a sediment toxicity test method. The "relative precision of response" category was deemed to be the least important and was assigned a weighting factor of 1. All of the remaining categories were considered to be of intermediate importance and were assigned a weighting factor of 2.

A numeric value was assigned for each of the performance and cost characteristics. The values for each category ranged from 0 to 3 and corresponded to the narrative categories assigned based on the data review. A value of zero was assigned when no data were available for a characteristic. Each individual value was multiplied by its respective weighting factor to produce a score for the characteristic. The scores were then summed to obtain final score for each candidate test method.

Table 2. List of candidate sediment toxicity tests, the citations containing testing protocols and whether quality assurance and test acceptability criteria have been established.

Species	Taxonomic Group	Duration (days)	Matrix	Endpoint(s)	Literature Level	Citations	QA Criteria[1]	State/National Program Use[2]
Ampelisca abdita *Eohaustorius estuarius* *Rhepoxynius abronius* *Leptocheirus plumulosus*	Amphipod	10	Whole sediment	Survival	Well established	(USEPA 1994, ASTM 1996)	Yes	EMAP NOAA USACE WA, RMP
L. plumulosus	Amphipod	28	Whole sediment	Growth, reproduction, survival	Well established	(USEPA 2001)	Yes	USACE
Neanthes arenaceodentata	Polychaete	28	Whole sediment	Growth, survival	Exposure method under revision	(ASTM 2002b) modified	Yes	USACE[3] WA
Strongylocentrotus purpuratus	Sea urchin	3	Sediment-water interface	Embryo development	Published	(Anderson *et al.* 1996)	Yes	
Mytilus galloprovincialis	Mussel	2	Sediment-water interface	Embryo development	Published	(Anderson *et al.* 1996)	Yes	RMP
Amphiascus tenuiremis	Copepod	14	Whole sediment	Reproduction, survival	Published	(Chandler and Green 1996)	No	NOAA
Mercenaria mercenaria	Clam	7	Whole sediment	Growth, survival	Journal	(Ringwood and Keppler 1998, Keppler and Ringwood 2002)	No	EMAP
Crassostrea virginica	Oyster	4	Whole sediment	Lysosomal stability	Exposure method not published	(Ringwood *et al.* 1998, Ringwood *et al.* 2003)	No	

[1]Information on acceptable water quality ranges, reference toxicants, guidelines, acceptable control parameters, and within test variability are available

[2]EMAP: Environmental Monitoring and Assessment Program; NOAA: NOAA National Status and Trends Program; USACE (U.S. Army Corps of Engineers: dredged material evaluation for disposal under USACE or USEPA guidance; WA: dredged material evaluation for disposal under Washington State guidance; RMP: San Francisco Bay Regional Monitoring Program

[3]The same species and endpoint is used in dredged material evaluations, but the duration and aspects of the test method differ

Results

Acute Test Method Evaluation

The four acute amphipod test species were similar in regards to the test feasibility characteristics of organism availability, method description, and technical difficulty (Table 3). Each of the species is available from commercial suppliers, test methods have been standardized, and the level of difficulty is generally low. All of the amphipod species were scored as having met the feasibility criteria (Table 4).

E. estuarius received the highest overall score for the performance and cost characteristics (Table 4). *E. estuarius* has an extensive history of use in toxicity testing studies on California sediments (Anderson *et al.* 1997, Bay *et al.* 2000, Bay and Brown 2003, Bay *et al.* 2005). The method has been shown to have good reproducibility between laboratories (Bay *et al.* 2003).

A slightly lower total score was obtained for *L. plumulosus* (Table 4), which was due to lower reproducibility within and among laboratories. *L. plumulosus* received a lower rating compared to *E. estuarius* and *Rhepoxynius abronius* regarding documentation of confounding factors due to a lack of information on sensitivity to hydrogen sulfide, which was available for *E. estuarius* and *R. abronius.* The high ranking for relative sensitivity compared to *E. estuarius* was based on limited data from a single study and may not represent overall trends. The *L. plumulosus* 10-day test has been conducted in California on a very limited basis. However, it has long been used in other parts of the country, especially on the Gulf coast for monitoring and assessment studies. In studies using diluted, contaminated field sediments or spiked sediments, it has been shown that *L. plumulosus* has a sensitivity similar to the other species (Schlekat *et al.* 1995, Boese *et al.* 1997, DeWitt *et al.* 1997). One of the most attractive attributes of *L. plumulosus* is that it is easily cultured in the laboratory and available year round from commercial suppliers who have them in culture.

The *R. abronius* 10-day test was ranked similarly to the other acute methods, except for a low score for relative sensitivity compared to *E. estuarius*. The relative sensitivity score was based on limited data for split samples from a single study and may not represent overall trends. *R. abronius* has been previously used in California sediment toxicity programs (Long *et al.* 1990, Anderson *et al.* 1998, Anderson *et al.* 2001). These studies found the *R. abronius* method to have equal or better sensitivity to contaminated sediments as compared to other methods tested simultaneously. An interlaboratory comparison exercise using this method found good agreement amongst the testing laboratories (Mearns *et al.* 1986). However, test organism availability has recently been a problem with *R. abronius*. Laboratories have had recent difficulty in locating a supplier of *R. abronius*. The only available source of animals is in Washington, which requires an export permit prior to receipt of the animals. These factors may interfere with the ability to conduct this amphipod test in a timely manner. Sediments with a silt-clay content of ≥80% have also been shown to be an adverse confounding factor for *R. abronius* (DeWitt *et al.* 1988). Care should be taken when planning a survey that sediment grain size will not be an issue and that an animal source is readily available.

The Ampelisca abdita 10-day test was assigned the lowest total score among the four acute test species. The low score was driven by a lack of sensitivity compared to *E. estuarius* and a lower reproducibility among laboratories (Table 3). Specifically, in tests of California sediments where *A. abdita* has been tested simultaneously with *E. estuarius* or *R. abronius* it has consistently been found to be less sensitive (Figures 1 and 2). Very few data are available to make direct comparisons between *E. estuarius* and *R. abronius*, but toxicity in southern California sediments has been detected at a similar frequency using either of these species. The lower apparent sensitivity of *A. abdita* may be due to the fact that this species does not burrow in sediment, but lives in a tube-like structure and does not ingest sediment.

A. abdita also received a lower rating regarding documentation of confounding factors due to a lack of information on sensitivity to hydrogen sulfide. In addition, it is difficult to obtain *A. abdita* during the winter months and if they are available, they are of a size that is smaller than desired for use in testing (Table 3). The *A. abdita* test was also rated as being more difficult to conduct than other 10-day amphipod survival tests, based on the experiences of several California laboratories in having a higher test failure rate when using *A. abdita*, compared other amphipod species (Table 3). These difficulties are not due to intrinsic problems with the test organism, but are likely due to problems in obtaining *A. abdita* from suppliers within California. *A. abdita* is widely used as an indicator of sediment toxicity in many monitoring programs and the data have been used to characterize sediment quality on a national scale (Long 2000, USEPA 2004). Laboratories outside of California have had a high rate of success in conducting tests with *A. abdita* and technical difficulties reported in California do not preclude the use of the test in other regions.

Table 3. Characteristics of candidate sediment toxicity test methods. Not applicable for test (NA).

	California Samples Tested	Organism Availability[1]	Method Description[2]	Technical Difficulty[3]	Concordance at Clearly Clean or Impacted Sites[4]	More Sensitive Than *Eohaustorius estuarius* (number of comparisons)[5]	Reproducibility Among Laboratories[6]	Reproducibility Within Lababoratories[6]	Relative Precision of Response[7]	Documentation of Confounding Factors[8]	Cost of Method[9]
Amphipod Acute											
Eohaustorius estuarius	1697	12 (+)	Standard	Low	NA	NA	Good	Good	NA	Good	Low
R. abronius	1026	12 (1)	Standard	Low	NA	Never (9)	Good	Good	NA	Good	Low
Leptocheirus plumulosus	15	12 (+)	Standard	Low	NA	Often (15)	Fair	Poor	NA	Fair	Low
A. abdita	710	8 (+)	Standard	Moderate	NA	Rarely (228)	Poor	Good	NA	Fair	Low
Sublethal Methods											
Mercenaria mercenaria	15	8(+)	Published	Low	Fair	Sometimes (15)	Fair	Fair	Similar	Good	Low
Neanthes arenaceodentata	15	12(1)	Published	Moderate	Fair	Sometimes (15)	Good	Good	Low	Good	High
Sediment-water Interface											
Mytilus galloprovincialis	117	12(++)	Published	Low	Fair	Rarely (117)	Fair	Good	Low	Fair	Low
Strongylocentrotus purpuratus	195	5(++)	Published	Low	Fair	Rarely (184)	None	Good	Low	Good	Low
L. plumulosus	15	12(+)	Standard	Moderate	Fair	Sometimes (15)	Fair	Good	Low	Good	High
A. tenuiremus	10	12(1)	Published	High	Good	Often (10)	None	Good	High	Fair	Very High
C. virginica	15	8(++)	Report	Moderate	Poor	Sometimes (15)	None	None	Low	Poor	Moderate

[1]Number of months (relative number of available suppliers, + +for many, + for few, 1 for one)
[2]Standard=Established method by government agency; Published = Peer reviewed publication of method; Report = In gray literature
[3]Low = Similar skills and equipment needed as for acute amphipod test; Moderate = More difficult to obtain acceptable controls, special techniques or more complex exposure system; High=Combination of special skills and more complex exposure system needed
[4]Concordance with acute amphipod test: Good = >75%; Fair = <75% and >50%; Poor <50%
[5]Of the stations found to be toxic by at least one endpoint: Often = >50% of stations; Sometimes = <50% and >20%, Rarely <20%; Never = 0%
[6] Good = CV <50%; Fair = CV >50% and <75%; Poor = CV>75% (CV = coefficient of variation; mean/standard deviation x 100)
[7]Categories based on the range of median acute amphipod standard deviations. High = below range; Similar = within range; Low = above range
[8]Data available for confounding factors: Good=Four or more factors; Fair= 2 or 3 factors; Poor= Less than 2 factors
[9]Low=150% or less the cost of acute amphipod; Moderate = 150% to 200% of amphipod; High = 200% to 300% of amphipod; Very High = >300% of amphipod.

Table 4. Numerically based rating matrix of acute and sublethal sediment toxicity methods. Final score is sum of ratings.

	Feasibility					Performance and Cost							
	Organisms Availability	Method Description	Technical Difficulty	Overall Feasibility		Concordance with Amphipods at Clearly Clean or Impacted Sites	More Sensitive Than Acute *Eohaustorius estuarius* Test	Reproducibility Among Laboratories	Reproducibility Within Laboratories	Relative Precision of Response	Documentation of Confounding Factors	Relative per Station Cost	Total Score
					Factor	2	4	2	2	1	2	2	
Amphipod Acute													
Eohaustorius estuarius	+	+	+	Yes		NA	8	6	6	2	6	6	34
Rhepoxynius abronius	+	+	+	Yes		NA	0	6	6	2	6	6	26
Leptocheirus plumulosus	+	+	+	Yes		NA	12	4	2	2	4	6	30
Ampelisca abdita	+	+	+	Yes		NA	4	2	6	2	4	6	24
Sublethal Methods													
Mercenaria mercenaria growth	+	-	+	No		4	8	4	4	2	6	6	34
Neanthes arenaceodentata survival and growth	+	+	+	Yes		4	8	6	6	1	6	2	33
Sediment-water Interface													
Mytilus galloprovincialis	+	+	+	Yes		4	4	4	6	1	4	6	29
Strongylocentrotus purpuratus	+	+	+	Yes		4	4	0	6	1	6	6	27
L. plumulosus –28-day	+	+	+	Yes		4	8	4	6	1	6	2	31
Amphiascus tenuiremus Life Cycle	-	+	-	No		6	12	0	6	3	4	0	31
Crassostrea virginica lysosomal stability	+	-	-	No		2	8	0	0	1	2	4	17

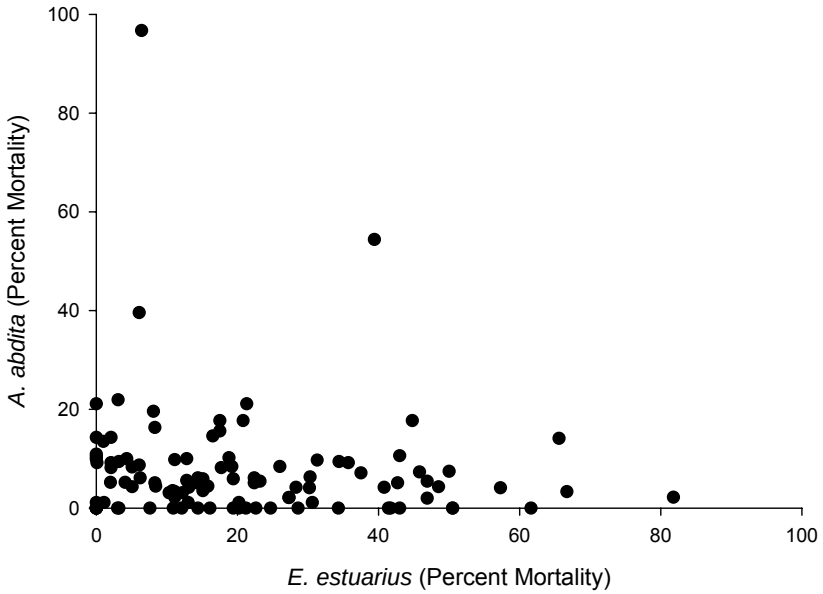

Figure 1. Comparison of mortality data between *Ampelisca abdita* and *Eohaustorius estuarius* on split samples. Data were obtained from multiple regional assessment studies in California.

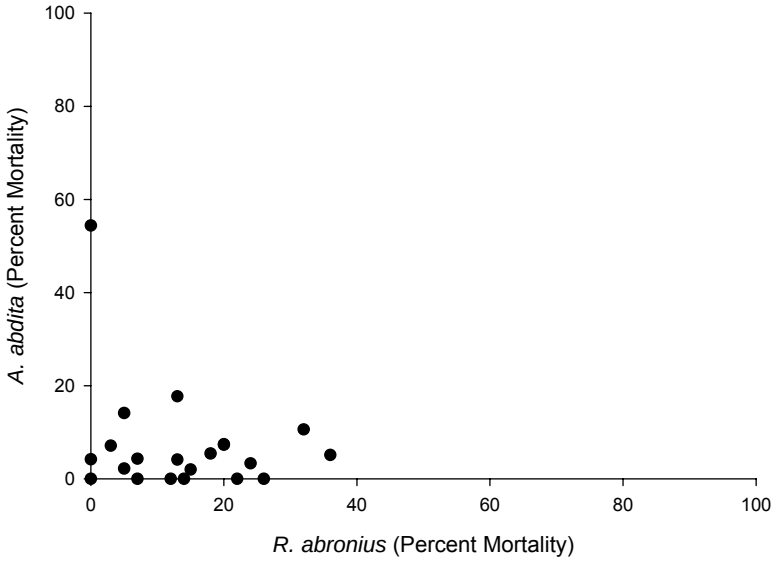

Figure 2. Comparison of mortality data between *Ampelisca abdita* and *Rhepoxynius abronius* on split samples. Data were obtained from multiple regional assessment studies in California.

Sublethal Test Method Evaluation

The candidate sublethal tests were more variable in regards to feasibility, performance, and cost than the acute methods (Table 3). Three of the sublethal test methods had substantial limitations in regard to method documentation, organism availability, or technical difficulty that resulted in an overall rating of not feasible for use in a statewide assessment program at this time (Table 4). These methods were the bivalve *M. mercenaria* growth test, the copepod *Amphiascus tenuiremus* life cycle test, and the lysosomal destabilization test using the oyster *Crassostrea virginica*.

The *M. mercenaria* growth test received the highest total performance and cost score of any of the methods, based on average to slightly above ratings in all of the categories (Tables 2 and 3). However, there is not a single, cohesive document that completely details the protocol and there are no published test acceptability criteria. The test is economical (Table 5) and is not technically difficult to perform. The method exhibited fair reproducibility between laboratories in a round-robin study (Appendix B). In a previous study in the EMAP Carolinian Province, the clam test found no toxicity in reference areas, but did well at identifying areas that were clearly degraded as being toxic; it did better in both these regards than did the *A. abdita* 10-day test (Hyland *et al.* 1998). However, in testing on California sediments the clam test proved to be less sensitive than the *E. estuarius* 10-day test and was one of the least sensitive tests overall (Appendix A). In the Carolinian Province it was found that the *M. mercenaria* test was the best of the toxicity tests conducted at predicting expected bioeffects (Van Dolah *et al.* 1999).

The life cycle test with the copepod *Amphiascus* was by far the most sensitive of the sublethal methods compared to amphipod acute tests (Table 4, Appendix A). This method was also shown to very sensitive compared to an amphipod acute test in a previous study in Florida (Long *et al.* 1999). Nevertheless, the *Amphiascus* test did not pass two of the feasibility criteria. There is no established commercial supplier of the test animals. Only one laboratory in the country maintains a culture of the animals that can be used by other laboratories to start their own cultures. The necessity to culture the animals in individual laboratories leads to the second feasibility limitation, which is technical difficulty. In order to conduct this toxicity test, a laboratory must maintain a copepod culture, and cultures of three algal species used to feed the copepods. In addition, the protocol requires specialized exposure containers and a finely-controlled seawater flow through system. The *Amphiascus* life cycle test is also approximately three times more expensive than other tests (Table 5) and has received no interlaboratory testing to document reproducibility.

Table 5. Per sample cost of performing sediment toxicity tests. Prices are based on quotes from a minimum of three laboratories.

Test	Low Quote ($)	High Quote ($)
Amphipod Acute		
Ampelisca abdita	600	800
Eohaustorius estuarius	600	800
Leptocheirus plumulosus	600	800
Rhepoxynius abronius	600	800
L. plumulosus 28-day Growth and Reproduction	1,375	1,800
Neanthes arenaceodentata 28-day Growth	800	1,400
Sediment-water Interface Embryo Development		
Strongylocentrotus purpuratus	550	1,100
Mytilus galloprovincialis	700	1,200
Mercenaria mercenaria Survival and Growth	600	750
Crassostrea virginica Lysosomal Stability	400	1,500
Amphiascus tenuiremus Life Cycle	2,200	2,800

The oyster lysosomal destabilization test had the lowest total score of any of the test methods (Table 4). Besides the low ranking, this test method does not have a complete protocol that is published (Table 2). In preliminary tests of the procedure, we also found the endpoint determination to be very difficult to discern without significant training from someone very experienced in the procedure, leading to the acceptability failure for technical difficulty. Further, this method has had very limited testing with individual chemicals and until this project, had not been used in field studies along side other test methods. In the testing conducted to date, the oyster lysosomal destabilization test has not been demonstrated to be particularly sensitive compared to acute amphipod tests (Appendix A).

The remaining three sublethal test methods, the *N. arenaceodentata* 28-day growth test, the SWI test with either *M. galloprovincialis* or *Strongylocentrotus purpuratus*, and the *L. plumulosus* 28-day growth and reproduction test, met all of the feasibility criteria. The ranking scores of these tests covered a fairly narrow range of 27 to 33 (Table 4).

The *N. arenaceodentata* growth test received the highest ranking of the remaining sublethal tests (Table 4). It is fairly well established with an ASTM method, although the method documentation is currently under revision to reflect some changes in the procedure. It has been used in multiple field studies and individual chemical exposures to spiked sediments (Dillon *et al.* 1993, Green *et al.* 1999, Lotufo *et al.* 2000, Lotufo *et al.* 2001b, Moore *et al.* 2003, Kennedy *et al.* 2004). The *N. arenaceodentata* 28-day test has also been the subject of considerable refinement efforts considering animal age, test duration and food ration (Bridges and Farrar 1997, Bridges *et al.* 1997). For the methods comparison study using California sediments, the *N. arenaceodentata* test was the second most sensitive test (Appendix A). In that study, the *N. arenaceodentata* test either agreed with the *E. estuarius* test or identified stations as toxic that the *E. estuarius* did not; there were no stations that were found to be toxic by *E. estuarius,* but not *N. arenaceodentata*. When compared to the *L. plumulosus* 10-day results, the *N. arenaceodentata* test was about equal in its ability to detect toxicity, and was second only to the copepod test in sensitivity. While the *N. arenaceodentata* test is one of the more expensive to conduct (Table 5) it has relatively high sensitivity, reliability, and technical feasibility.

The SWI test using mussel embryos received a lower total score than the *N. arenaceodentata* test (Table 4). The SWI test using either developing sea urchin or mussel embryos is an established test method that has been used by multiple laboratories to assess California sediments. The exposure protocol for this procedure is published in a well respected compendium of toxicity test methods (Table 2) and the embryo testing methods are based on standard EPA procedures (USEPA 1995). The protocol has previously been successfully employed in multiple studies within California (Hunt *et al.* 2001, Bay *et al.* 2004, Brown and Bay 2005). The cost of conducting the test is relatively low and the mussels are available in spawning condition year round from multiple suppliers. The test protocol also addresses an important pathway of toxicant effects: exposure of water column organisms to chemicals released from contaminated sediments. The relative sensitivity of this protocol compared to amphipod acute tests is uncertain since the results of side-by-side testing have been mixed. The SWI tests were classified as having relatively low precision (Table 3). This low score reflects increased variability among replicates due to the SWI test design, where the replicates often represent discrete sediment core samples as opposed to replicates of a homogenized sample. Between replicate precision of mussel or sea urchin embryo tests in water only tests is much higher than the SWI results.

The SWI test has been used in the past with both sea urchin and mussel embryos; however, review of the data available for the sea urchin method led to a lower score than for the mussels (Table 4). This low score was due to low interlaboratory reproducibility. Greater technical difficulty is associated with conducting the SWI test with sea urchins. One issue is that sea urchins have a short spawning season in the field and it is cumbersome to extend the spawning season by maintaining the animals in the laboratory. Second, laboratories have reported greater difficulty in recovering the sea urchin embryos at the end of the exposure period. This may be due to a more delicate structure of the sea urchin embryos, which may cause them to stick to the exposure chamber. The reduced embryo recovery success may produce higher between replicate variability for the sea urchins, which may account for the lower sensitivity and reproducibility scores. Compared to sea urchins, *M. galloprovincialis* embryos provide advantages of being available year round in spawning condition and having an endpoint that is easier to measure with precision.

The *L. plumulosus* 28-day test received a relatively high total score that was only two points below the *N. arenaceodentata* test method. This test is both well established and documented (USEPA 2001). The method has been used in multiple field studies and individual chemical exposures to spiked sediments (DeWitt *et al.* 1997, McGee *et al.* 1999, Lotufo *et al.* 2001a, McGee *et al.* 2004). The *L. plumulosus* 28-day test was the third most sensitive of the sublethal methods tested using California sediments (Appendix A). However, there were several California stations where the acute amphipod tests detected toxicity and *L. plumulosus* 28-day did not. Also during this testing, the *L. plumulosus* 28-test experienced a test failure and there were questions regarding the reliability of the reproduction data (Appendix A). Inconsistent reliability of the *L. plumulosus* 28-day test reproductive endpoint has also been reported in another study (Kennedy *et al.* 2004). In a study of sediments in Chesapeake Bay, it was found that the 28-day test did not provide more information regarding toxicity than the 10-day test with the same species and that the 10-day test data had a better correlation with changes in the benthic

community (McGee *et al.* 2004). The *L. plumulosus* 28-day test is the second most expensive test to perform (Table 5).

Discussion

The evaluation of the candidate acute and sublethal tests identified five methods that had the best overall combination of technical feasibility and relatively high performance. These methods include three acute amphipod and two sublethal test methods (Table 6). Each of these methods is well suited for use in a California statewide sediment quality assessment program where feasibility, sensitivity, reliability, and cost are all important factors.

Table 6. Sediment toxicity test methods with the highest overall ranking with respect to the evaluation characteristics.

Species	Taxonomic Group	Matrix	Duration (days)	Endpoint(s)
Acute				
Eohaustorius estuarius *Leptocheirus plumulosus* *Rhepoxynius abronius*	Amphipod	Whole sediment	10	Survival
Sublethal				
Neanthes arenaceodentata	Polychaete	Whole sediment	28	Growth, survival
Mytilus galloprovincialis	Bivalve	Sediment-water interface	2	Embryo development

The two sublethal tests in Table 6 provide important features not present in the suite of amphipod acute tests that are most commonly used to assess sediment quality. The use of a polychaete worm in the *N. arenaceodentata* test provides greater taxonomic diversity among the test organisms and is representative of one the most abundant taxonomic groups comprising the benthic community. The SWI test also represents a different taxon that is also a dominant member of most benthic macrofaunal communities, and the use of an early life-stage may provide enhanced sensitivity to different contaminants. The incorporation a SWI exposure in the *M. galloprovincialis* test also provides a means to evaluate the significance of sediment contaminant impacts on organisms residing in the water column, and thus increases the chance that the testing program will detect toxicity that is present under a diversity of conditions.

Only one of several sediment toxicity methods using the polychaete *N. arenaceodentata* was evaluated in this study. The two methods that are the most established are a 20-day growth test used in the Pacific Northwest (PSWQA 1995), California and many other regions for dredged material characterization, and a 28-day test (ASTM 2002b) that has been optimized to achieve a more sensitive growth endpoint (Bridges *et al.* 1997). The 20-day method has been successfully used in the state of Washington for over 15 years. However, some researchers have found it to be less sensitive than amphipod survival tests (Anderson *et al.* 1998, Pinza *et al.* 2002). In side-by-side testing, one study found the 28-day test to be more sensitive than the 20-day method (Gardiner and Niewolny 1998). Based on the results of these studies, it was decided to focus the evaluation on the 28-day method.

The *L. plumulosus* 28-day test is also a feasible test that had a relatively high total score and could be used in a statewide assessment program. This method was judged to have lower overall suitability because the test is fairly costly to perform, provides no increase in taxonomic diversity, and an uncertain increase in sensitivity relative to the acute amphipod methods already in widespread use.

This study was restricted to toxicity tests where whole sediment samples were included in the exposure. Tests on sediment pore water or elutriate samples were not considered for evaluation because of technical limitations in the methods and a greater uncertainty in the relationship between the test exposure and sediment contaminant concentrations. Pore water tests are a widely used method for testing sediment toxicity (Carr and Nipper 2003), but it is often difficult to collect enough sample for testing. There are other issues associated with pore water toxicity tests that make these methods problematic for use as an initial test of sediment toxicity, including potential changes in metal toxicity due to oxidation, change in sample pH, sorption of contaminants to test chambers, confounding effects of ammonia toxicity, and elimination of sediment ingestion as a route of uptake (Chapman *et al.* 2002, Ho *et al.* 2002). While many of these issues may also be associated with whole sediment tests, they are magnified with the use of pore water.

While elutriate tests are used in several assessment programs, the relationship of the results to direct sediment exposure is not clear. Elutriate tests were developed for testing the effects of the resususpension of the dredged sediment on water column toxicity, not the toxicity of bedded sediment. The proportions of sediment and water and the method of agitation used to prepare the elutriate are operationally defined and the relationship of the resulting exposure experienced by a test organism to that from a whole sediment exposure is unknown. The State of Washington uses a modified elutriate toxicity test method that includes the whole sediment after mixing with the water and tests bivalve or echinoderm larvae (PSWQA 1995, ASTM 2002a). These methods have been used successfully in Washington for over a decade. The Puget Sound method was not included in the present study because of concerns that the organism's response to the whole sediment in the test chamber would be confounded by the presence of the elutriate.

The *A. abdita*, *M. mercenaria,* and *A. tenuiremus*, tests showed good potential as tests that might be feasible for statewide application in the future. For now, more work needs to be performed on issues regarding animal availability, method development, relative sensitivity, and interlaboratory variation to make these protocols viable choices. Although the oyster lysosome test scored poorly in our ratings, the endpoint represents an important indicator of cellular stress that is responsive to toxicant exposure. The applicability of this method to assess sediment toxicity would be improved through the use of an organism with a greater direct exposure to the sediment, such as a crustacean, polychaete or deposit-feeding bivalve.

The use of multiple toxicity tests is needed to provide a complete and confident evaluation of sediment toxicity. None of the methods identified in Table 6 has been shown to be consistently the most sensitive or reliable test. This situation is to be expected, since there are species-specific variations in contaminant sensitivity and mode of exposure among the test organisms, and many different combinations of chemical type and magnitude may produce sediment toxicity. The use of multiple tests provides two key advantages. First, this approach provides a more reliable assessment of toxicity by reducing the chance that a spurious result in any one test will determine the toxicity classification. The influence of potentially confounding factors such as sediment grain size and organic carbon content are still not entirely known for many tests. Confidence in the results is increased when the results of multiple toxicity tests are similar. Second, the use of multiple test methods increases the sensitivity of the testing program by using

a variety of species, response endpoints, and exposure methods. This combination reduces the chance of a false negative (failure to detect sediment toxicity) due to species-specific variations in contaminant sensitivity or mode of exposure. Multiple toxicity tests were used in NOAA's National Status and Trends Program (Long *et al.* 1996) and are currently used in Washington's Puget Sound Ambient Monitoring Program (Long *et al.* 2005).

TOXICITY RESPONSE THRESHOLDS

Approach

An ordinal scoring system consisting of four categories of response was developed for each of the toxicity tests listed in Table 6. The use of multiple categories, as opposed to a simple binary approach (nontoxic/toxic) retains more information about the toxicity response and thus provides greater potential resolution when combining the toxicity data with other lines of evidence in a sediment quality triad approach. Each category was based on a narrative description of condition that incorporated both the degree of confidence that a toxic effect was present and the magnitude of mean response to the sample.

- **Nontoxic:** Response not substantially different from that expected in sediments that are uncontaminated and have optimum characteristics for the test species
- **Low Toxicity**: A response that is of relatively low magnitude; the response may not be greater than test variability
- **Moderate Toxicity**: High confidence that a statistically significant effect is present
- **High Toxicity**: Highest confidence that a toxic effect is present and the magnitude of response is among the strongest effects observed for the test

This four-category system is an adaptation of the three-category system that is often used to classify sediment toxicity (Long *et al.* 2000), where the test response is classified as nontoxic, marginal, or toxic. The nontoxic and marginal categories correspond to the nontoxic and low toxicity categories of the scoring system used here. The toxic category used in many studies usually represents a reliably statistically significant response that encompasses a wide range of effect (e.g., 20 to 100% mortality) and thus provides little discrimination among the majority of the toxic samples. Two categories of response, moderate and high, were established to represent these toxic samples in order to provide the ability to distinguish severe effects from more moderate responses.

A conceptual approach was developed to relate each of the above categories to a series of numeric thresholds and statistical criteria (Figure 3). This approach relies on the comparison of the test result (e.g., % survival) to Low, Moderate, and High thresholds, corresponding to the upper bound of the response range for the Low Toxicity, Moderate Toxicity, and High Toxicity categories. The thresholds were developed using test-specific characteristics, such as test variability (minimum significant difference (MSD)) and distribution of the toxicity response data. A statistical criterion was also used in the classification scheme (Figure 3). Samples qualifying for the Low or Moderate categories based on test response magnitude were classified into the next lower category if the response was not significantly difference relative to the control (t test, $p \leq 0.05$). A statistical significance criterion was not applied to the highest toxicity category because the derivation of the high toxicity threshold already incorporated a high degree of statistical confidence.

The methodology used to derive the numeric thresholds is described in the following sections.

Low Threshold

The threshold separating the Nontoxic and Low categories was defined as the lowest acceptable control response value for the given test, as established in the test protocols. The response value is defined as the mean value for the endpoint for a given test method (i.e., survival, growth). Any test sample having a response value that is greater (e.g., greater survival) than or equal to the low threshold will be classified as nontoxic, regardless of whether a statistical difference from the control is present. A test response that is less (e.g., lower survival) than the low threshold will be classified as low, moderate, or high, depending on the magnitude of response and statistical significance (Figure 3).

This threshold was based on the rationale that any response that fell within the range expected of animals exposed to optimum sediment conditions (i.e., controls) should indicate a nontoxic condition in the test sample. The control acceptability criteria were obtained from the appropriate protocol for each test method.

Moderate Threshold

The intent of the Moderate Threshold is to distinguish between samples producing a small response of uncertain significance and larger responses representing a reliably significant difference relative to the control. This threshold was based on the Minimum Significant Difference (MSD), which was specific to each test method. The MSD represents the minimum difference between the control and sample mean response that is necessary to be statistically different at $p \leq 0.05$ level. The moderate threshold was equal to the 90th percentile of the MSDs for a given toxicity test method. This approach for calculating a toxicity threshold has been used by other researchers (Phillips *et al.* 2001). Use of the 90th percentile results in a threshold with a high degree of confidence that the sample is different from the nontoxic condition.

The MSD values were calculated using a dataset of replicate control and sample data that were compiled from the SQO database and from laboratories outside of California. Details of this calculation can be found in Phillips *et al*. (2001). An MSD was calculated for each combination of a control and a sample using the following equation:

$$MSD = t_{critical} \left(s_1^2/n_1 + s_2^2/n_2\right)^{-½}$$

where $t_{critical}$ = t value from the standard statistical table ($\alpha = 0.05$); s_1^2, s_2^2 = variances for control and field sample; and n_1, n_2 = numbers replicates. All of the MSD values in the dataset for each toxicity test method were then sorted in rank order. The 90th percentile value of this set of data was then calculated (MSD_{90}). The MSD_{90} values were calculated using all available data for each toxicity test method. Finally, the moderate threshold value was calculated by subtracting the MSD_{90} from 100% in order to produce a value that could be compared to the control-adjusted test response value.

Sample response values (i.e., survival or growth) between the low and moderate thresholds are classified as Low Toxicity if they are significantly different from the control response (Figure 3). Sample response values that are less than the moderate threshold and are significantly different from the control are categorized as moderately toxic.

High Threshold

The narrative intent of the High Threshold is to identify samples producing a severe and highly significant effect from those samples producing lesser effects. No precedent for this threshold was available from the literature, so this threshold was based on a combination of test variability and response distribution that corresponded to the category definition.

The 99th percentile MSD value was used to link the high threshold to test variability. A sample having a response that falls below this limit (e.g., lower survival) would be expected to be significantly different from the control 99% of the time. This value therefore represents a response that is associated with a very high level of confidence of statistical significance. The 99th percentile MSD for the high threshold was calculated using the same data and methodology described for the calculation of the MSD_{90} for the moderate threshold.

The response distribution component of the high threshold was based on the distribution of toxic samples from California. For purposes of this calculation, toxic samples were defined as samples having a mean response that was significantly different from the control response. The toxic samples were ranked in descending order based on the control-adjusted mean survival. The response magnitude component of the high threshold corresponded to the 75th percentile of the data. The value obtained from this calculation represents the response associated with the most strongly affected 25% of the toxic samples found in California. It was required that data for this calculation be from stations within California in order to obtain a response value that was relevant to the characteristics of sediments in California.

Both the variability and data distribution response values represented important, but partial, aspects of the High Threshold. Therefore, the mean of the two values was used as the High Threshold. Response values (i.e., survival or growth) below the high threshold are classified as high toxicity regardless of whether they are significantly different from the control response or not (Figure 3).

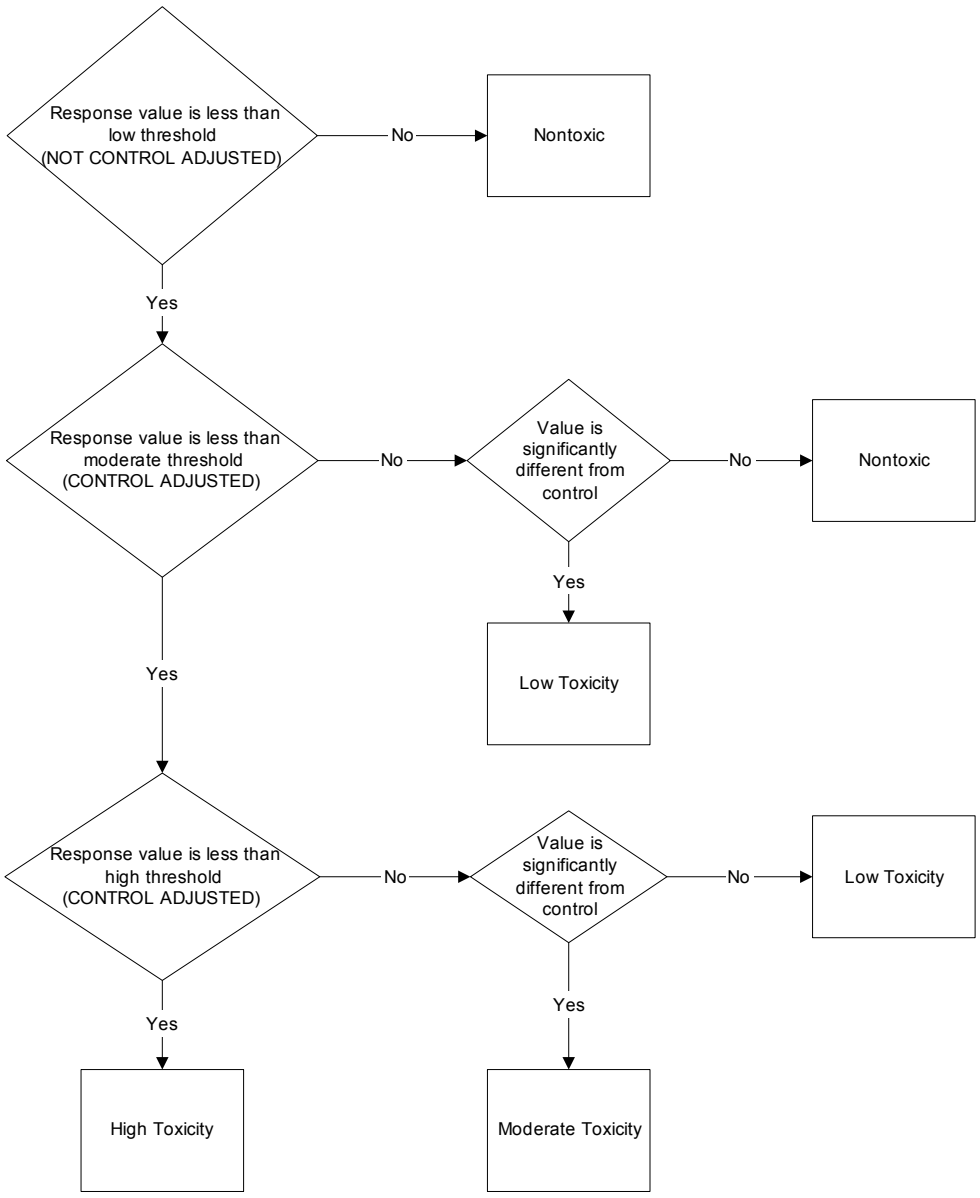

Figure 3. Conceptual approach for assigning the category of toxic effect from exposure response data. The test response value is expressed as survival, embryo development or growth.

Results

Low Threshold

For the amphipod acute survival tests the threshold is 90% survival (USEPA 1994). For the *N. arenaceodentata* growth endpoint, the threshold is 90% of the mean control growth, according to the revised ASTM protocol that is in preparation (J. D. Farrar, personal communication). For the SWI test using *M. galloprovincialis* embryos, the low threshold is 80% normal-alive (not control adjusted). The control criterion for the *M. galloprovincialis* test was established by the Marine Pollution Studies Laboratory, Granite Canyon (B. Phillips, personal communication).

Moderate Threshold

The moderate threshold for the *E. estuarius* 10-day survival test was calculated using data from the California Sediment Quality Objectives database, which included 876 MSD values. The 90th percentile of the MSD values was 18%, which corresponds to a control adjusted survival of 82% (Figure 4).

The *R. abronius* 10-day acute test threshold was also calculated using data from the California database. The dataset included 264 data points (Figure 5). The calculated control adjusted survival threshold for *R. abronius* was 83%, very similar to the *E. estuarius* value.

The threshold for the *L. plumulosus* 10-day survival test was calculated using data from tests on sediment from throughout the U.S. The data were provided by multiple laboratories. Few of the 199 samples in the data set were from stations located in California. The calculated control adjusted survival threshold for the *L. plumulosus* acute test was 78% (Figure 6).

Like the *L. plumulosus* 10-day value, the threshold of the *N. arenaceodentata* growth test was calculated from tests of samples from throughout the United States, with few California stations included. There were less data available for this test method; the calculation was based on 92 data points. The threshold value for the *N. arenaceodentata* growth endpoint was 68% of the mean weight of the control animals (Figure 7).

The threshold for the SWI test with *M. galloprovincialis* embryos was calculated using data from the statewide SQO database. The threshold value of 77% was calculated from 118 MSD values (Figure 8).

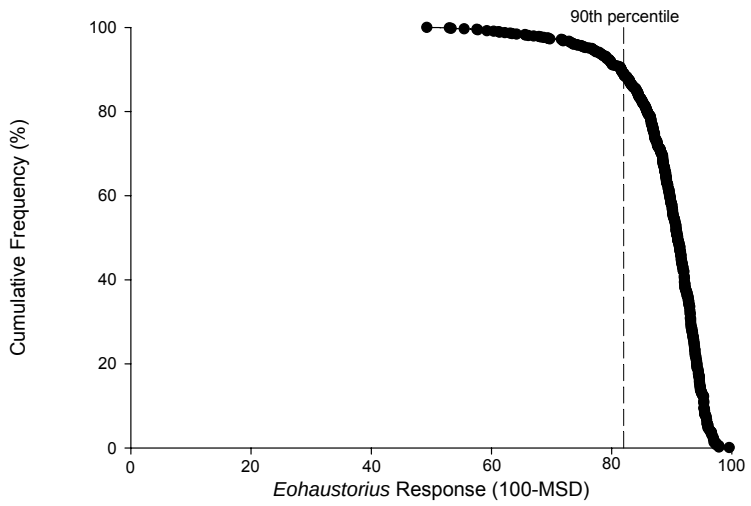

Figure 4. Cumulative frequency of *Eohaustorius estuarius* response (100-MSD) values expressed as a percentage of control survival. The 90th percentile value is the moderate response threshold. Sample size = 876.

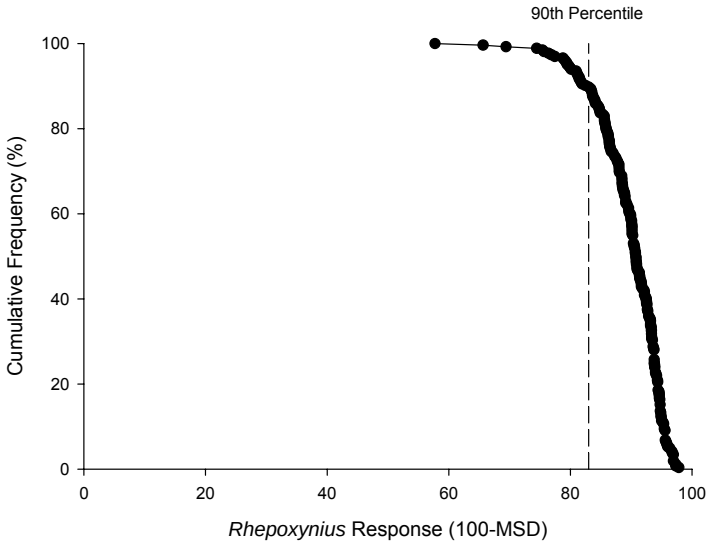

Figure 5. Cumulative frequency of *Rhepoxynius abronius* response (100-MSD) values expressed as a percentage of control survival. The 90th percentile value is the moderate response threshold. Sample size = 264.

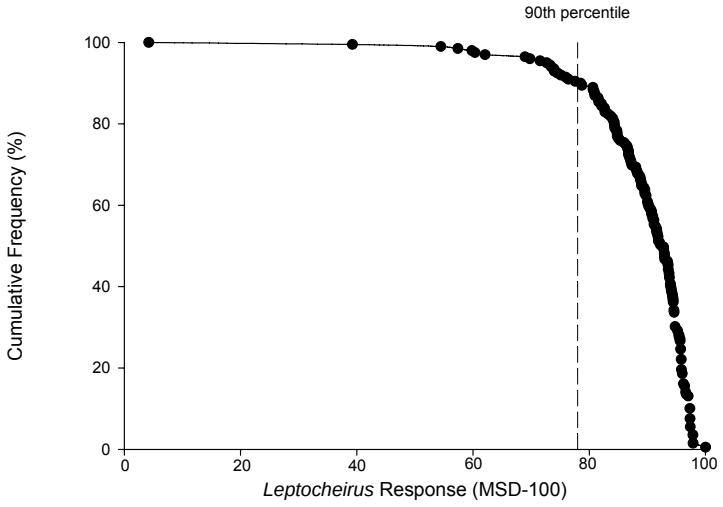

Figure 6. Cumulative frequency of *Leptocheirus plumulosus* response (100-MSD) values expressed as a percentage of control survival. The 90th percentile value is the moderate response threshold. Sample size = 199.

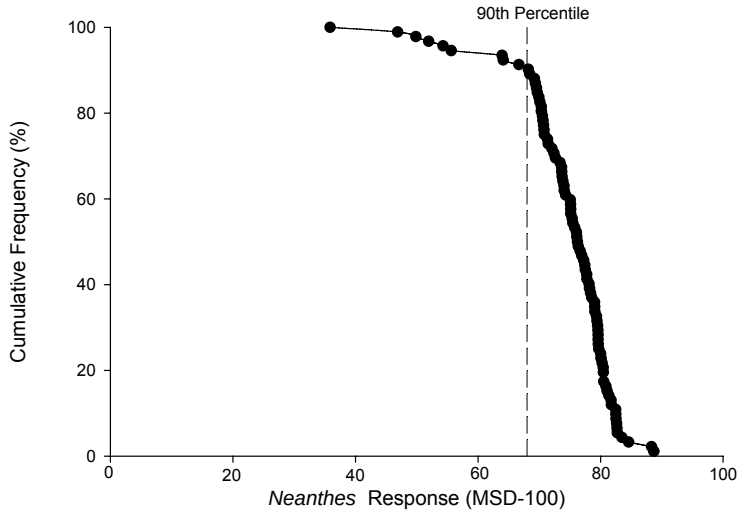

Figure 7. Cumulative frequency of *Neanthes arenaceodentata* growth response (100-MSD) values expressed as a percentage of control growth. The 90th percentile value is the moderate response threshold. Sample size = 92.

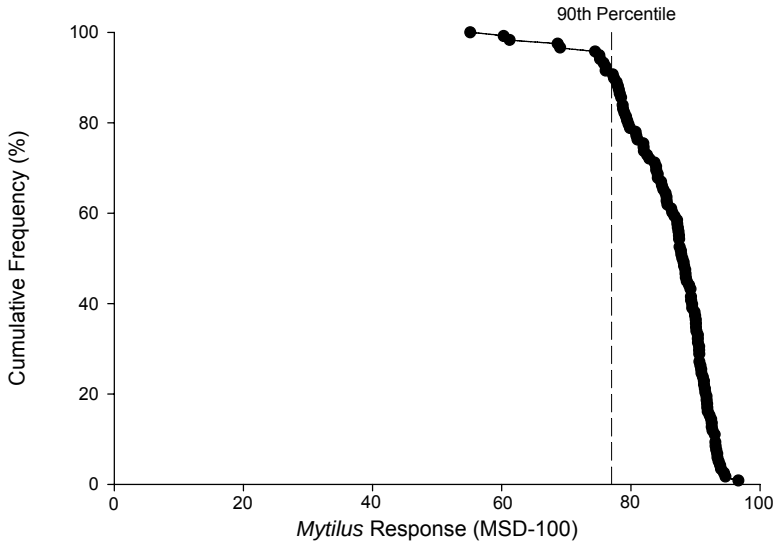

Figure 8. Cumulative frequency of *Mytilus galloprovincialis* sediment-water interface normal-alive response (100-MSD) values expressed as a percentage of response. The 90th percentile value is the moderate response threshold. Sample size = 118.

High Threshold

The species-specific MSD_{99} values were calculated using the same data described for the moderate threshold (Figures 4 through 8). The MSD_{99} values (expressed as the control normalized response) ranged from 46% for *N. arenaceodentata* to 73% for *R. abronius* (Table 7).

The 75th percentile of the toxic *E. estuarius* samples corresponded to a control-adjusted survival of 57% (Figure 9). The 75th percentile value for *R. abronius* was 66% (Figure 10). The data distribution of the toxic *M. galloprovincialis* samples from California produced the lowest 75th percentile value: 24%. This relatively low value may have been related to the small number of toxic samples available for analysis (Figure 11). The toxic data distribution approach could not be used for the *L. plumulosus* and *N. arenaceodentata* tests since most of the samples in the dataset were from outside of California. For *L. plumulosus*, the 75th percentile value of 57% from the *E. estuarius* dataset was substituted for the threshold calculation.

Calculation of the mean of the MSD_{99} and 75th percentile values produced high threshold values ranging from 42% for *Mytilus* to 70% for *R. abronius* (Table 7). This threshold was more variable than the Moderate or Low thresholds, which had ranges of 14% and 10% respectively.

The calculated toxicity test thresholds are summarized in Table 8. For application of the moderate and high thresholds, the data from each exposure must first be normalized to the control response ((sample÷control) x 100). The low threshold is evaluated using the raw data (not normalized), except for the *N. arenaceodentata* 28-day growth endpoint. Normalized data are used for the low, moderate, and *N. arenaceodentata* thresholds because these thresholds are defined relative to the control response, which can vary among tests.

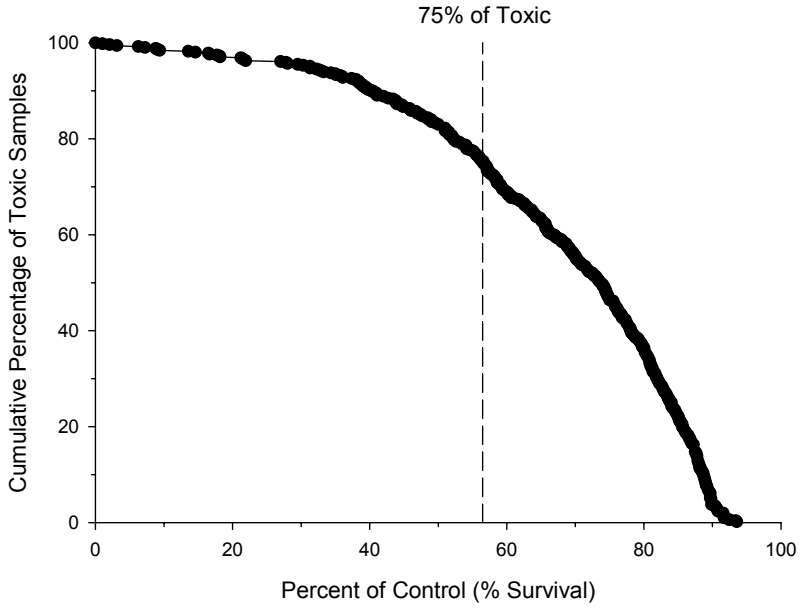

Figure 9. Cumulative frequency distribution plot of *Eohaustorius estuarius* survival data used for 75th percentile of toxic stations calculations. Sample size = 333.

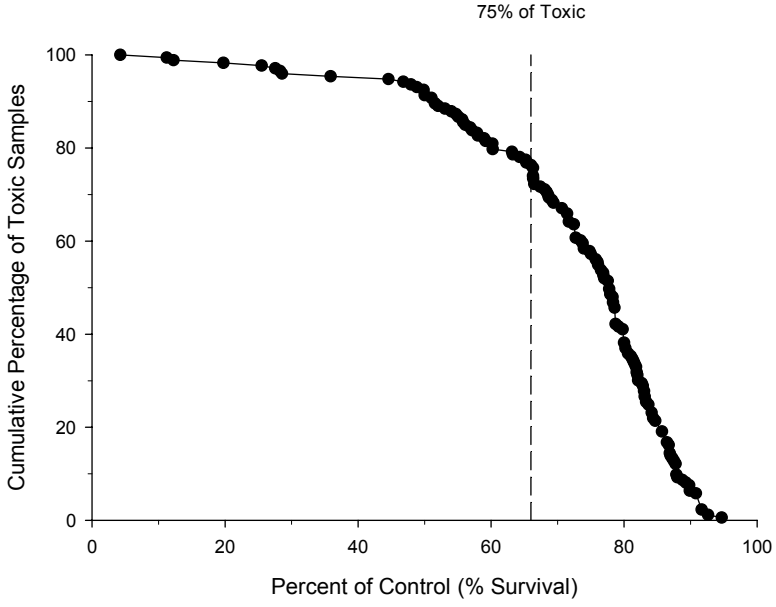

Figure 10. Cumulative frequency distribution plot of *Rhepoxynius abronius* survival data used for 75th percentile of toxic stations calculations. Sample size = 114.

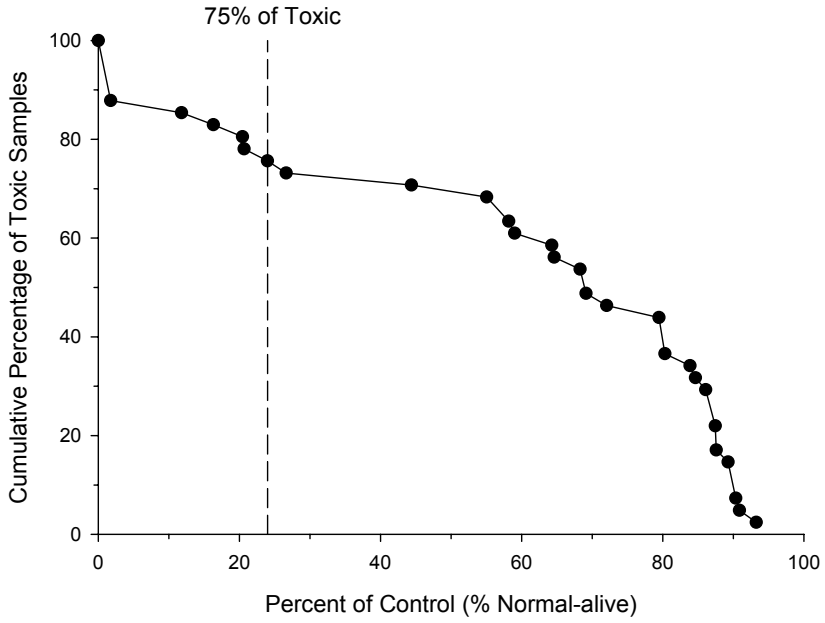

Figure 11. Cumulative frequency distribution plot of sediment-water interface method *Mytilus galloprovincialis* embryo percent normal-alive data used for threshold calculations. Sample size = 28.

Table 7. Data used in calculation of high threshold values for acute and sublethal sediment toxicity test methods. The high threshold is the mean of the two response values shown in the table.

Species	99th MSD	75th of Toxic	High Threshold
Eohaustorius estuarius	61	57	59
Rhepoxynius abronius	73	66	70
Leptocheirus plumulosus	54	57[1]	56
Neanthes arenaceodentata	46	-[2]	46
Mytilus galloprovincialis	60	24	42

[1] No California data available, so *E. estuarius* data was used for this calculation
[2]No California data available

Table 8. Toxicity threshold values for the proposed sediment toxicity test methods.

Species	Low (% Control)	Moderate (% Control)	High (% Control)
Eohaustorius estuarius	90	82	59
Rhepoxynius abronius	90	83	70
L. plumulosus	90	78	56
Neanthes arenaceodentata	90[1]	68	46
Mytilus galloprovincialis	80	77	42

[1] % of control growth.

Discussion

The thresholds derived in this study represent a unique combination of established and new approaches to achieve the goal of being able to classify sediment toxicity into multiple clearly delineated categories. By incorporating both magnitude of response and statistical uncertainty, these categories represent the two factors that are essential to describing a toxicity test response.

Thresholds based on minimum significant difference (MSD_{90}) values have been used by others to establish a threshold representing a test response associated with moderate to strong toxicity (Phillips *et al.* 2001, Field *et al.* 2002). Control acceptability criteria are also frequently used to characterize test responses. This study represents the first known application of the MSD_{99} and 75th percentile of toxic samples for classifying samples in a high toxicity category.

The thresholds developed for this study are similar to comparable thresholds calculated by others. The calculated value of 82% for the *E. estuarius* test is within the range of thresholds of 83% calculated for the Bight'03 regional monitoring project in southern California (Bay *et al.* 2005) and 75% for data from the California Bay Protection and Toxic Cleanup Program (Phillips *et al.* 2001). The moderate threshold of 77% for the SWI test with *M. galloprovincialis* is similar to the MSD value of 80% reported by Phillips *et al.* (2001) for a larger dataset for *M. galloprovincialis* that included pore water and water column data.

The *M. galloprovincialis* SWI test low and moderate thresholds appear to represent a very narrow range of response (Table 8). This response window is not as small as it first seems because the low and moderate thresholds are expressed differently. The low threshold value is not control adjusted while the moderate threshold is adjusted. The average control value for *M. galloprovincialis* SWI tests in the statewide database is 85% normal-alive. Therefore, the control-adjusted value of 77% for the moderate threshold represents a noncontrol-adjusted value of 65% (77% x 85% = 65%), representing a response window of about 15% for the low toxicity category.

Little data from California stations was available to calculate the MSD for the *L. plumulosus* and *N. arenaceodentata* test methods. This is of little concern since the MSD is a measurement of the inherent variability of the test method and should not be affected to a great extent by sample source. However, as more data becomes available the MSD should be recalculated to provide a more confident value. The thresholds for the SWI test with *M. galloprovincialis* should also be recalculated when more data become available, since the number of data points was limited in comparison to the *E. estuarius* and *R. abronius* datasets.

The greatest amount of uncertainty is associated with the high threshold values. The approach used to calculate these values is new so there is no basis of comparison to help identify spurious values. In addition, this threshold is based on the analysis of extreme portions of data distributions (99th and 75th percentiles), which are more sensitive to data quantity and may be more variable. Confidence in the high threshold values would be improved by the availability of more data collected on samples from within California. For the calculation of the 75th percentile of toxic stations, it is vital that the data is generated using California samples so that future comparison of samples from within the State will be evaluated in the correct context.

Currently, there is a very limited amount of California data for the *L. plumulosus* 10-day and *N. arenaceodentata* growth tests.

RESEARCH NEEDS

The analyses described in this report were used to select a suite of test methods for use in sediment toxicity testing. These represent a minimum suite of test methods that had the best available combination of feasibility and performance. Several data limitations were encountered in the course of this study that either restricted the suite of suitable test methods or complicated the calculation of the classification thresholds. The following research activities are needed to improve the use of toxicity tests for evaluating sediment quality:

- **Refine thresholds for the *L. plumulosus* and *N. arenaceodentata* tests as new data become available.** Limited data were available to calculate the toxicity thresholds for these species. More toxicity data from California samples are needed to refine calculation of the 75th percentile of toxic stations, which would improve confidence in the calculation of the high toxicity threshold values.

- **Evaluate additional sublethal test methods for inclusion in the suite of recommended test methods.** A wider variety of sublethal test methods that are feasible and sensitive should be available. Use of a wider variety of toxicity tests would help ensure that the toxicity information addresses variations in routes of exposure and sensitivity to sediment contaminants among the sediment-dwelling organisms. Some of the methods evaluated in the current study showed promise for future use, but were lacking in protocol development, had little field testing, and had not been compared in sensitivity to more established methods. Research is needed to fully document these tests and develop quality assurance criteria, such as required pH, salinity and temperature ranges. Research should be conducted to field test any additional methods side by side with the methods already evaluated in this document in order to evaluate relative sensitivity and produce the data needed for threshold development.

REFERENCES

Anderson, B., J. Hunt, S. Tudor, J. Newman, R. Tjeerdema, R. Fairy, J. Oakden, C. Bretz, C. Wilson, F. LaCaro, G. Kapahi, M. Stephenson, M. Puckett, J. Anderson, E. Long, T. Fleming and K. Summers. 1997. Chemistry, toxicity and benthic community conditions in sediments of selected southern California bays and estuaries. Report to California State Water Resources Control Board. Sacramento, CA.

Anderson, B.S., J.W. Hunt, M. Hester and B.M. Phillips. 1996. Assessment of sediment toxicity at the sediment-water interface. pp. 609-624 *in*: G.K. Ostrander (ed.), Techniques in aquatic toxicology. CRC Press Inc. Boca Raton, FL.

Anderson, B.S., J.W. Hunt, B.M. Phillips, R. Fairey, C.A. Roberts, J.M. Oakden, H.M. Puckett, M. Stephenson, R.S. Tjeerdema, E.R. Long, C.J. Wilson and J.M. Lyons. 2001. Sediment quality in Los Angeles Harbor, USA: A triad assessment. *Environmental Toxicology and Chemistry* 20:359-370.

Anderson, B.S., J.W. Hunt, B.M. Phillips, S. Tudor, R. Fairey, J. Newman, H.M. Puckett, M. Stephenson, E.R. Long and R.S. Tjeerdema. 1998. Comparison of marine sediment toxicity test protocols for the amphipod *Rhepoxynius abronius* and the polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environmental Toxicology and Chemistry* 17:859-866.

American Society for Testing and Materials ASTM. 1996. Standard guide for conducting 10-day static sediment toxicity tests with marine and estuarine amphipods. pp. 769-794, 1996 Annual Book of ASTM Standards, Vol. 11.05. American Society for Testing and Materials. West Conshohocken, PA.

American Society for Testing and Materials ASTM. 2002a. E724 Standard guide for conducting static acute toxicity tests starting with embryos of four species of bivalve molluscs. pp. 148-168, Section 11, Vol. 11.05, 2002 Annual Book of ASTM Standards. American Society for Testing and Materials. West Conshohocken, PA.

American Society for Testing and Materials ASTM. 2002b. E 1611 Standard guide for conducting sediment toxicity tests with Polychaetous Annelids. pp. 987-1012, Annual book of ASTM standards, Vol. 11.05. American Society for Testing and Materials. West Conshohocken, PA.

Bay, S. and J. Brown. 2003. Chemistry and toxicity in Rhine Channel sediments. Technical Report 391. Southern California Coastal Water Research Project. Westminster, CA.

Bay, S., D. Greenstein and J. Brown. 2004. Newport Bay sediment toxicity studies: Final report. Technical Report #433. Southern California Coastal Water Research Project. Westminster, CA.

Bay, S.M., D.J. Greenstein, A.W. Jirik and J.S. Brown. 1998. Southern California Bight 1994 Pilot Project: VI. Sediment toxicity. Southern California Coastal Water Research Project. Westminster, CA.

Bay, S.M., A. Jirik and S. Asato. 2003. Interlaboratory variability of amphipod sediment toxicity tests in a Cooperative Regional Monitoring Program. *Environmental Monitoring and Assessment* 81:257-268.

Bay, S.M., D. Lapota, J. Anderson, J. Armstrong, T. Mikel, A. Jirik and S. Asato. 2000. Southern California Bight 1998 Regional Monitoring Program: IV. Sediment toxicity. Southern California Coastal Water Research Project. Westminster, CA.

Bay, S.M., T. Mikel, K. Schiff, S. Mathison, B. Hester, D. Young and D. Greenstein. 2005. Southern California Bight 2003 regional monitoring program: I. Sediment toxicity. Southern California Coastal Water Research Project. Westminster, CA.

Boese, B.L., J.O. Lamberson, R.C. Swartz and R.J. Ozretich. 1997. Photoinduced toxicity of fluoranthene to seven marine benthic crustaceans. *Archives of Environmental Contamination and Toxicology* 32:389-393.

Bridges, T.S. and J.D. Farrar. 1997. The influence of worm age, duration of exposure and endpoint selection on bioassay sensitivity for *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environmental Toxicology and Chemistry* 16:1650-1658.

Bridges, T.S., J.D. Farrar and B.M. Duke. 1997. The influence of food ration on sediment toxicity in *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environmental Toxicology and Chemistry* 16:1659-1665.

Brown, J. and S. Bay. 2005. Temporal assessment of chemistry, toxicity and benthic communities in sediments at the mouths of Chollas Creek and Paleta Creek, San Diego Bay. Draft Report. Southern California Coastal Water Research Project. Westminster, CA.

Carr, R.S. and M. Nipper (eds.). 2003. Porewater toxicity testing: Biological, chemical, and ecological considerations. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

Chandler, G.T. and A.S. Green. 1996. A 14-day harpacticoid copepod reproduction bioassay for laboratory and field contaminated muddy sediments. pp. 23-39 *in*: G.K. Ostrander (ed.), Techniques in aquatic toxicology. CRC Press. Boca Raton, FL.

Chapman, P.M., F. Wang, J.D. Germano and G.E. Batley. 2002. Pore water testing and analysis: the good, the bad, and the ugly. *Marine Pollution Bulletin* 44:359-366.

DeWitt, T.H., G.R. Ditsworth and R.C. Swartz. 1988. Effects of natural sediment features on survival of the Phoxocephalid Amphipod, *Rhepoxynius abronius*. *Marine Environmental Research* 25:99-124.

DeWitt, T.H., M.R. Pinza, L.A. Niewolny, V.I. Cullinan and B.D. Gruendell. 1997. Development and evaluation of standard marine/estuarine chronic sediment toxicity test method using *Leptocheirus plumulosus*. Battelle Marine Sciences Laboratory. Sequim, WA.

Dillon, T.M., D.W. Moore and A.B. Gibson. 1993. Development of a chronic sublethal bioassay for evaluating contaminated sediment with the marine polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environmental Toxicology and Chemistry* 12:589-605.

Fairey, R., C. Roberts, M. Jacobi, S. Lamerdin, R. Clark, J. Downing, E. Long, J. Hunt, B. Anderson, J. Newman, R. Tjeerdema, M. Stephenson and C. Wilson. 1998. Assessment of sediment toxicity and chemical concentrations in the San Diego Bay region, California, USA. *Environmental Toxicology and Chemistry* 17:1570-1581.

Farrar, J.D., T.S. Bridges and G.R. Lotufo. 1998. Comparative sediment toxicity in the marine polychaetes *Polydora cornuta, Boccardia proboscidea* and *Neanthes arenaceodentata*, the estuarine amphipod *Leptocheirus plumulosus* and copepod *Schizopera knabeni*. Paper presented at the Society of Environmental Toxicology and Chemistry Annual Meeting. Charlotte, NC.

Field, L.J., D.D. MacDonald, S.B. Norton, C.G. Ingersoll, C.G. Severn, D. Smorong and R. Lindskoog. 2002. Predicting amphipod toxicity from sediment chemistry using logistic regression models. *Environmental Toxicology and Chemistry* 21:1993-2005.

Fulton, M.H., G.I. Scott, P.B. Key, G.T. Chandler, R.F. Van Dolah and P.P. Maier. 1999. Comparative toxicity testing of selected benthic and epibenthic organisms for the development of sediment quality test protocols. EPA/600/R-99/085. U.S. Environmental Protection Agency. Washington, D.C.

Gardiner, W.W. and L.A. Niewolny. 1998. Results of an interlaboratory evaluation of the 28-day chronic sediment toxicity test using *Neanthes arenaceodentata*. Battelle. Sequim, WA.

Green, A.S., D. Moore and D. Farrar. 1999. Chronic toxicity of 2,4,6-trinitrotoluene to a marine polychaete and an estuarine amphipod. *Environmental Toxicology and Chemistry* 18:1783-1790.

Ho, K.T., R.M. Burgess, M.C. Pelletier, J.R. Serbst, S.A. Ryba, M.G. Cantwell, A. Kuhn and P. Raczelowski. 2002. An overview of toxicant identification in sediments and dredged materials. *Marine Pollution Bulletin* 44:286-293.

Hunt, J.W., B.S. Anderson, B.M. Phillips, R.S. Tjeerdema, K.M. Taberski, C.J. Wilson, H.M. Puckett, M. Stephensen, R. Fairey and J. Oakden. 2001. A large-scale categorization of sites in San Francisco Bay, USA, based on the sediment quality triad, toxicity identification evaluations, and gradient studies. *Environmental Toxicology and Chemistry* 20:1252-1265.

Hyland, J.L., L. Balthis, C.T. Hackney, G. McRae, A.H. Ringwood, T.R. Snoots, R.F. Van Dolah and T.L. Wade. 1998. Environmental quality of estuaries of the Carolinian Province: 1995. NOS ORCA 123. National Oceanic and Atmospheric Administration. Charleston, SC.

Kennedy, A., J.D. Farrar, J.A. Steevens and M. Reiss. 2004. Evaluation of the applicability of standard toxicity test methods to dredged material management. Paper presented at the Society of Environmental Toxicology and Chemistry Annual Meeting. Portland, OR.

Keppler, C.J. and A.H. Ringwood. 2002. Effects of metal exposures on juvenile clams, *Mercenaria mercenaria*. *Bulletin of Environmental Contamination and Toxicology* 68:43-48.

Lamberson, 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, Inc. Boca Raton, FL.

Long, E., A. Robertson, D. Wolfe, J. Haeedi and G. Sloane. 1996. Estimates of the spatial extent of sediment toxicity in major U.S. estuaries. *Environmental Science and Technology* 30:3585-3592.

Long, E.R. 2000. Spatial extent of sediment toxicity in U.S. estuaries and marine bays. *Environmental Monitoring and Assessment* 64:391-407.

Long, E.R., M.F. Buchman, S.M. Bay, R.J. Breteler, R.S. Carr, P.M. Chapman, J.E. Hose, A.L. Lissner, J. Scott and D.A. Wolfe. 1990. Comparative evaluation of five toxicity tests with sediments from San Francisco Bay and Tomales Bay, California. *Environmental Toxicology and Chemistry* 9:1193-1214.

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* 16:405-415.

Long, E.R., M. Dutch, S. Aasen, K. Welch and M.J. Hameedi. 2005. Spatial extent of degraded sediment quality in Puget Sound (Washington state, USA) based upon measures of the sediment quality triad. *Environmental Monitoring and Assessment* 111:173-222.

Long, E.R., D.D. MacDonald, C.G. Severn and C.B. Hong. 2000. Classifying probabilities of acute toxicity in marine sediments with empirically derived sediment quality guidelines. *Environmental Toxicology and Chemistry* 19:2598-2601.

Long, E.R., G.M. Sloane, G.I. Scott, B. Thompson, R.S. Carr, J. Biedenback, T.L. Wade, B.J. Presley, K.J. Scott, C. Mueller, G. Brecken-Fols, B. Albrecht, J.W. Anderson and G.T. Chandler. 1999. Magnitude and extent of chemical contamination and toxicity in sediments of Biscayne Bay and vicinity. NOS NCCOS CCMA 141. National Oceanic and Atmospheric Administration. Silver Spring, MD.

Lotufo, G.R., J.D. Farrar and T.S. Bridges. 2000. Effects of exposure source, worm density, and sex on DDT bioaccumulation and toxicity in the marine polychaete *Neanthes arenaceodentata*. *Environmental Toxicology and Chemistry* 19:472-484.

Lotufo, G.R., J.D. Farrar, B.M. Duke and T.S. Bridges. 2001a. DDT toxicity and critical body residue in the amphipod *Leptocheirus plumulosus* in exposures to spiked sediment. *Archives of Environmental Contamination and Toxicology* 41:142-150.

Lotufo, G.R., J.D. Farrar, L.S. Inouye, T.S. Bridges and D.B. Ringelberg. 2001b. Toxicity of sediment-associated nitroaromatic and cyclonitramine compounds to benthic invertebrates. *Environmental Toxicology and Chemistry* 20:1762-1771.

McGee, B.L., D.J. Fisher, D.A. Wright, L.T. Yonkos, G.P. Ziegler, S.D. Turley, J.D. Farrar, D.W. Moore and T.S. Bridges. 2004. A field test and comparison of acute and chronic sediment toxicity tests with the estuarine amphipod *Leptocheirus plumulosus* in Chesapeake Bay, USA. *Environmental Toxicology and Chemistry* 23:1751-1761.

McGee, B.L., D.J. Fisher, L.T. Yonkos, G.P. Ziegler and S. Turley. 1999. Assessment of sediment contamination, acute toxicity, and population viability of the estuarine amphipod *Leptocheirus plumulosus* in Baltimore Harbor, Maryland, USA. *Environmental Toxicology and Chemistry* 18:2151-2160.

Mearns, A.J., R.C. Swartz, J.M. Cummins, P.A. Dinnel, P. Plesha and P.M. Chapman. 1986. Inter-laboratory comparison of a sediment toxicity test using the marine amphipod, *Rhepoxynius abronius*. *Marine Environmental Research* 19:13-37.

Moore, D.W., M.A. Irwin, B. Hester, D. Diener and J.Q. Word. 2003. Field validation of chronic sublethal dredged material laboratory bioassays. Society of Environmental Toxicology and Chemistry Annual Meeting. Austin, TX.

Phillips, B.M., J.W. Hunt, B.S. Anderson, H.M. Puckett, R. Fairey, C.J. Wilson and R. Tjeerdema. 2001. Statistical significance of sediment toxicity results: Threshold values derived by the detectable significance approach. *Environmental Toxicology and Chemistry* 20:371-373.

Pinza, M.R., J.A. Ward and N.P. Kohn. 2002. Results of interspecies toxicity comparison testing associated with contaminated sediment management. Paper presented at Sediment Management Annual Review Meeting. Seattle, WA.

Puget Sound Water Quality Authority (PSWQA). 1995. Recommended guidelines for conducting laboratory bioassays on Puget Sound sediments. Puget Sound Water Quality Authority for U.S. Environmental Protection Agency Region 10. Olympia, WA.

Ringwood, A.H., D.E. Conners and J. Hoguet. 1998. Effects of natural and anthropogenic stressors on lysosomal destabilization in oysters *Crassostrea virginica*. *Marine Ecology Progress Series* 166:163-171.

Ringwood, A.H., J. Hoguet, C.J. Keppler, M.L. Gielazyn, B.P. Ward and A.R. Rourk. 2003. Cellular biomarkers (lysosomal destabilization, glutathione & lipid peroxidation) in three common estuarine species: A methods handbook. Marine Resources Research Institute. Charleston, SC.

Ringwood, A.H., A.F. Holland, R.T. Kneib and P.E. Ross. 1996. EMAP/NS&T pilot studies in the Carolinian Province: Indicator testing and evaluation in the Southeastern estuaries. NOS ORCA 102. National Atmospheric and Oceanic Administration. Silver Springs, MD.

Ringwood, A.H. and C.J. Keppler. 1998. Seed clam growth: An alternative sediment bioassay developed during EMAP in the Carolinian Province. *Environmental Monitoring and Assessment* 51:247-257.

Schlekat, C.E., K.J. Scott, R.C. Swartz, B. Albrecht, L. Antrim, K. Doe, S. Douglas, J.A. Ferretti, D.J. Hansen, D.W. Moore, C. Mueller and A. Tang. 1995. Interlaboratory comparison of a 10-day sediment toxicity test method using *Ampelisca abdita, Eohaustorius estuarius* and *Leptocheirus plumulosus*. *Environmental Toxicology and Chemistry* 14:2163-2174.

State Water Resources Control Board (SWRCB). 2006. Development of Sediment Quality Objectives for Enclosed Bays and Estuaries. California State Water Resources Control Board. Sacramento, CA.

U.S. Environmental Protection Agency (USEPA). 1991. Evaluation of dredged material proposed for ocean disposal (Testing Manual). EPA/503/8-91/001. United States Environmental Protection Agency and Department of The Army U.S. Army Corps of Engineers. Washington, DC.

USEPA. 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA/600/R-94/025. Office of Research and Development, U.S. Environmental Protection Agency. Narragansett, RI.

USEPA. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. Office of Research and Development. Cincinnati, OH.

USEPA. 1998. Guidelines for Ecological Risk Assessment. EPA/630/R-95/002F. U.S. Environmental Protection Agency. Washington, D.C.

USEPA. 2001. Methods for assessing the chronic toxicity of marine and estuarine sediment-associated contaminants with the amphipod *Leptocheirus plumulosus*. U.S. Environmental Protection Agency. Washington, D.C.

USEPA. 2004. National coastal condition report II. EPA-620/R-03/002. U.S. Environmental Protection Agency, Office of Water. Washington, DC.

USEPA and U.S. Army Corps of Engineers. 1998. Evaluation of Dredged Material Proposed for Discharge in Waters of the U.S. - Testing Manual. EPA-823-B-98-004. U.S. Environmental Protection Agency. Washington, D.C.

Van Dolah, R.F., J.L. Hyland, A.F. Holland, J.S. Rosen and T.R. Snoots. 1999. A benthic index of biological integrity for assessing habitat quality in estuaries of the southeastern USA. *Marine Environmental Research* 48:269-283.

APPENDIX A

Comparison of Methods for Evaluating Acute and Chronic Toxicity in Marine Sediments

Darrin Greenstein[1], Steven Bay[1], Brian Anderson[2], G. Thomas Chandler[3], J. Daniel Farrar[4], Charles Keppler[5,] Bryn Phillips[2], Amy Ringwood[6], and Diana Young[1]

[1] Southern California Coastal Water Research Project, Westminster CA
[2] University of California Davis, Monterey, CA
[3] University of South Carolina, Columbia, SC
[4] US Army Engineer Research and Development Center, Vicksburg, MS
[5] Marine Resources Research Institute, Charleston, SC
[6] University of North Carolina-Charlotte, Charlotte, NC

ABSTRACT

Sublethal test methods are being used with increasing frequency to measure sediment toxicity, but little is known about the relative sensitivity of these tests compared to the more commonly used acute tests. A study was conducted to compare the sensitivity of several acute and sublethal toxicity methods, and investigate their correlations with sediment chemistry and benthic community condition. Six sublethal methods (amphipod, *Leptocheirus plumulosus* 28-day survival, growth and reproduction; polychaete, *Neanthes arenaceodentata* 28-day survival and growth; benthic copepod, *Amphiascus tenuiremis,* 14-day life cycle; seed clam, *Mercenaria mercenaria* 7-day growth; oyster, *Crassostrea virginica* lysosome destabilization; and sediment-water interface (SWI) testing with embryos of the mussel *Mytilus galloprovincialis*) and two acute methods (10-day amphipod survival with *Eohaustorius estuarius* and *Leptocheirus plumulosus*) were used to test split samples of sediment from stations in southern California and San Francisco Bay. The most sensitive sublethal test, and most sensitive overall, was the life cycle test with the copepod, *Amphiascus*. The *L. plumulosus* 10-day survival test was the most sensitive of the acute tests. The sublethal tests were not, in general, more sensitive to the sediments than the acute tests. Of the sublethal tests only the *A. tenuiremus* endpoints and polychaete growth correlated with sediment chemistry. There was poor correspondence between the toxicity endpoints and indicators of benthic community condition. Differences in test characteristics such as mode of exposure, species-specific contaminant sensitivity, changes in contaminant bioavailability, and the influence of noncontaminant stressors on the benthos may have been responsible for the variations in response among the tests and low correspondence with benthic community condition. The influence of these factors cannot be easily predicted and underscores the need to use multiple toxicity methods in combination with other lines of evidence to provide an accurate and confident assessment of sediment toxicity.

ACKNOWLEDGMENTS

The authors would like to the field crews and analytical laboratories participating in the Southern California Bight 2003 Regional Survey and Sarah Lowe (San Francisco Estuary Institute) for their assistance in sample collection and analysis. We also wish thank J. Ananda Ranasinghe and Bruce Thompson for providing analysis of the benthic community data for this study.

INTRODUCTION

Acute sediment toxicity testing has been routinely conducted as part of monitoring and assessment programs, such as the USEPA's Environmental Monitoring and Assessment Program (Strobel *et al.* 1995). The toxicity tests are usually conducted on whole sediments using amphipod 10-day survival tests in accordance with standard protocols (USEPA 1994). Sublethal testing has been conducted on a much more limited basis, but there is increased interest in using sublethal methods due to the assumption that they are more sensitive to contaminated sediments than the acute methods (Adams *et al.* 2005). Sublethal methods include embryo development tests and other tests with various life stages of animals having endpoints such as growth and reproduction in addition to survival. A wide variety of sublethal methods have been described (Lamberson *et al.* 1992), but only a few such methods have been used commonly; they include the amphipod *Leptocheirus plumulosus* 28-day growth and reproduction test (USEPA 2001), a 20-day polychaete growth test using *Neanthes arenaceodentata* (PSWQA 1995), pore water testing using echinoderm gametes or embryos (Carr and Nipper 2003) and a SWI test using sea urchin or mussel embryos (Anderson *et al.* 1996). Additional promising sublethal tests that have been developed recently and include the measurement of copepod reproduction (Chandler and Green 1996), juvenile clam growth (Ringwood and Keppler 1998), and oyster biomarker responses (Ringwood *et al.* 1998).

Because sublethal toxicity methods have been used less commonly, there are questions regarding whether these test methods are practical, reproducible, and more sensitive than the acute methods already in use (Anderson *et al.* 1998, Pinza *et al.* 2002). Few studies have been conducted that were designed specifically to compare the relative attributes of various sublethal tests. Studies conducted to date have only compared two or three methods together (DeWitt *et al.* 1997, Anderson *et al.* 1998, Green *et al.* 1999), or have focused more on sublethal elutriate or pore water tests rather than whole sediment tests (Long *et al.* 1990). Important factors to consider in the selection and interpretation of toxicity tests include the degree of exposure to whole sediment, the relative sensitivity to sediment contaminants, and the level of concordance with benthic community impacts. Information on these factors is extremely limited for many sublethal tests.

This study was designed to investigate relative performance of several acute and sublethal test methods with whole sediments. Three specific points were examined. First, the relative sensitivity of the toxicity test methods was compared. Sensitivity was defined as the relative ability of a test method to detect toxicity in a sample. Sensitivity comparisons were made both between acute and sublethal methods and among the sublethal methods. Secondly, the relationship between sediment chemical concentrations and toxicity of each method was examined. Finally, this study investigated the relationship between changes in benthic community condition and toxicity.

METHODS

Six candidate whole sediment sublethal methods were selected (Table 1). These methods appeared to be technically feasible and had data available that indicated some level of sensitivity to contaminated sediments. Methods were first selected that had established, published methods by a government or scientific agency (e.g., USEPA methods, ASTM methods). Additional methods were selected from the scientific literature and from recommendations by toxicologists with experience in sediment quality assessment. Acute amphipod testing was also conducted for comparison with sublethal methods using two species, *E. estuarius* and *L. plumulosus*.

Table 1. Characteristics of the sublethal sediment toxicity methods included in the comparison study. Duration given in days.

Species	Taxon	Test endpoint(s)	Duration
Mytilus galloprovincialis	mussel	sediment-water interface, embryo development	2
Mercenaria mercenaria	clam	growth	7
Crassostrea virginica	oyster	lysosomal destabilization	4
Leptocheirus plumulosus	amphipod	growth, reproduction, survival*	28
Neanthes arenaceodentata	polychaete	growth, survival*	28
Amphiascus tenuiremis	benthic copepod	reproduction, survival*	14

* Secondary endpoint

The sediment samples that were tested were collected as part of two regional monitoring surveys, Southern California Bight 2003 Regional Monitoring Program (Figure 1) and the San Francisco Estuary Institute Regional Monitoring Program (RMP; Figure 2). The stations represented a wide range of expected contamination levels and habitat types with the aim being to target stations expected to have a low to moderate level of acute toxicity. Stations expected to have a high degree of acute toxicity were not included in the study because they would be less effective in eliciting different sublethal responses among the tests. The stations from southern California were selected to include a range of geographical location, proximity to sources of contamination, and expected sediment grain size. The RMP sites have been monitored for about 10 years and were chosen based on their wide geographic distribution and a range of acute toxicity to amphipods.

Tests on split samples were conducted by laboratories with extensive experience using the various tests. The *L. plumulosus* and *N. arenaceodentata* testing was conducted at the Army Corps of Engineers, Research and Development Center, Environmental Laboratory in Vicksburg, MS. The *A. tenuiremis* assays were performed at the University of South Carolina in Columbia, SC. The *M. mercenaria* growth test and *C. virginica* lysosomal destabilization procedures were done at the South Carolina Department of Natural Resources, Marine Resources Research Institute in Charleston, SC. The SWI testing was conducted at the University of California, Davis, Marine Pollution Studies Laboratory in Carmel, CA. Ten-day *E. estuarius* acute survival tests were performed on sediment from each station. These acute tests were performed by multiple laboratories, as part of the regional monitoring efforts. The laboratories that performed the *E. estuarius* tests on southern California stations participated in intercalibration procedures, which showed

reasonable agreement between laboratories (Bay *et al.* 2005). The laboratory testing the San Francisco Bay stations did not participate in this intercalibration. A summary of the characteristics of all of these test methods can be found in Bay *et al.* (2007). Samples were also analyzed for organic and metals chemistry, total organic carbon (TOC), grain size and benthic infauna.

Sediments were collected in July through August 2003. A Van Veen grab was used to collect whole sediment from the surface (top 2 cm) and subcores. Surface sediment was obtained from multiple grabs at each site, composited, transferred to plastic containers, and stored at 5°C. Sediment-water interface subcores were also collected from the Van Veen grab by inserting a polycarbonate core tube into the sediment to a depth of 5 cm and capping the bottom and top of the tube. All sediment samples were transported to Southern California Coastal Water Research Project (SCCWRP) within 24 hours of collection. The core samples were then transported with ice packs to the testing laboratory within 24 hours. Core samples from the San Francisco Bay stations were transported directly to the testing laboratory.

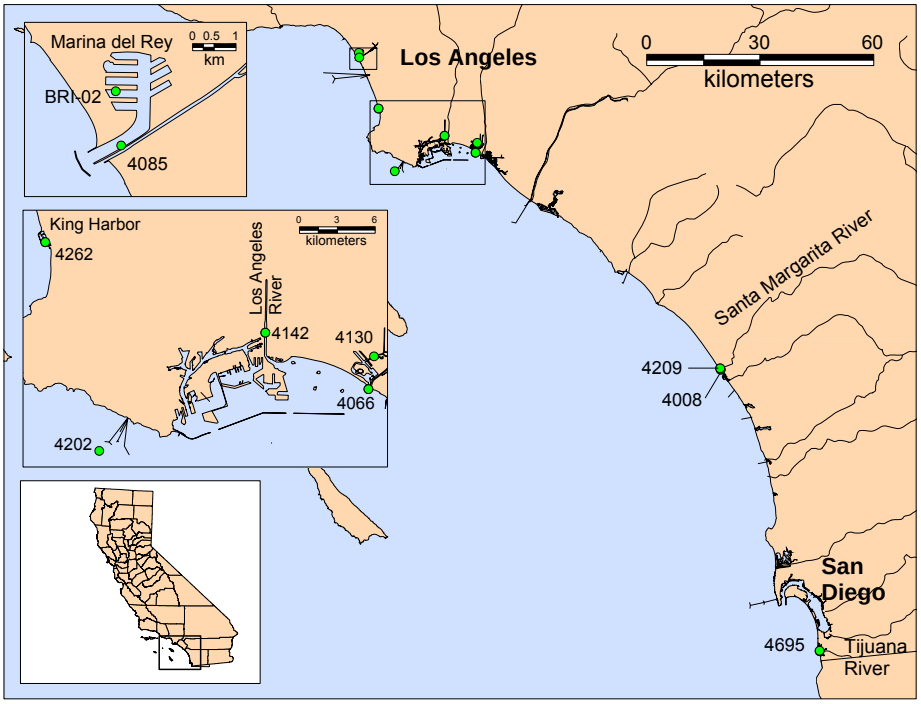

Figure 1. Location of southern California stations used for the sediment toxicity methods comparison study.

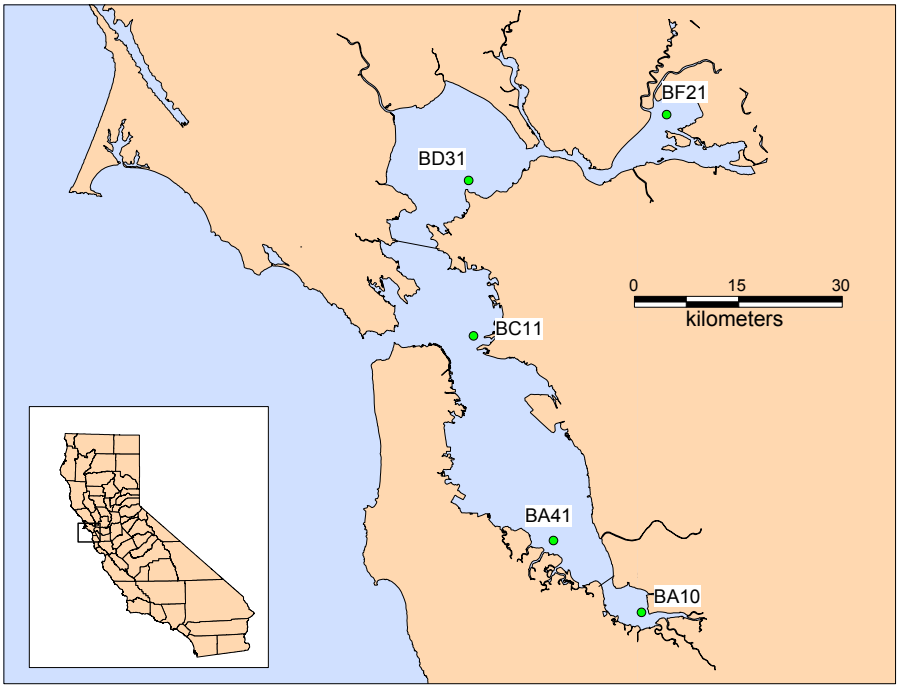

Figure 2. Location of San Francisco Bay stations used for the sediment toxicity methods comparison study.

The subcores were shipped to the testing laboratory within 48 hours of collection and the SWI tests were initiated within 10 days of collection (Table 2). The whole sediment samples were shipped to the testing labs in two batches, one with six of the southern California stations, the other with the remaining four southern California and all five San Francisco stations. Before shipment of each batch, all of the sediment from each station was placed in a large polycarbonate bowl and homogenized with a polycarbonate spoon. Samples for each laboratory were then aliquoted into polyethylene containers and shipped overnight with sufficient quantities of ice packs to maintain temperature at 5°C. Holding time between collection and testing of the composites varied from 6 to 116 days (Table 2).

Table 2. Holding times (number of days) for sediment samples tested with acute and sublethal toxicity methods. *Eohaustorius estuarius* (*Eohaustorius*), *Leptocheirus plumulosus* (*Leptocheirus*), Sediment-water Interface (SWI), *Merceneria mercenaria* (*Mercenaria*), *Crassostrea virginica* (Lysosome), *Neanthes arenaceodentata (Neanthes)*, and *Amphiascus tenuiremus* (*Amphiascus*).

		Leptocheirus						
Station	*Eohaustorius*	10-Day	28-Day	SWI	*Mercenaria*	Lysosome	*Neanthes*	*Amphiascus*
Batch 1								
4066	27	26	116	6	13	26	32	-
4130	26	26	116	6	13	26	32	12
4142	27	26	116	6	13	29	32	-
4008	11	22	112	10	9	25	28	8
4209	11	22	112	10	9	22	28	8
4695	10	21	111	9	8	24	27	7
Batch 2								
4202	13	41	90	6	21	37	58	19
4262	12	40	89	5	20	36	57	18
BRI-02	14	28	77	1	8	24	45	6
4085	7	28	77	1	8	24	45	6
BA10	8	36	85	1	16	32	53	-
BA41	11	39	88	4	19	35	56	17
BC11	13	41	90	6	21	34	58	19
BD31	13	41	90	6	21	34	58	-
BF21	15	43	92	8	23	36	60	-

Toxicity Testing

Eohaustorius estuarius 10-day survival

Ten day survival tests with *E. estuarius* were conducted using standard USEPA testing procedures (1994). Sediment samples were pre-sieved through a 2-mm mesh screen and homogenized in the laboratory before testing. Sediment was placed in 1-L glass jars to a depth of 2 cm. The samples were aerated and allowed to equilibrate overnight before addition of 20 adult amphipods to each of five replicates. All of the laboratories obtained the amphipods from Northwestern Aquatic Sciences (Yaquina Bay, OR). The exposures took place at 15°C, at a salinity of 20 g/kg with constant lighting. The animals were not fed and the water was not renewed during the exposures. At the end of the exposure, the sediment from each jar was sieved and the surviving animals were counted and recorded. Water quality measurements (dissolved oxygen, pH, salinity and overlying water ammonia) were determined at day 0 and prior to test termination.

Leptocheirus plumulosus 10-day survival

The experimental design followed guidelines set forth by the USEPA (1994). Sediment was added to each of 5 replicate 1-L beakers to obtain a 2 cm depth. Sediment was then overlain with 20 g/kg synthetic seawater. Temperature was maintained at 25°C with constant illumination and the beakers were aerated during the exposure. At day 0, 20 *L. plumulosus* (500- to 750-µm sieve size class) obtained from in-house cultures were gently transferred to each replicate beaker. The animals were not fed and the water was not renewed during the exposures. Water quality measurements (dissolved oxygen, pH, salinity and overlying water ammonia) were determined at day 0 and prior to test termination. On day 10, the sediment in each beaker was sieved and the surviving animals recovered. The number of surviving organisms was counted and recorded.

Leptocheirus plumulosus 28-day survival, growth and reproduction

The *L. plumulosus* 28-day experiments were conducted following the guidelines provided by the USEPA (2001). Due to conflicts in the laboratory schedule and a test failure, the samples for this test method were held for a much longer period than the other test methods (Table 2). Sediment was added to 5 replicate 1-L beakers to obtain the required depth of 2 cm. Sediment was then overlain with 20 g/kg synthetic seawater and gently aerated. Temperature was maintained at 25°C and the light cycle was set at 16:8 h light:dark. At day 0, *L. plumulosus* (250- to 600-µm sieve size class) were obtained from in-house cultures. Twenty animals were transferred to each replicate beaker. Water quality measurements (dissolved oxygen, pH, salinity and overlying water ammonia) were determined at day 0, prior to test termination and in one replicate per sediment three times per week. Water was changed in each beaker after water quality parameters were measured. Each beaker was provided with 20 mg of Tetramin® three times per week for the first two weeks and 40 mg per beaker the final two weeks of testing. On day 28, the sediment in each beaker was sieved and surviving animals were recovered. Surviving adults and neonates were counted and recorded. The surviving adults from each replicate were placed on a tared pan and dried at 60°C for 24 hours. The pans were then removed, allowed to cool and weighed to obtain total dry-weight for each replicate. The reproductive endpoint had an acceptability criteria failure in batch and an abnormal response in another. It was therefore decided that the reproductive data would not be used for analysis.

Neanthes arenaceodentata 28-day survival and growth

The 28-d *N. arenaceodentata* experiments were conducted following guidelines developed by the US Army ERDC (Bridges and Farrar 1997, Bridges *et al.* 1997). Sediment was added to 10 replicate 300 ml tall-form beakers to obtain the required depth of 2 cm. Sediment was then overlain with 30 g/kg synthetic seawater and gently aerated. Temperature was maintained at 20°C and light cycle was set at 12:12 hour light:dark. Organisms were obtained from Dr. Don Reish (California State University, Long Beach, CA). On day 0, one *N. arenaceodentata* (≤7 days old) was gently transferred to each replicate beaker. Water quality measurements (dissolved oxygen, pH, salinity and overlying water ammonia) were determined at day 0, prior to test termination, and in three replicates per sample weekly. Water was changed in each beaker once per week after water quality parameters were measured. Each beaker was provided 2 mg of

Tetramarin® once per week and 2 mg of Tetramarin® plus 2 mg of Alfalfa once per week. On day 28, the sediment contained in each beaker was sieved and surviving worms were recovered, counted and recorded. Surviving animals from each replicate were put on a pre-weighed pan and placed in a drying oven at 60°C for 24 hours. The pans were then removed, allowed to cool, and weighed to obtain the individual dry weight for each replicate/animal.

Amphiascus tenuiremis 14-day life cycle

Testing of the copepods followed the methods of Chandler and Green (1996). A sediment reference sample was collected from Oyster Landing at North Inlet, SC. Stations BA41, BC11, BRI2, 4085, 4202, 4262, were press-sieved through a 125 µm sieve in order to facilitate recovery of the animals at the conclusion of the exposure. A larger sieve size was used for some of the larger grained stations in order to obtain a sufficient volume of sediment for testing. Sediment samples 4008 and 4695 were screened with a 250 µm sieve while 4209 and 4130 were sieved through a 212 and 180 µm sieve, respectively. Sediment samples 4066 and 4142 were too sandy to pass a 250 µm sieve, and could not be tested with the copepod method. A total of ten stations were tested with *Amphiascus*. Teflon 50-ml Erlenmeyer flasks with mesh-covered outflow holes were filled with 0.45-µm filtered, aerated seawater. Press-sieved sediment samples were then packed into Teflon syringes and slowly extruded onto the bottoms of their respective chambers (4 replicates per sediment sample). Adult non-gravid female and adult male copepods (*Amphiascus tenuiremis*) were then counted into each quadruplicate test chamber (25/sex). Chambers were placed in an incubator at 20°C under continuous dripping flow for 14 days with a 12:12 hour light:dark cycle. Chambers were fed every third day a mixture of frozen algal stock (10^7 cells of 1:1:1 *Isochrysis galbana, Phaeodactylum tricornutum* and *Dunaliella tertiolecta*). Water quality parameters (dissolved oxygen, pH, and salinity) were measured every third day. Overlying water ammonia was measured once, at the end of each exposure period. At the end of 14 days of exposure, copepods were collected on a 63-µm sieve. Samples were checked/counted for dead bodies. Copepods were stained with Rose Bengal and preserved in 5% borate-buffered formalin. Non-gravid adult females, gravid adult females, adult males, copepodites, nauplii, and clutch sizes were enumerated under a Nikon SMZ-U stereo dissection microscope. Two endpoints of the *A. tenuiremus* test were calculated: the number of copepodites produced and the realized offspring production (output of new animals normalized to the number of females surviving at the end of the test).

Mercenaria mercenaria 7-day growth

The clam tests measured growth during a 7-day exposure to whole sediment (Ringwood and Keppler 1998, Keppler and Ringwood 2002). Sediment samples were pressed through a 500-µm sieve, homogenized, and 50-ml aliquots were placed into four replicate 250-ml beakers. The sediment was then overlain with clean 25 g/kg seawater. The replicates were gently aerated for the duration of the experiment, and the assays were conducted at room temperature (22 to 25°C) for 7 days with a 16:8 light cycle. Juvenile clams (*M. mercenaria*) used for all experiments were obtained from Atlantic Littleneck Clam Farm, Charleston, SC. Clams were sieved through two mesh sizes (1.0 mm and 1.2 mm) to ensure that the clams were of a similar size range. Twenty-five clams were used

for each replicate. Pre-assay wet weights of each clam group were taken for growth rate estimates, and to ensure that all replicate groups had similar initial weights. Replicate subsets of clams were also counted, wet weighed, dried overnight and reweighed to verify the wet:dry weight ratio used to estimate initial dry weights. Each replicate was fed on the first, third and sixth days of the assay (50:50 mix of *I. galbana* and *Chaetoceros gracilis*: 20 x 10^6 cells / replicate). The overlying water was not renewed during the exposure. At the end of the exposures, clams were sieved from the sediments and placed in fresh 25 g/kg seawater for approximately 2 hours to depurate. Dead clams were counted and removed, and percent mortalities were calculated. The surviving clams were counted and rinsed with distilled water to remove excess salts. Post-assay wet weights were determined, and clams were then dried for 48 hours (at 70°C). Each clam replicate was recounted and final dry weight per clam was determined. Initial dry weights were subtracted from the final dry weights, and the results were expressed as growth rates (µg/clam/day). Sediment pore water chemistry parameters (salinity, pH, and total ammonia – nitrogen [TAN]) were measured for each sediment sample prior to use in any assay. Overlying water quality was also measured.

Crassostrea virginica 4-day lysosomal destabilization

The lysosomal destabilization assay was conducted following the methods described in Ringwood et al. (1998). Sediment samples were homogenized and 100-ml aliquots were placed into three replicate 1L beakers. The sediment was topped with clean 25 g/kg seawater. The beakers were allowed to settle for 2 hours, and then 3 clean-scrubbed oysters were gently added to each replicate. Oysters (5.3 $\pm$0.7 cm) used for laboratory sediment exposures were collected from control sites and acclimated to laboratory conditions for at least 24 hours prior to the start of the experiment. The replicates were gently aerated for the duration of the experiment, and the assays were conducted at room temperature (22 to 25°C) for 4 days with a 16:8 light cycle. Each replicate was fed on the first and third days of the assay (algal paste mixed into filtered sea water, 70 x 10^6 cells / replicate). The overlying water was not renewed during the exposure. Water quality parameters for both the pore and overlying waters were measured in the same way as for the *M. mercenaria* testing. Digestive gland tissue from the exposed oysters was diced and treated with trypsin to produce a cell suspension. A cell suspension aliquot was mixed with an equal aliquot of neutral red (NR) solution, placed on a microscope slide and examined under a light microscope to evaluate NR retention by digestive gland cells containing lysosomes. At least 50 cells were scored as stable (NR retention in the lysosomes) or destabilized (NR leaking into the cytoplasm), and the data were expressed as the percentage of cells with destabilized lysosomes per oyster.

Mytilus galloprovincialis 2-day embryo development at the sediment-water interface

Exposure procedures followed those detailed by Anderson *et al.* (1996). One day prior to the start of the test, 300 ml of clean seawater (1-µm filtered, approximately 34 g/kg) was added over the sediment to each of five replicate core tubes. Samples were then aerated overnight to equilibrate. On test day 0, water quality samples were collected from the core tubes and tubes containing a 25-µm screen were placed on the sediment surface. The screen was approximately 1 cm above the sediment. Mussel embryos were unavailable to test stations 4008, 4209, and 4695, so sea urchin embryos were used

instead. Embryos were prepared following USEPA protocols (USEPA 1995) and added to the screen tubes. Mussels were exposed for 48 hours and sea urchins for 96 hours. Exposures were carried out at 15°C with gentle aeration. Water quality parameters of dissolved oxygen, total ammonia, pH, and salinity were measured at the beginning and end of the exposure period. Temperature was measured continuously. The exposures were terminated by removing the screen tube, rinsing the embryos into a vial, and adding formalin to fix and preserve embryos. The samples were then examined microscopically for normal embryo development. Data were expressed as percentage normal-alive. This endpoint is calculated by dividing the number of normal embryos by initial number of embryos inoculated into the chambers.

Chemical Analysis

Sediment samples were analyzed for a suite of parameters that included metals, organics, grain size and TOC. Analyses were conducted by a variety of laboratories participating in the regional monitoring programs and used standardized EPA recommended methods (Bight'03 Coastal Ecology Committee 2003, SFEI 2005). The laboratories had achieved acceptable comparability during pre-project intercalibration exercises and the data were subjected to rigorous post survey review. Quality assurance samples were included in each sample batch and included method blanks, duplicates, matrix spikes, and a certified reference material. Sediment particle size was measured by light-scattering technology using either a Coulter LS230 or a Horiba LA900 instrument. The sediment samples analyzed for all metal analytes except mercury were digested in strong acid according to the procedures described in EPA Method 3050B. Metals were quantified using either inductively coupled plasma mass spectrometry, inductively coupled plasma emission spectroscopy, flame atomic absorption, or graphite furnace atomic absorption. Mercury was analyzed using cold vapor atomic absorption spectroscopy. Samples for organic chemistry analysis were solvent extracted using accelerated solvent extraction, sohxlet, or roller table. The extracts obtained were subjected to each laboratory's own clean-up procedures and were analyzed by gas chromatographic method (e.g., dual-column GC-ECD or GC-MS in the selected ion monitoring mode).

Benthic Community Analysis

A separate grab sample was taken for benthic community analysis at all the stations. The contents of the grab were washed through a 1.0-mm screen and all of the retained animals identified to species or the lowest possible taxon. Different benthic indices were used to assess community status for the San Francisco Bay and southern California stations because of habitat differences between the two regions that affected species composition. The benthic community condition of the southern California stations was assessed using the Benthic Response Index (BRI; Ranasinghe *et al.* 2003). The BRI is the abundance-weighted average pollution tolerance score of organisms occurring in a sample. The Index of Biotic Integrity (IBI) was used to determine benthic community condition for the San Francisco Bay stations (Thompson and Lowe 2004). The IBI uses a multimetric index to discriminate between impacted and reference areas.

Data Analysis

Toxicity data were control normalized ((station value/control) x 100) to facilitate comparisons among the test methods. Statistical significance was tested using Student's t-test ($p \leq 0.05$) assuming unequal variance (Zar 1999). For sublethal methods having more than one endpoint, if either or both endpoints were significantly different from control, the station was designated as toxic.

The mean ERM quotient (ERMq; Long *et al.* 1998) was calculated for each station to integrate a subset of the analyzed chemicals into a value that is predictive of toxic effects. The ERM for DDT was not used in calculations because it has been found to be unreliable (Long *et al.* 1995). Relationships between sediment chemistry parameters or benthic community condition and toxicity response were analyzed using a non-parametric Spearman's rank correlation.

RESULTS

The experimental batches for all toxicity data that is presented passed test control acceptability criteria, except for one SWI batch with *M. galloprovincialis*. That batch contained the only sample with a significant toxic response for the SWI test and had a low control normal-alive percentage. Because the difference between the control and sample response was very large, we have chosen to include the data.

There were two quality assurance issues with the *L. plumulosus* 28-day test. First, there was a test failure based on insufficient reproduction in the controls. When the test was repeated, the controls reproduced sufficiently, but all of the other samples had greatly less reproduction than the controls. This situation had never been encountered by the testing laboratory and led to our decision to not use the reproduction data for analysis. The second issue with this test was the very long holding time of the sediments before testing began compared to the other methods (Table 2). The effects of this prolonged holding time are unknown. The data are presented for the purposes of comparison, but may have differed had the holding times been identical between methods.

Water quality measurements made during testing indicated that the values were within acceptable range for the majority of sample/test combinations. For the *M. mercenaria* test, station 4130 exhibited elevated pore water ammonia (37.5 mg/L total ammonia-nitrogen). While the tolerance of *M. mercenaria* to ammonia is not known, there is correlative evidence that the ammonia level in the sample may have been the cause of toxicity. For the SWI test, station BC11 had an overlying water ammonia concentration of 0.145 mg/L un-ionized ammonia, which is very near the EC50 (approximately 0.17 mg/L, unpublished data).

Comparisons Among Sublethal Tests

There was a wide range in the percentage of stations that each of the sublethal methods identified as toxic (Figure 3). The highest percentage was for the copepod, *Amphiascus* that found 9 out of the 10 stations tested to be toxic, followed by *N. arenaceodentata* with 8 out of 15 stations. The proportion of stations identified as toxic was much lower for the remaining test methods, with the lowest percentage for the SWI testing which identified 1 out of 15 stations as toxic.

Comparisons Between Acute Tests

The *E. estuarius* method was the less sensitive of the two amphipod acute protocols, identifying 4 out of 15 stations as toxic (Figure 4). Overall the *E. estuarius* method was near the mid-point in sensitivity relative to the sublethal tests. The *L. plumulosus* 10-day method identified 9 out of the 15 sites as toxic and was more sensitive than all but one of the sublethal methods.

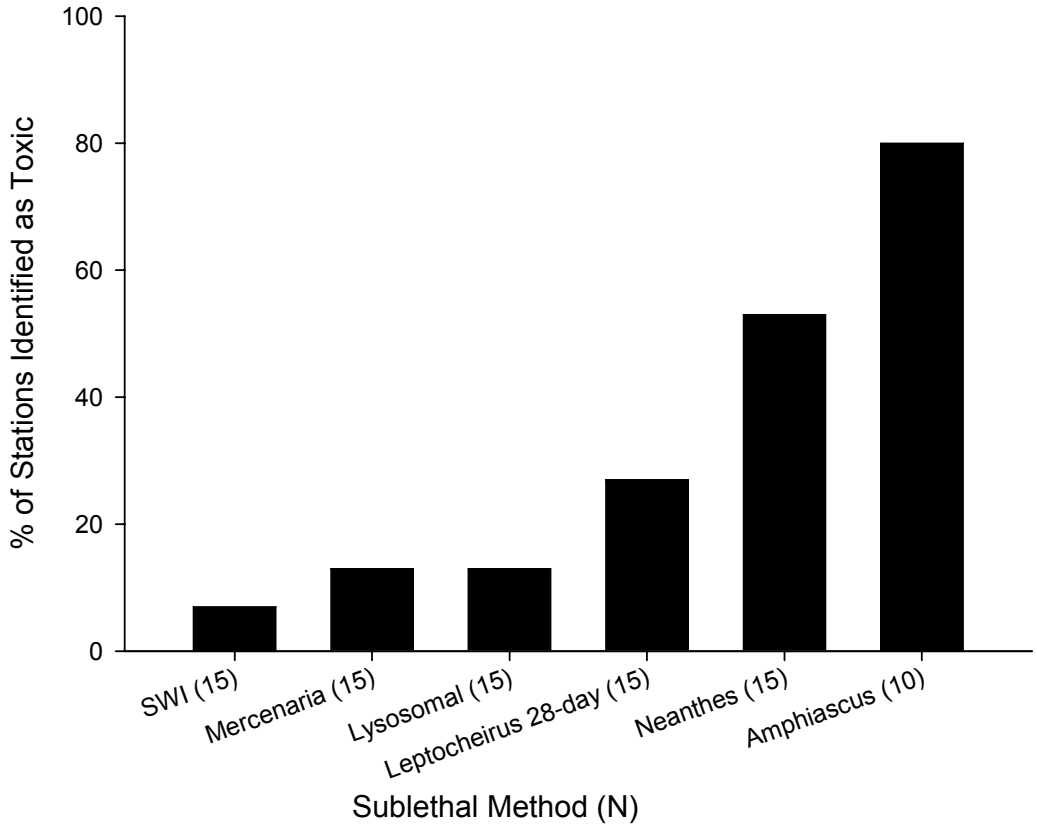

Figure 3. Percentage of stations that each sublethal method identified as being toxic. Number of samples tested is in parentheses. *Leptocheirus plumulosus* (*Leptocheirus*), Sediment-water Interface (SWI), *Merceneria mercenaria* (*Mercenaria*), *Crassostrea virginica* (Lysosomal), *Neanthes arenaceodentata (Neanthes)*, and *Amphiascus tenuiremus* (*Amphiascus*).

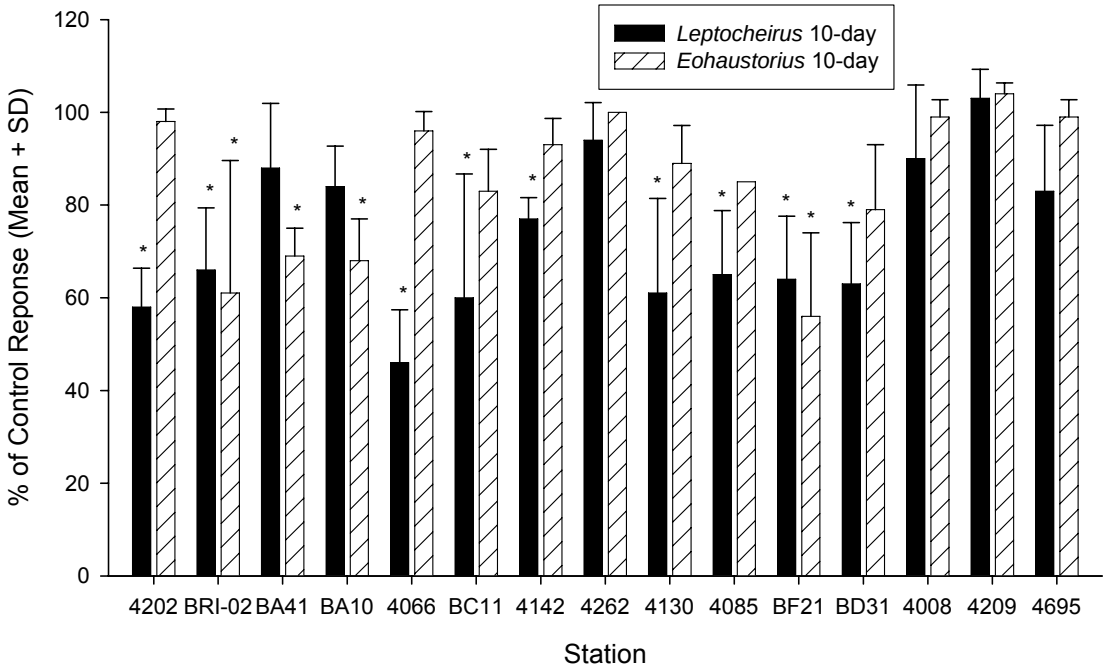

Figure 4. Results of *Eohaustorius estuarius* (*Eohaustorius*) 10-dayand *Leptocheirus plumulosus* (*Leptocheirus*) 10-day survival tests conducted as part of the Bight'03 and RMP regional monitoring programs. Stations marked with * are significantly different from control values (p <0.05) and less than 80% of the control response.

Comparisons Between Sublethal and Acute Tests

The *N. arenaceodentata* and *Amphiascus* tests detected toxicity at stations where *E. estuarius* did not at 27% and 70% of the stations, respectively; while in no cases did *E. estuarius* demonstrate toxicity where either of these two tests did not (Table 3). Alternatively, the *E. estuarius* test identified a higher percentage of stations as toxic than did the SWI, *M. mercenaria* and *C. virginica* tests. The *E. estuarius* test identified toxicity in 27% of the samples that the other tests classified as nontoxic.

Table 3. Comparative ability of acute and sublethal sediment toxicity test methods to detect toxicity in stations from southern California and San Francisco Bay. Numeric values are expressed as percentage of stations tested. The station order is based on a combined ranking of chemical contamination and benthic community health, with the most contaminated/impacted stations listed first. *Eohaustorius estuarius* (*Eohaustorius*), *Leptocheirus plumulosus* (*Lepto*), Sediment-water Interface (SWI), *Merceneria mercenaria* (*Mercenaria*), *Crassostrea virginica* (Lysosome), *Neanthes arenaceodentata (Neanthes)*, and *Amphiascus tenuiremus* (*Amphiascus*).

	Acute Methods		Sublethal methods					
Station	*Eohaustorius*	*Lepto* 10-day	SWI	*Mercenaria*	Lysosome	*Lepto*28-day	*Neanthes*	*Amphiascus*
4202	N	Y	N	N	N	Y	N	Y
BRI-2	Y	Y	N	N	N	Y	Y	Y
BA41	Y	N	N	N	N	N	Y	Y
BA10	Y	N	N	N	N	N	Y	--
4066	N	Y	N	N	Y	Y	N	--
BC11	N	Y	Y	N	Y	N	Y	Y
4142	N	Y	N	N	N	N	N	--
4262	N	N	N	N	N	N	Y	Y
4130	N	Y	N	Y	N	N	Y	Y
4085	N	Y	N	N	N	N	N	Y
BF21	Y	Y	N	N	N	N	Y	--
BD31	N	Y	N	N	N	N	N	--
4008	N	N	N	Y	N	N	Y	Y
4209	N	N	N	N	N	N	N	Y
4695	N	N	N	N	N	Y	N	N
% Sublethal Toxic, *Eohaustorius* Not Toxic	--	--	7	13	13	20	27	70
% *Eohaustorius* Toxic, Sublethal Not Toxic	--	--	27	27	27	20	0	0
% Agree Toxic	--	--	0	0	0	7	27	20
%Agree Not Toxic	--	--	67	60	60	53	47	10
	--	--						
% Sublethal Toxic, *Leptocheirus* Not Toxic	--	--	0	7	0	7	27	40
% *Leptocheirus* Toxic, Sublethal Not Toxic	--	--	53	53	47	40	33	0
% Agree Toxic	--	--	7	7	13	20	27	50
%Agree Not Toxic	--	--	40	33	40	33	13	10

Y = Station identified as toxic
N = Station not identified as toxic
-- = Station or comparison not tested

The *L. plumulosus* 10-day test found a higher percentage of toxic stations than all of the sublethal methods except for the *Amphiascus* test (Table 3). The *Amphiascus* test found four stations (40%) to be toxic that were not identified by the *L. plumulosus* acute test. There was concordance between the *L. plumulosus* 10-day test and the *Amphiascus* test for the remaining stations with both finding five stations to be toxic and one not. The *N. arenaceodentata* test found four stations to be toxic that were not identified by the *L. plumulosus* acute test. However, there were five stations that were toxic in the *L. plumulosus* acute test, but were not toxic in the *N. arenaceodentata* test. For the *M. mercenaria, C. virginica, L. plumulosus* 28-day and SWI tests there was a high percentage of stations (40% or more) that the *L. plumulosus* acute test found to be toxic that the sublethal methods did not.

Combining the data from either a lethal and sublethal test or two lethal tests provided more information regarding toxicity than conducting just one test of either kind. The greatest sensitivities (most toxic stations detected) were found with the combinations of *L. plumulosus* 10-day and *N. arenaceodentata* or *Amphiascus* methods (Table 3). Nearly as sensitive was the combination of the two acute tests (see Figure 4 where 11 out of fifteen stations were identified as toxic by one or both tests).

Chemistry

Sediment physical parameters were very wide ranging with grain sizes that were nearly 100% fines (silt + clay) to 100% sand (Table 4). TOC values ranged from 0.02% to 2.9%.

Sediment contaminant concentrations also were variable among stations (Table 4). Three stations had elevated chemistry compared to the other stations. Station 4202, on the Palos Verdes shelf, had a very high concentration of total DDTs. Station BRI-02, in Marina Del Rey, had low concentrations of organic contaminants, but substantial concentrations of copper, lead and zinc. Station 4085 contained intermediate concentrations of several metals and organics. Based on the mean ERMq calculations, all of the stations tested fell into what would be considered the low to moderate range of contaminant concentrations with all mean quotients less than 0.7 (Table 4). Five samples had mean ERMq values below 0.1, a level not expected to be toxic. The mean quotients for the remaining stations fell between 0.11 and 1.0, a range that has been found to be toxic in about half of the cases (Long *et al.* 1998).

Eohaustorius estuarius survival, both *Amphiascus* endpoints and *N. arenaceodentata* growth had significant Spearman correlations with sediment chemistry (Table 5). Correlations with various metals were present, but none with organics. All of the significant correlations were negative, indicating that as the concentration increased the endpoint decreased (e.g., decreased survival or growth). All the toxicity test methods that correlated with chemistry also had significant correlations with sediment grain size. The chemical constituents that correlated with toxicity also correlated with the grain size parameters.

Table 4. Selected chemistry data from southern California and San Francisco Bay sediment samples on which toxicity tests were performed.

Station	Arsenic (mg/kg)	Cadmium (mg/kg)	Chromium (mg/kg)	Copper (mg/kg)	Lead (mg/kg)	Mercury (mg/kg)	Nickel (mg/kg)	Silver (mg/kg)	Tin (mg/kg)	Zinc (mg/kg)
4202	8.5	6.6	136	5	30.0	0.46	29.0	1.9	NA	180
BRI-02	13.0	0.3	94	362	113.0	0.98	41.6	2.0	6.3	382
BA41	4.5	0.2	NA	30	17.4	0.34	58.2	0.1	NA	90
BA10	4.1	0.1	NA	24	11.3	0.24	46.9	0.2	NA	70
4066	1.0	0.1	7	7	4.7	0.10	4.0	0.6	0.4	22
BC11	4.0	0.3	NA	39	29.7	0.23	65.9	0.1	NA	108
4142	1.0	0.2	5	6	4.3	0.06	4.1	0.3	0.5	49
4262	4.0	0.6	46	3	38.9	0.23	21.2	0.7	NA	92
4130	7.0	0.8	49	87	61.6	0.40	25.8	0.8	3.8	248
4085	11.6	1.7	78	101	130.0	0.41	33.1	2.9	6.3	315
BF21	8.5	0.2	NA	53	16.4	0.27	88.6	0.2	NA	126
BD31	7.5	0.2	NA	51	17.8	0.24	87.8	ND	NA	126
4008	2.5	0.1	34	14	4.7	0.08	10.2	0. 7	1.6	48
4209	1.5	ND	10	3	1.4	0.02	3.0	0.2	0.5	14
4695	1.1	ND	5	1	1.2	0.02	0.9	0.2	0.3	6

Table 4. (continued)

Station	TOC %	Sand %	Silt %	Clay %	ΣPAHs µg/kg	ΣDDTs µg/kg	ΣPCBs µg/kg	Mean ERMq*	ERMq Ranking
4202	2.06	39	50	11	678	2301.3	193.9	0.68	1
BRI-02	1.99	8	74	18	76	2.2	ND	0.26	4
BA41	1.09	20	22	49	1923	0.2	2.5	0.14	7.5
BA10	2.34	44	15	36	724	0.6	2.3	0.10	10
4066	0.02	100	0	0	52	1.0	ND	0.02	12
BC11	1.80	22	22	48	740	0.6	111.3	0.34	2
4142	0.27	62	NA	NA	73	ND	ND	0.02	13
4262	1.50	56	36	8	625	49.8	66.0	0.29	3
4130	2.04	44	46	10	1206	9.9	15.7	0.17	6
4085	2.93	30	57	13	578	14.6	22.6	0.24	5
BF21	1.37	1	39	60	582	0.8	0.8	0.14	7.5
BD31	1.33	9	32	59	450	1.4	0.8	0.14	9
4008	0.67	54	40	6	12	1.3	ND	0.04	11
4209	0.04	98	2	ND	ND	ND	ND	0.01	14
4695	ND	100	ND	ND	ND	ND	ND	0.01	15

* DDT concentrations not included in ERMq calculation.

Table 5. Spearman rank correlations on selected sediment parameters and toxicity endpoints. Boxed values are significant (p ≤0.05). *Eohaustorius estuarius* (*Eohaus*), *Leptocheirus plumulosus* (*Lepto*), Sediment-water Interface (SWI), *Crassostrea virginica* (Lysosome), ERMq (effects range mean quotient).

r	Eohaus Survival	Lepto 10 Survival	SWI Mussel	Clam Growth	Lysosome	Lepto 28 Survival	Lepto 28 Growth	Worm Survival	Worm Growth	Number of Copepodites	Realized Offspring
Arsenic	-0.604	-0.239	0.274	-0.145	-0.080	-0.0502	-0.422	0.136	-0.542	-0.585	-0.806
Cadmium	-0.155	-0.401	0.264	-0.295	-0.099	-0.307	-0.295	0.132	-0.264	-0.206	-0.488
Copper	-0.786	-0.375	0.196	-0.354	-0.059	0.039	-0.293	-0.051	-0.565	-0.829	-0.952
Lead	-0.366	-0.350	0.337	-0.306	-0.025	-0.251	-0.247	0.233	-0.390	-0.482	-0.842
Mercury	-0.596	-0.406	0.476	-0.143	-0.093	-0.196	-0.351	0.059	-0.514	-0.572	-0.742
Nickel	-0.836	-0.289	-0.386	-0.382	0.136	0.222	-0.111	-0.022	-0.594	-0.866	-0.709
Silver	0.220	-0.089	0.533	0.012	-0.225	-0.373	-0.209	0.188	-0.080	-0.043	-0.455
Zinc	-0.549	-0.434	0.250	-0.301	-0.085	-0.196	-0.443	0.138	-0.476	-0.567	-0.842
TOC (%)	-0.440	-0.250	0.119	-0.268	0.070	-0.043	-0.181	-0.012	-0.424	-0.390	-0.661
Sand (%)	0.820	0.237	0.091	0.349	0.081	-0.120	0.288	-0.069	0.653	0.933	0.794
Clay (%)	-0.823	-0.229	-0.320	-0.326	0.139	0.228	-0.116	-0.032	-0.596	-0.881	-0.717
∑PAHs	-0.491	-0.259	0.032	-0.354	0.222	0.104	0.181	-0.314	-0.490	-0.520	-0.486
∑DDTs	-0.013	-0.333	0.123	-0.264	0.014	-0.201	-0.100	0.382	-0.320	-0.086	-0.365
∑PCBs	-0.124	-0.295	-0.078	-0.192	0.339	-0.052	0.043	0.062	-0.211	-0.066	-0.125
Mean ERMq	-0.288	-0.268	0.018	-0.402	0.124	-0.221	-0.150	0.306	-0.449	-0.329	-0.370

Benthic Community

A range of benthic community condition was present among the stations. Most stations were classified as being in reference condition (8/15) or having an intermediate level of disturbance (5/15 stations at Level 2 or 3). Two stations (4066 and 4142) had Level 4 designations (Table 6), which indicated severe effects to the benthic community. The variations in benthic community condition did not correspond with the sediment contamination gradient. The average mean ERMq of all stations in each benthic condition category was lowest for the Level 4 stations and highest for the Level 2 stations (Table 6).

There was little correspondence between changes in benthic community condition and toxicity for most of the test methods. *L. plumulosus* 10-day survival was the only test that consistently detected toxicity at the Level 4 stations (Table 6). Most of the stations that did show toxicity were in the Reference or Level 2 categories for benthic community condition. Four of the test methods (*E. estuarius*, *L. plumulosus* 10-day and 28-day and *Amphiascus*) showed an increased incidence of toxicity among all impacted stations (Levels 2 through 4 combined) compared to stations classified as having a reference benthic condition. Correlations of BRI values for the southern California stations showed that only the *L. plumulosus* 10-day test method had a significant correlation with benthic community condition (Table 6). The correlation coefficients were negative for all but the *C. virginica* lysosome method, indicating that as the BRI value increased the toxicity endpoint value decreased (i.e., survival or growth decreased).

Table 6. Incidence of toxicity within benthic index categories and Spearman's Rank Correlation values for toxicity test endpoints. Boxed values are statistically significant ($p \leq 0.05$).

	Benthic Index Category					
Test	Ref[1]	Level 2[2]	Level 3[3]	Level 4[4]	Levels 2-4	r[5]
Number of Stations	8	4	1	2	7	
Benthic Station Rank	11.5	5.5	3.0	1.5		
Mean ERMq	0.15	0.31	0.10	0.02	0.20	
	Incidence of Toxicity (%)					
Eohaustorius estuarius 10-day Survival	12	50	100	0	42	-0.52
Leptocheirus plumulosus 10-day Survival	50	75	0	100	71	**-0.64**
Mytilus galloprovincialis Sediment-water Interface	12	0	0	0	0	-0.27
Mercenaria mercenaria Growth	12	25	0	0	14	-0.20
Crassostrea virginica Lysosome	12	0	0	50	14	0.04
L. plumulosus 28-day Growth	12	50	0	50	50	-0.25
Neanthes arenaceodentata Growth	50	75	100	0	57	-0.12
Amphiascus No. Copepodites	83	100	na	na	100	-0.44

[1] Reference stations: BC11, 4262, 4085, BF21, BD31, 4008, 4209, 4695
[2] Level 2 (Loss of biodiversity): 4202, BRI-02, BA41, 4130
[3] Level 3 (Loss of community function): BA10
[4] Level 4 (Defaunation): 4066, 4142
[5] Correlation calculated using southern California data only

Ranking of Stations to Reflect Sediment Condition

Since most of the stations in this study had not been previously sampled, there was not a known gradient of expected sediment condition. To put the data into this context, the stations were ranked by a combination of chemical contamination and benthic community health. To achieve this the stations were ranked by their mean ERMq values (Table 4). The stations were also ranked similarly by the benthic community analysis results (Table 6). These two rankings were then summed and the stations re-ranked to get the combined effect. The data presented in Figure 5 have the stations with the lowest rankings (highest chemistry and most degraded benthos) on the left and highest rankings on the right. Station 4202 had the highest concentrations of the most chemical constituents and showed a toxic response to two of the sublethal test endpoints. It ranked as having the worst sediment condition of all the stations even without the high value of DDT taken into consideration. Station BRI-2 with high concentrations of three metals and with a Level 2 benthic designation ranked as the second worst. Station 4085 with moderate levels of several chemicals ranked in the middle. Although stations 4066 and 4142 had Level 4 benthic designations, they fell in the middle of the ranks because their chemical concentrations were lower.

DISCUSSION

The sensitivity of the toxicity methods were variable within the two broad categories of tests evaluated, indicating that general classifications of tests as either acute or sublethal do not reliably indicate their relative sensitivity. For example, the most sensitive test in this study was the sublethal *Amphiascus* life cycle method, but the acute *L. plumulosus* survival test was more sensitive that any of the other sublethal tests compared. This variation in sensitivity between acute and sublethal tests is consistent with other studies, suggesting that the relative sensitivity of acute and sublethal tests to whole sediment samples varies according to the combination of tests and sample types evaluated. Comparative studies using the *L. plumulosus* 28-day test have shown that the sublethal endpoints from this test are not consistently more sensitive than acute amphipod tests to field and spiked sediments (DeWitt *et al.* 1997). Another study found that the acute *A. abdita* test was more sensitive than the *L. plumulosus* 28-day test, which was more sensitive than the *N. arenaceodentata* 28-day test (Kennedy *et al.* 2004). In contrast to the results of the present study, the *M. mercenaria* test was found to be more sensitive than the acute *A. abdita* survival test when sediment samples from the Carolinian Province were tested (Ringwood *et al.* 1996).

Our finding that *Amphiascus* was the most sensitive method overall is consistent with other studies indicating the high sensitivity of this life cycle test. Tests using sediments from Biscayne Bay in Florida by Long *et al.* (Long *et al.* 1999) found a greater incidence of toxicity with the *Amphiascus* life cycle method (73%) than with the *A. abdita* 10-day survival test (7%). The high sensitivity, chronic exposure and multiple endpoints that are characteristic of this test are desirable qualities, however, more investigation is needed to determine whether the high level of response of the test to southern California samples having low contaminant concentrations and reference benthic community condition reflect chemical toxicity or the effects of potentially confounding factors such as ammonia or organic carbon.

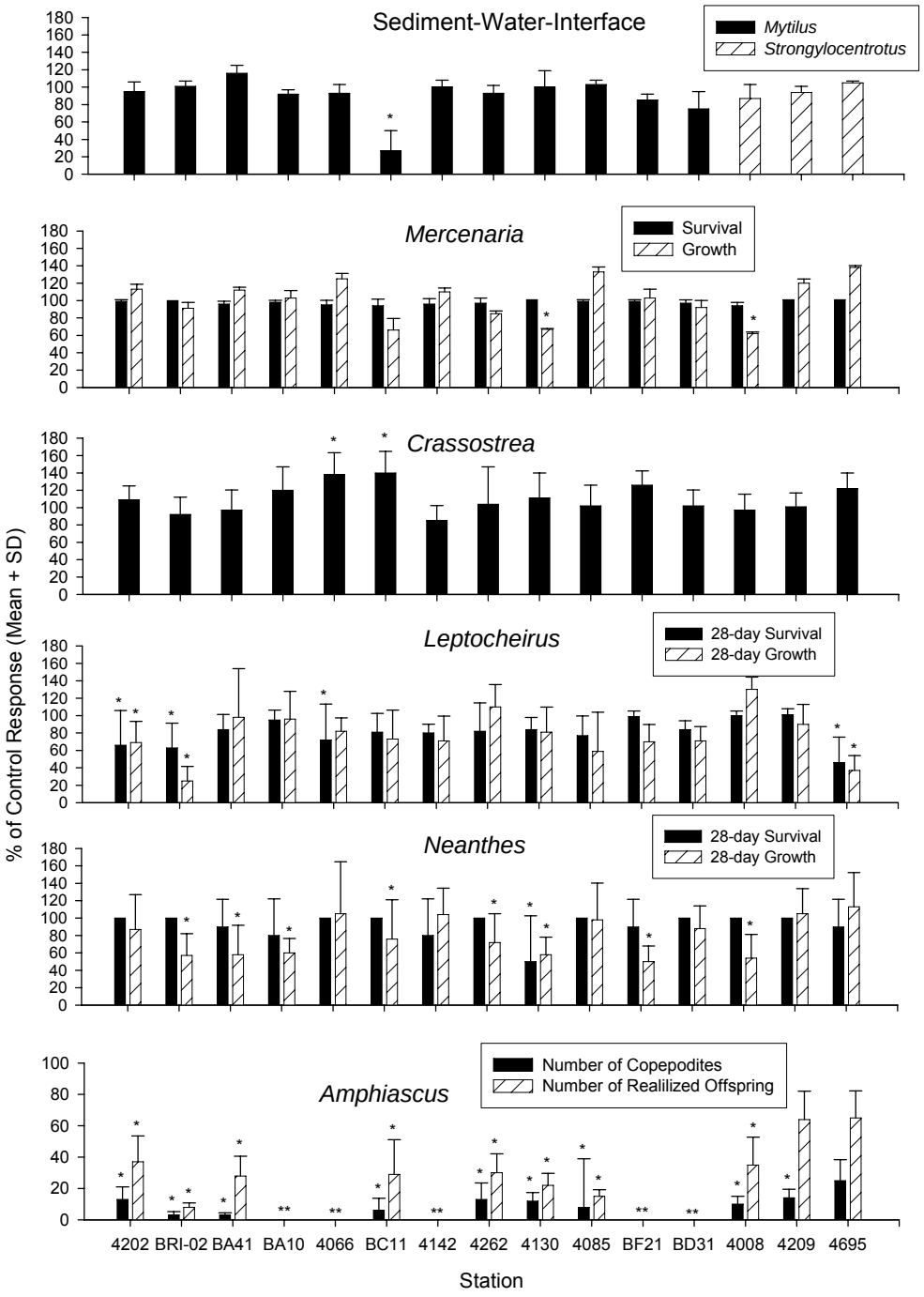

Figure 5. Results of sublethal test methods conducted on samples from southern California and San Francisco Bay. Stations marked with * are significantly different from control values (p <0.05). Stations with ** indicate that the station was not tested using that method. *Eohaustorius estuarius* (*Eohaustorius*), *Leptocheirus plumulosus* (*Leptocheirus*), *Merceneria mercenaria* (*Mercenaria*), *Crassostrea virginica* (Crassostrea), *Neanthes arenaceodentata (Neanthes)*, *Mytilus galloprovincialis* (*Mytilus*), *Strongylocentrotus pupuratus* (*Strongylocentrotus*) and *Amphiascus tenuiremus* (*Amphiascus*).

Several factors may have accounted for the variation in sensitivity among methods observed in this study, including: mode of exposure, species-specific sensitivity to contaminants, and the influence of confounding factors. The mode of exposure varied greatly among tests and those tests with the longest exposure duration and most direct contact with the sediment (i.e., *Amphiascus* and *N. arenaceodentata*) tended to be most sensitive. For the SWI method, which was least sensitive, the organisms are in the water column directly above the sediment, and are exposed for a relatively short period of time to only those contaminants diffusing into the overlying water. These differences in exposure method and sample response can be used to advantage to investigate the mode of contaminant exposure or identify the cause of toxicity.

Differences in contaminant sensitivity among test methods have been documented for some of the test species and may have influenced the results of this study. Several studies have been conducted that compared the *L. plumulosus* 10-day and 28-day tests and the *N. arenaceodentata* 28-day test to various chemicals and found varying patterns of response. The *N. arenaceodentata* test was more sensitive than *L. plumulosus* to sediments contaminated with metals or the explosive TNT, both of these sublethal tests were more sensitive than the acute *L. plumulosus* test to PCBs, yet the *L. plumulosus* acute method was more sensitive to PAH contaminated sediments than *N. arenaceodentata* (Farrar *et al*. 2005, Green *et al*. 1999). Comparisons among acute tests using *A. abdita*, *E. estuarius* and *R. abronius* showed that *E. estuarius* was the most sensitive to DDT, while *A. abdita* and *R. abronius* were more sensitive to cadmium (Weston 1996). Sediment contaminant mixtures varied among the stations in the present study, with differences of up to two orders of magnitude in metals, PCB, and PAH concentrations, and up to three orders of magnitude in DDT. These differences may have contributed to the variation in response among the test methods.

Variations in holding time or sediment handling that occurred among the laboratories are potential confounding factors that may have altered the toxicity of the samples through changes in bioavailability or chemical composition. The nature and magnitude of such effects was not determined in this study, but an analysis of the data indicates that the patterns of relative sensitivity observed among the test methods were independent of holding time. For example, holding times were shortest and similar for the SWI and *Amphiascus* methods, yet these two tests had very different patterns of response to the samples (Table 7). The patterns of relative response among the tests were also similar for the two batches of whole sediment tested (e.g., *Amphiascus* most sensitive, *M. mercenaria* and *C. virginica* usually least sensitive), indicating that variations in holding time or sediment handling among the tests and batches were not major confounding factors.

The most responsive of the acute and sublethal toxicity tests showed a general correspondence with the gradient of sediment condition described by a combination of the chemistry and benthic community data. The *Amphiascus* and *N. arenaceodentata* tests reflected the expected pattern of decreasing toxicity with improving sediment condition (Figure 5), as did both of the acute tests (Figure 4). These relationships were inconsistent for stations having intermediate rankings of sediment conditions, indicating

substantial uncertainty in the relationships among the different indicators of sediment quality. In addition to the sources of variability mentioned previously for the toxicity tests, measures of sediment chemistry and benthic community condition also have inherent uncertainty and sources of error that may have accounted for the inconsistent relationships.

Significant correlations with chemistry concentrations were found in the present study for the *E. estuarius* survival, *Amphiascus* reproduction and *N. arenaceodentata* growth tests. Similar relationships have also been documented in many other studies for a variety of test organisms and form the basis for empirical sediment quality guidelines (Long *et al.* 1995, Fairey *et al.* 2001). There were also significant correlations with grain size for each test. The chemistry values also correlated with grain size and many of the chemical constituents also correlated with one another. These intercorrelations make determining whether toxicity is associated with chemistry or the confounding factor of grain size a difficult matter. Grain size is not known to be a confounding factor for *E. estuarius* (USEPA 1994). Grain size should not have been an issue for *Amphiascus* since all samples were sieved to remove large particles and optimize the sediments for the animals. *Neanthes arenaceodentata* have been tested in grain sizes ranging from 5 to 100% sand with no effects on either survival or growth (Dillon *et al.* 1993). These factors indicate that there is a likelihood of an association between sediment contamination and toxicity for these three methods in the current study, rather than a grain size effect.

The lack of correlations with sediment chemistry for some of the test methods may have several causes. There was little observed toxicity for many of the tests making the detection of correlations difficult. In addition, no measure of bioavailability of chemical constituents was made for the sediments, adding uncertainty regarding the actual chemical dose received by the test animals. Sediment chemistry analyses do not quantify all possible toxicants, so it is possible that unmeasured chemical constituents or interactions between compounds may have caused the observed toxicity. Another potential source of uncertainty is toxicity from confounding factors such as ammonia or sulfides. While the sensitivity of some of the test methods to these factors is poorly known, water quality data from the tests showed that dissolved ammonia concentrations were low and below concentrations of concern for most of the samples, indicating that these factors did not have a significant influence on the results.

A strong relationship between the toxicity results and benthic community condition was not found in this study, suggesting that these indicators were responding to different aspects of sediment quality. Other studies have reported similar results. Analyses of Chesapeake Bay sediment toxicity using the *L. plumulosus* 10-day and 28-day tests found a similar lack of correspondence with benthic community response (McGee *et al.* 2004). A statistically significant correlation between *E. estuarius* mortality and benthic community impact was found for southern California embayment sediments, but the relationship accounted for only 10% of the variation in community condition (Ranasinghe *et al.* 2003). Toxicity tests differ from the *in situ* benthic environment in many aspects, such as the exposure duration, species type, and laboratory handling of the

sediment. These factors can affect contaminant bioavailability or the sensitivity of the response and may have accounted for the relatively high frequency of toxicity detected in samples containing an unimpacted benthic community. It is not possible for toxicity tests to perfectly replicate environmental exposure conditions and provide a substitute for assessment of biological effects on resident organisms; these tests are intended to provide a measure of potential contaminant effects that is complementary to chemical and biological measures.

The effects of noncontaminant factors on the benthic community analyses may have also influenced the correlation analyses with toxicity. Changes in benthic community condition did not correspond with increasing contamination levels, as represented by the mean ERMq (Table 6). This finding contrasts with studies in other regions of the United States that have shown an increase in the incidence of degraded benthos within the mean ERMq range present among the southern California samples (Hyland *et al.* 2003). It is possible that variations in noncontaminant factors related to the diversity of habitats and sediment types included in this study may have influenced the benthic community results and confounded the ability of to discern impacts due to toxicity.

This study and others have shown marked differences in sensitivity among toxicity tests that cannot be easily predicted on the basis of biological endpoint and mode of exposure. This diversity presents both a challenge and opportunity for sediment toxicity evaluation. The challenge lies in selecting the most appropriate tests for use in a particular study. Variations in relative sensitivity related to contaminant type and uncertainties in the interpretation of chemistry and benthic community data suggest that the use of just a single test method, selected on the basis of high sensitivity to a subset of samples, is unlikely to provide a complete or confident assessment of toxicity. Data from multiple toxicity tests that represent a diversity of species, endpoints, and exposure modes, in addition to sediment chemistry and benthic community analyses, are needed to assess sediment quality to the level of confidence needed to support management decisions (Chapman and Anderson 2005). The use of a diverse suite of toxicity tests also provides an opportunity to improve our understanding of the causes of sediment toxicity, as differences in the patterns or symptoms of response between tests can be used to help identify the case of toxicity (USEPA 1993).

LITERATURE CITED

Adams, W.J., A.S. Green, W. Ahlf, S.S. Brown, G.A. Burton Jr., D.B. Chadwick, M. Crane, R. Gouguet, K.T. Ho, C. Hogstrand, T.B. Reynoldson, A.H. Ringwood, J.D. Savitz and P.K. Sibley. 2005. Using sediment assessment tools and a weight-of-evidence approach. pp. 163-225 *in*: R.J. Wenning, G.E. Batley, C. Ingersoll and D.W. Moore (eds.), Use of Sediment Quality Guidelines and Related Tools for the Assessment of Contaminated Sediments. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

Anderson, B.S., J.W. Hunt, M. Hester and B.M. Phillips. 1996. Assessment of sediment toxicity at the sediment-water interface. pp. 609-624 *in*: G.K. Ostrander (ed.), Techniques in aquatic toxicology. CRC Press Inc. Boca Raton, FL.

Anderson, B.S., J.W. Hunt, B.M. Phillips, S. Tudor, R. Fairey, J. Newman, H.M. Puckett, M. Stephenson, E.R. Long and R.S. Tjeerdema. 1998. Comparison of marine sediment toxicity test protocols for the amphipod *Rhepoxynius abronius* and the polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environmental Toxicology and Chemistry* 17:859-866.

Bay, Steven, Darrin Greenstein and Diana Young. 2007. Evaluation of methods for measuring sediment toxicity in California bays and estuaries. Technical Report 503. Southern California Coastal Water Research Project. Costa Mesa, CA.

Bay, S.M., T. Mikel, K. Schiff, S. Mathison, B. Hester, D. Young and D. Greenstein. 2005. Southern California Bight 2003 regional monitoring program: I. Sediment toxicity. Southern California Coastal Water Research Project. Westminster, CA.

Bight'03 Coastal Ecology Committee. 2003. Quality assurance manual. Southern California Coastal Water Research Project. Westminster, CA.

Bridges, T.S. and J.D. Farrar. 1997. The influence of worm age, duration of exposure and endpoint selection on bioassay sensitivity for *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environmental Toxicology and Chemistry* 16:1650-1658.

Bridges, T.S., J.D. Farrar and B.M. Duke. 1997. The influence of food ration on sediment toxicity in *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environmental Toxicology and Chemistry* 16:1659-1665.

Carr, R.S. and M. Nipper. 2003. Porewater toxicity testing: Biological, chemical, and ecological considerations. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

Chandler, G.T. and A.S. Green. 1996. A 14-day harpacticoid copepod reproduction bioassay for laboratory and field contaminated muddy sediments. pp. 23-39 *in*: G.K. Ostrander (ed.), Techniques in aquatic toxicology. CRC Press. Boca Raton, FL.

Chapman, P.M. and J. Anderson. 2005. A decision-making framework for sediment contamination. *Integrated Environmental Assessment and Management* 1:163-173.

DeWitt, T.H., M.R. Pinza, L.A. Niewolny, V.I. Cullinan and B.D. Gruendell. 1997. Development and evaluation of standard marine/estuarine chronic sediment toxicity test method using *Leptocheirus plumulosus*. Battelle Marine Sciences Laboratory. Sequim, WA.

Dillon, T.M., D.W. Moore and A.B. Gibson. 1993. Development of a chronic sublethal bioassay for evaluating contaminated sediment with the marine polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environmental Toxicology and Chemistry* 12:589-605.

Fairey, R., E.R. Long, C.A. Roberts, B.S. Anderson, B.M. Phillips, J.W. Hunt, H.R. Puckett and C.J. Wilson. 2001. An evaluation of methods for calculating mean sediment quality guideline quotients as indicators of contamination and acute toxicity to amphipods by chemical mixtures. *Environmental Toxicology and Chemistry* 20:2276-2286.

Farrar, J.D., A. Kennedy, G. Lotufo and C. McNemar. 2005. Comparative evaluation of the efficacy of acute and chronic sublethal sediment tests for assessing sediment quality. SETAC North America 26th Annual Meeting. Baltimore, MD.

Green, A.S., D. Moore and D. Farrar. 1999. Chronic toxicity of 2,4,6-trinitrotoluene to a marine polychaete and an estuarine amphipod. *Environmental Toxicology and Chemistry* 18:1783-1790.

Hyland, J.L., W.L. Balthis, V.D. Engle, E.R. Long, J.F. Paul, J.K. Summers and R.F. Van Dolah. 2003. Incidence of stress in benthic communities along the U.S. Atlantic and Gulf of Mexico coasts within different ranges of sediment contaminations from chemical mixtures. *Environmental Monitoring and Assessment* 81:149-161.

Kennedy, A., J.D. Farrar, J.A. Stevens and M. Reiss. 2004. Evaluation of the applicability of standard toxicity test methods to dredged material management. Society of Environmental Toxicology and Chemistry Annual Meeting. Portland, OR.

Keppler, C.J. and A.H. Ringwood. 2002. Effects of metal exposures on juvenile clams, *Mercenaria mercenaria*. *Bulletin of Environmental Contamination and Toxicology* 68:43-48.

Lamberson, 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, Inc. Boca Raton, FL.

Long, E.R., M.F. Buchman, S.M. Bay, R.J. Breteler, R.S. Carr, P.M. Chapman, J.E. Hose, A.L. Lissner, J. Scott and D.A. Wolfe. 1990. Comparative evaluation of five

toxicity tests with sediments from San Francisco Bay and Tomales Bay, California. *Environmental Toxicology and Chemistry* 9:1193-1214.

Long, E.R., L.J. Field and D.D. MacDonald. 1998. Predicting toxicity in marine sediments with numerical sediment quality guidelines. *Environmental Toxicology and Chemistry* 17:714-727.

Long, E.R., D.D. MacDonald, S.L. Smith and F.D. Calder. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environmental Management* 19:81-97.

Long, E.R., G.M. Sloane, G.I. Scott, B. Thompson, R.S. Carr, J. Biedenback, T.L. Wade, B.J. Presley, K.J. Scott, C. Mueller, G. Brecken-Fols, B. Albrecht, J.W. Anderson and G.T. Chandler. 1999. Magnitude and extent of chemical contamination and toxicity in sediments of Biscayne Bay and vicinity. NOS NCCOS CCMA 141. National Oceanic and Atmospheric Administration. Silver Spring, MD.

McGee, B.L., D.J. Fisher, D.A. Wright, L.T. Yonkos, G.P. Ziegler, S.D. Turley, J.D. Farrar, D.W. Moore and T.S. Bridges. 2004. A field test and comparison of acute and chronic sediment toxicity tests with the estuarine amphipod *Leptocheirus plumulosus* in Chesapeake Bay, USA. *Environmental Toxicology and Chemistry* 23:1751-1761.

Pinza, M.R., J.A. Ward and N.P. Kohn. 2002. Results of interspecies toxicity comparison testing associated with contaminated sediment management. Sediment Management Annual Review Meeting. Seattle, WA.

Puget Sound Water Quality Authority PSWQA. 1995. Recommended guidelines for conducting laboratory bioassays on Puget Sound sediments. Puget Sound Water Quality Authority for U.S. Environmental Protection Agency Region 10. Olympia, WA.

Ranasinghe, J.A., D.E. Montagne, R.W. Smith, T.K. Mikel, S.B. Weisberg, D.B. Cadien, R.G. Velarde and A. Dalkey. 2003. Southern California Bight 1998 Regional Monitoring Program: VII. Benthic Macrofauna. Southern California Coastal Water Research Project. Westminster, CA.

Ringwood, A.H., D.E. Conners and J. Hoguet. 1998. Effects of natural and anthropogenic stressors on lysosomal destabilization in oysters *Crassostrea virginica*. *Marine Ecology Progress Series* 166:163-171.

Ringwood, A.H., D.E. Conners, J. Hoguet and L.A. Ringwood. 2005. Lysosomal destabilization assays in estuarine organisms. pp. 287-300 *in*: G.K. Ostrander (ed.), Techniques in Aquatic Toxicology, Volume 2. CRC Press, Taylor and Francis. Boca Raton, FL.

Ringwood, A.H., A.F. Holland, R.T. Kneib and P.E. Ross. 1996. EMAP/NS&T pilot studies in the Carolinian Province: Indicator testing and evaluation in the Southeastern

estuaries. NOS ORCA 102. National Atmospheric and Oceanic Administration. Silver Springs, MD.

Ringwood, A.H. and C.J. Keppler. 1998. Seed clam growth: An alternative sediment bioassay developed during EMAP in the Carolinian Province. *Environmental Monitoring and Assessment* 51:247-257.

San Francisco Estuary Institute SFEI. 2005. 2003 annual monitoring results. The San Francisco Estuary regional monitoring program for trace substances (RMP). San Francisco Estuary Institute. Oakland, CA.

Strobel, C.J., D.J. Klemm, L.B. Lobring, J.W. Eichelberger, A. Alford-Stevens, B.B. Potter, R.F. Thomas, J.M. Lazorchak, G.B. Collins and R.L. Graves. 1995. Environmental monitoring and assessment program (EMAP) laboratory methods manual estuaries. Volume 1- Biological and physical analyses. EPA/620/R-95/008. U.S. Environmental Protection Agency, Office of Research and Development. Narragansett, RI.

Thompson, B. and S. Lowe. 2004. Assessment of macrobenthos response to sediment contamination in the San Francisco Bay estuary, California, USA. *Environmental Toxicology and Chemistry* 23:2178-2187.

U.S. Environmental Protection Agency (USEPA). 1993. Methods for Aquatic Toxicity Identification Evaluations: Phase III Toxicity Confirmation Procedures for Samples Exhibiting Acute and Chronic Toxicity. EPA/600/R-92/081. U. S. Environmental Protection Agency. Duluth, MN.

USEPA. 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA/600/R-94/025. Office of Research and Development, U.S. Environmental Protection Agency. Narragansett, RI.

USEPA. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. Office of Research and Development. Cincinnati, OH.

USEPA. 2001. Methods for assessing the chronic toxicity of marine and estuarine sediment-associated contaminants with the amphipod *Leptocheirus plumulosus*. U.S. Environmental Protection Agency. Washington, D.C.

Weston, D.P. 1996. Further development of a chronic *Ampelisca abdita* bioassay as an indicator of sediment toxicity. RMP Contribution #17. San Francisco Estuary Institute. Richmond, CA.

Zar, J.H. 1999. Biostatistical analysis (4th ed.). Simon & Schuster. Upper Saddle River, NJ.

APPENDIX B

Interlaboratory Comparison of Sublethal Sediment Toxicity Test Methods Using *Mercenaria mercenaria* and *Mytilus galloprovincialis*

Steven Bay, Diana Young and Darrin Greenstein

Southern California Coastal Water Research Project, Westminster, CA

INTRODUCTION

Many sediment quality monitoring and assessment programs use a combination of acute amphipod survival and sublethal sediment toxicity test methods. The acute amphipod methods are usually conducted using standard protocols for a small number of species (USEPA 1994) and several studies have been conducted that document important aspects of the tests such as relative sensitivity and interlaboratory variability. A greater diversity of sublethal sediment toxicity test methods have been applied in various studies (Lamberson *et al.* 1992), yet few studies have been conducted that compare the relative performance of these methods.

A significant data gap for some sublethal toxicity tests is information on interlaboratory variability. An understanding of the amount of variation associated with conducting the test in different laboratories is needed to assist in decisions regarding the selection of test methods for use in a study and for determining the significance of various ranges in the organism's response to the test samples. Interlaboratory variability data are not available for two sublethal methods that are promising candidates for use in regional monitoring programs: the seven-day growth test using the seed clam, *Mercenaria mercenaria* (Ringwood and Keppler 1998), and the two-day sediment-water interface (SWI) test using embryos of the mussel *Mytilus galloprovincialis* (Anderson *et al.*, 1996). Interlaboratory variability for these two test methods is needed to support the evaluation of these methods for use in sediment testing programs.

The objective of this study was to measure the interlaboratory variability associated with the seed clam and mussel embryo sediment toxicity tests. Interlaboratory comparison tests were conducted with both test methods using field and spiked sediments.

METHODS

Concurrent sediment toxicity tests were conducted by two laboratories for the mussel embryo test and by three laboratories for the seed clam test. Two types of samples were tested in each set of experiments: dilutions a contaminated field sediment and several concentrations of sediment spiked with nonylphenol. In both cases, one of the participating laboratories was the originator of the test method. For both test methods, each laboratory also conducted reference toxicant exposures to demonstrate laboratory comparability. Additionally, range finding tests were conducted to determine the proper concentrations of the spiked and diluted samples.

Field

The sediment used for spiking with nonylphenol and for dilution of contaminated field sediment was collected by Orange County Sanitation District near their reference site 18. This station is located offshore and has low levels of chemical contamination and a moderate grain size (~50% sand). The contaminated field sediment was from Consolidated Slip (CS) in Los Angeles Harbor and had been in storage since collection in October 2002. Consolidated Slip has a long history of contamination from industrial sources with very high levels of PAHs, DDT and metals and very fine grain size. Both the sediment from CS and Orange County (OC) were stored in plastic containers at 5 ºC.

Test Sediment Preparation

Stock solutions of 4-n-nonylphenol (Alfa Aesar) in acetone were placed into 2 L glass jars and the carrier was allowed to volatilize on a Wheaton roller apparatus (Distworth *et al*., 1990). After volatilization, OC sediment was added to the containers in amounts in amounts corresponding to nominal nonylphenol concentrations of 0.1-1000 mg/kg and rolled for the first 24 hours of the seven-day equilibration time. Sediment was stored at 5ºC in amber glass jars for the remainder of the equilibration period. Chemical verification of the final sediment concentrations was not preformed.

The CS dilutions were made as 10, 25, and 50 percent wet weight:wet weight CS sediment diluted with OC sediment. Mixing was accomplished with a polycarbonate spoon in a large polycarbonate bowl. A control sample consisting of 100% OC sediment was also tested. Aliquots of the mixtures were placed into separate containers for each laboratory. The samples were then stored in plastic containers at 4ºC and allowed to equilibrate for seven days, before being used in the interlaboratory experiments.

Mussel Embryo Development Test

The University of California Davis Marine Pollution Studies Laboratory (MPSL) and Southern California Coastal Water Research Project (SCCWRP) conducted the laboratory intercalibration for the sediment-water interface (SWI) mussel embryo development test. The mussels (*M. galloprovincialis*), obtained from Carlsbad Aquafarms in Carlsbad, CA, were acclimated in 32 g/kg seawater at 15ºC overnight. The procedure for the mussel development test and the exposure procedures followed methods described in Appendix A. To simulate a core sample, the core tubes were filled with 5 cm of the sediment samples, with five replicate tubes per treatment. Seawater was

added over the sediments, aeration was added and the system was allowed to equilibrate overnight.

Both laboratories also performed a 48-hour water only reference toxicant experiment with copper. A stock solution of $CuCl_2$ was provided by SCCWRP. Each laboratory prepared dilutions of the stock to achieve concentrations of 4.5, 6.5, 9.5, 13.9, 20.4, and 30.0 µg/L copper plus a water only seawater control. Four replicates of each concentration were tested.

At the end of the experiment all normal and abnormal embryos were counted. The %Normal-Alive endpoint was calculated by dividing the number of normal embryos in each vial by the mean initial embryo count and then multiplying by 100.

Juvenile Clam Growth Test

Three laboratories participated in the seed clam (*M. mercenaria*) interlaboratory calibration experiment: South Carolina Marine Resources Research Institute (MRRI), SCCWRP, and Weston Solutions (Carlsbad, CA). Exposure methods followed those described in Appendix A. The clams were fed the algae *Isochrysis galbana* during all exposures. For the interlaboratory experiment all laboratories used live *I. galbana* cultures. However, during the range finding tests, a concentrated *I. galbana* solution obtained from Reed Mariculture was used for feeding after proper dilution.

All laboratories performed a water only 7-day reference toxicant test exposure to copper with the same feeding regime as for the sediment experiment. The reference toxicant experiment used a 10,000 µg/L stock solution of $CuCl_2$ provided by SCCWRP. Dilutions were prepared at each of the laboratories to achieve concentrations of 6.25, 12.5, 25, 50, and 100 µg/L copper.

Although OC sediment was used as a control, it had never been previously tested using the juvenile clams. Therefore, a second control was included that has historically been used as a reference for this clam test, to ensure reasonable control response. MRRI provided this reference sediment (coded LTH), which was sandy sediment from a clean site in South Carolina.

Data analysis

Data for all tests were adjusted to control response within each laboratory. For the SWI test, the data was adjusted to the water only control value from the reference toxicant test. For the *M. mercenaria* test, the data was adjusted to the response in the LTH sample. Significant differences between controls and treatments were calculated by t-tests assuming unequal variance ($p \leq 0.05$). Differences between laboratories were calculated with either t-tests (SWI) assuming unequal variance or ANOVAs (*M. mercenaria*) followed by Tukey's multiple comparison test. EC50s for reference toxicant exposure for the mussel embryos were calculated using probit analysis. For the clam reference toxicant exposure, the IC50 (the inhibition concentration where a 50% reduction in growth is predicted to occur) was calculated using the EPA ICP program.

RESULTS

Mussel Embryo Development

Range Finding

For the SWI mussel embryo development test, range finding experiments using nonylphenol and CS samples were completed at SCCWRP. An initial selected series of 10, 100, and 1000 mg/L nonylphenol produced a dose response with 87% of control normal-alive embryos in the 10mg/L nonylphenol, 80% in the 100 mg/L nonylphenol, and 21% in the 1000 mg/L nonylphenol sample. Because this was a suitable dose-response, these concentrations were selected for use in the intercalibration exercise.

The CS dilutions were tested at 5, 10, and 25% of CS sediment. The percent of control normal-alive embryos at 5 and 10% CS was 99%, and at 25% CS was 79%. In order to increase the range of response, the percentage of CS in the samples was increased to 10, 25 and 50% for the intercalibration exercise.

Interlaboratory Calibration

MPSL results showed a significant difference between all three concentrations of the nonylphenol spiked sediments and the non-spiked OC control station. MPSL obtained a good dose response, with each concentration showing substantially more toxicity than the previous one and severe toxicity at 1000 mg/L nonylphenol, with no normally developed embryos (Figure 1). SCCWRP found only the 1000 mg/L nonylphenol sample significantly different from the control with 0% of the embryos developed normally (Figure 1). SCCWRP found development in the other two nonylphenol concentrations was similar to the OC sediment.

MPSL found the highest two concentrations of the CS sediment to be significantly different from the OC station. However, the toxicity in the dilution series of 10, 25, and 50% CS was of moderate degree with 77, 70, and 57% normal-alive relative to the control, respectively (Figure 2). SCCWRP did not find a dose response for CS dilution sediments and did not find any of the dilutions to be significantly different from the OC station. The two higher concentrations of CS had normal development only slight less than that found in the water only controls.

There was little agreement between the two laboratories' results. Of the seven samples tested only two, OC and 10% CS, were not significantly different between the laboratories. The five other samples were significantly different from each other, and in all cases the MPSL %normal-alive results were lower than those of SCCWRP.

Reference Toxicant

The EC50s for the two laboratories were comparable with MPSL being 6.8 µg/L copper (upper and lower 95% confidence limits were 6.5 and 6.9) and SCCWRP 7.6 µg/L copper (lower and upper 95% confidence limits were 7.2 and 8.0). The dose-response plots of the copper exposure were remarkably similar between the two laboratories (Figure 3).

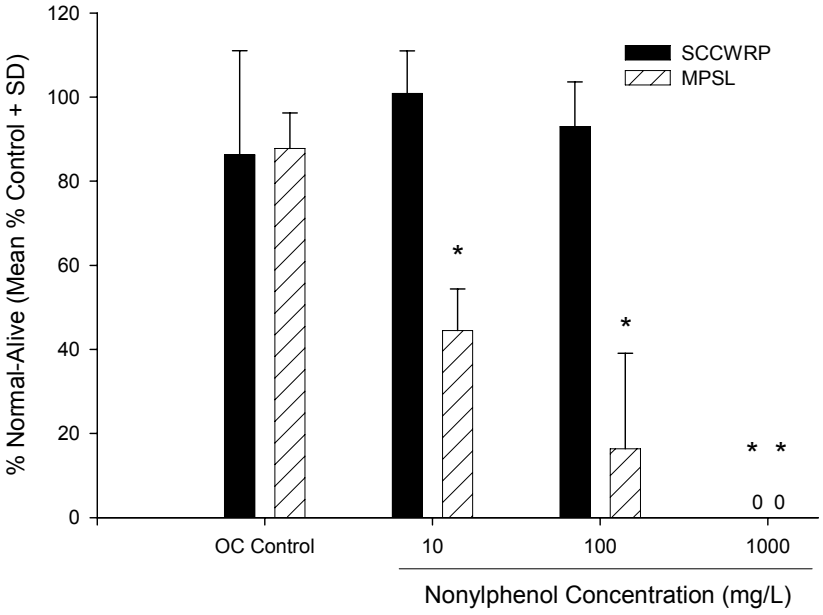

Figure 1. Water only control adjusted sediment-water interface mussel embryo development responses to nonylphenol from Marine Pollution Studies Laboratory (MPSL) and Southern California Coastal Water Research Project (SCCWRP). Results marked with * are significantly different from OC.

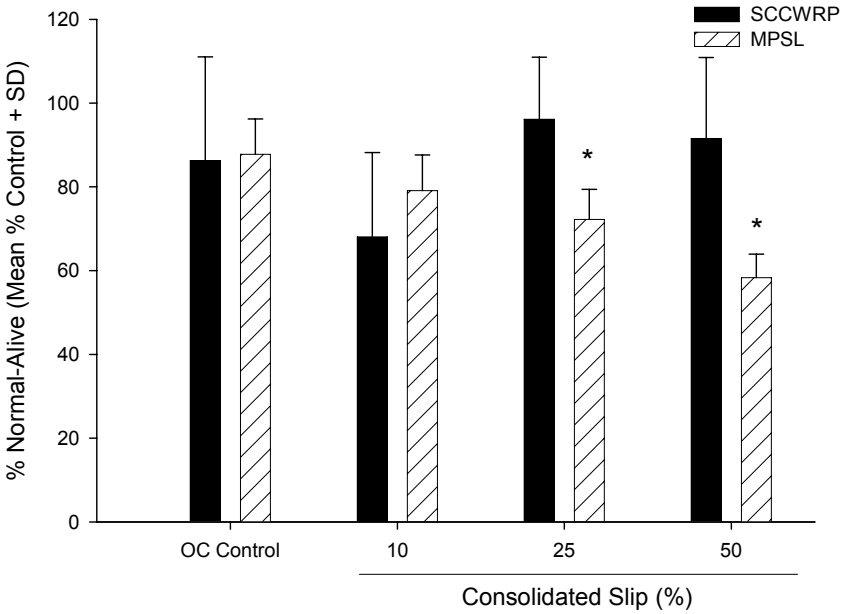

Figure 2. Water only control adjusted sediment-water interface embryo development responses to consolidated slip dilutions from Marine Pollution Studies Laboratory (MPSL) and Southern California Coastal Water Research Project (SCCWRP). Results marked with * are significantly different from OC.

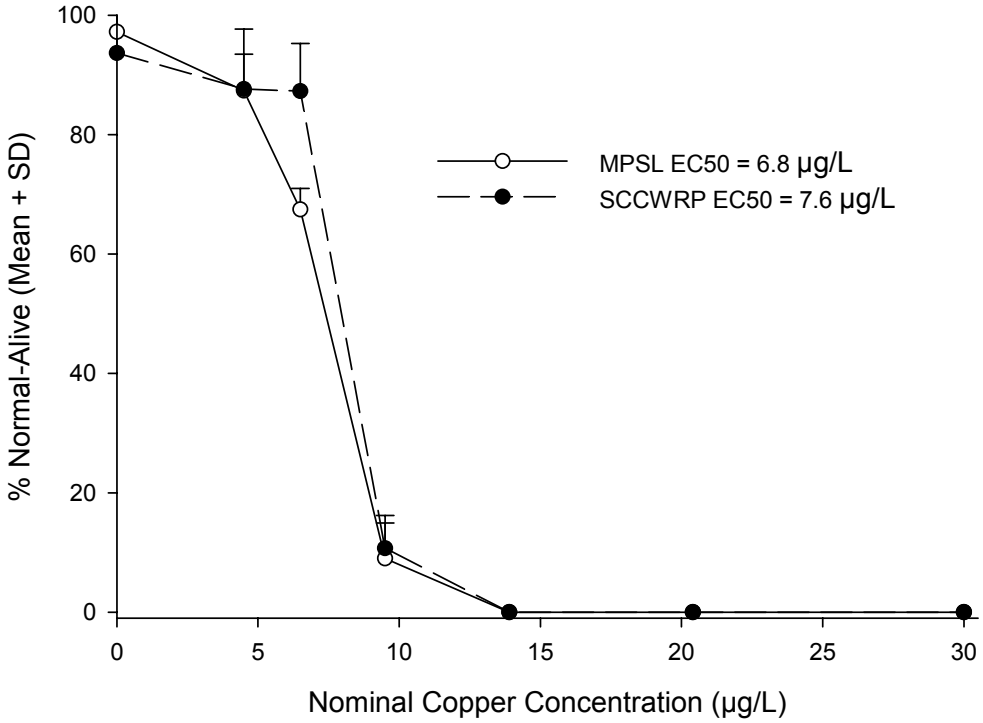

Figure 3. *Mytilus galloprovincialis* dose-response plot of copper reference toxicant exposures to copper for Marine Pollution Studies Laboratory (MPSL) and Southern California Coastal Water Research Project (SCCWRP).

Juvenile Clam Growth

Range Finding

Range finding tests were conducted at SCCWRP to select the concentrations for nonylphenol and the CS dilution samples in the interlaboratory exposures. Two experiments were needed to find concentrations of nonylphenol and CS that showed a dose response. In the first range finding experiment, the three nonylphenol concentrations (10, 100, and 1000) mg/L showed a very similar strong response, each producing growth between 25 to 30 % of control response. Both the CS dilutions tested, 0.5 and 2.0%, had no effect, with growth very similar to the control.

Adjustments were made for the second range finding experiment, decreasing the concentration of nonylphenol by a factor of ten and increasing the concentration of CS in the samples to 5, 10, and 25%. For the second nonylphenol experiment, there was a range of response from 10 to 40% of control growth among the three treatments and these concentrations were selected for use in the intercalibration experiment. In the CS dilution series a strong response in the 25% CS was still not present, therefore concentrations of 10, 25, and 50% CS were tested in the interlaboratory comparison.

Interlaboratory Calibration

None of the laboratories found a significant difference between any of the nonylphenol concentrations and the OC sediment. However, two of the laboratories found that the OC sediment had significantly less growth than the LTH sediment (Figure 4). Therefore, further comparisons were made between all nonylphenol treatments and the LTH sediment (Figure 4). MRRI found a significant difference between all the nonylphenol concentrations and the LTH sample. SCCWRP found a significant difference between the 0.1 and 1.0 mg/kg concentrations and the LTH sample. Weston found no significant difference between any of the nonylphenol samples and the LTH sample.

SCCWRP and Weston found no significant difference in clam growth between LTH and any of the three CS (10, 25, and 50%) dilutions. MRRI found only the 10% CS dilution to be significantly different with 74% of control growth (Figure 5). For all of the laboratories, the growth in the highest two concentrations was similar to or greater than what was observed in the LTH sediment.

The above comparisons detailed whether samples were significantly different from control values, which was deemed a reflection of whether a sample was toxic or not. Another method of comparison is to examine the differences in the growth values themselves between laboratories. For this analysis, the control adjusted means were compared using ANOVAs. There was only one sample (1.0 mg/L nonylphenol) where there was not statistical agreement between the laboratories for clam growth. However, the statistical agreement may be more due to between replicate variability rather than close agreement of the mean growth data from each laboratory. MRRI, SCCWRP and Weston had mean coefficients of variation of 26.6, 35.4, and 42.9, respectively (Table 1). While the mean coefficients of variation were not very different, the differences within individual samples were quite high in many cases. The variation is a little higher than for

the SWI tests where SCCWRP and MPSL had coefficients of variation of about 17 and 36 respectively.

Reference Toxicant

There was a large range of IC50s between MRRI, SCCWRP, and Weston, with 50.2 (95% CI = 43.1 and 58.3), 29.9 (95% CI = 11.8 and 37.5), and 13.5 (95% CI = 10.7 and 19.0) µg/L copper, respectively. All of the laboratories showed decreasing growth with increasing copper concentration (Figure 6). The values above were compared to previous data from Keppler and Ringwood (2002), of the MRRI laboratory. They published an IC50 for copper of 37.6 µg/L from five separate exposures. The IC50 data from MRRI, SCCWRP, and Weston were within one standard deviation of the mean of the five values from the published exposures. Therefore it was concluded that the three laboratories did not differ in reference toxicant outcomes.

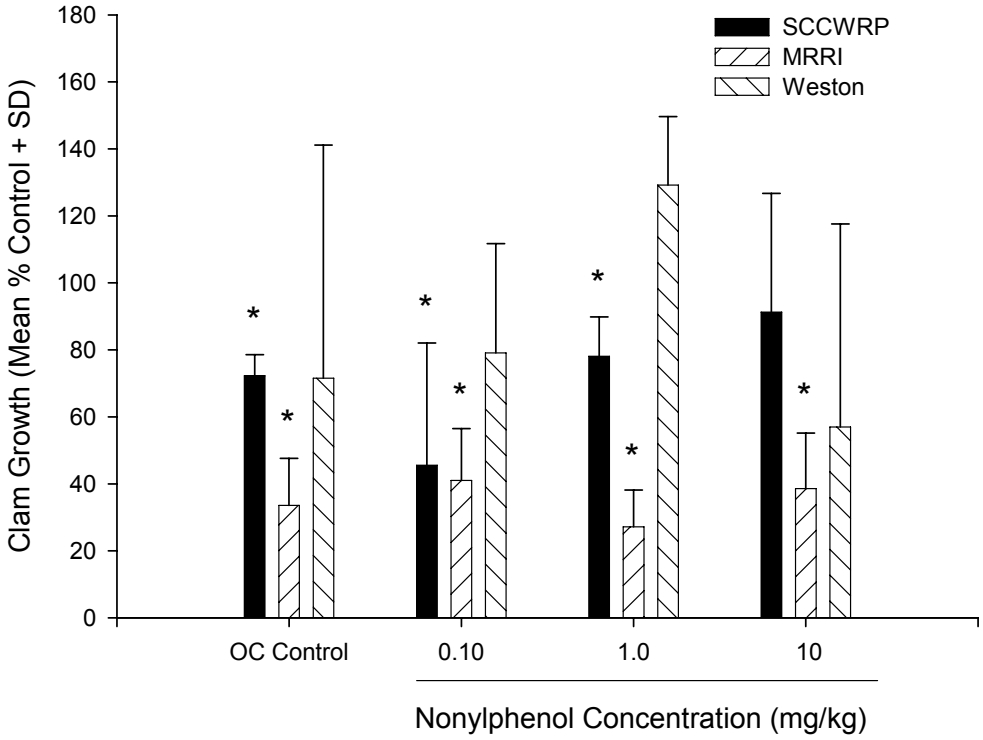

Figure 4. LTH sediment control adjusted juvenile clam 7-day growth test responses to nonylphenol from MRRI, SCCWRP, and Weston. * indicates values significantly different from LTH sediment. Marine Pollution Studies Laboratory (MPSL), Marine Resources Research Institute (MRRI), Weston Solutions (Weston), and Southern California Coastal Water Research Project (SCCWRP).

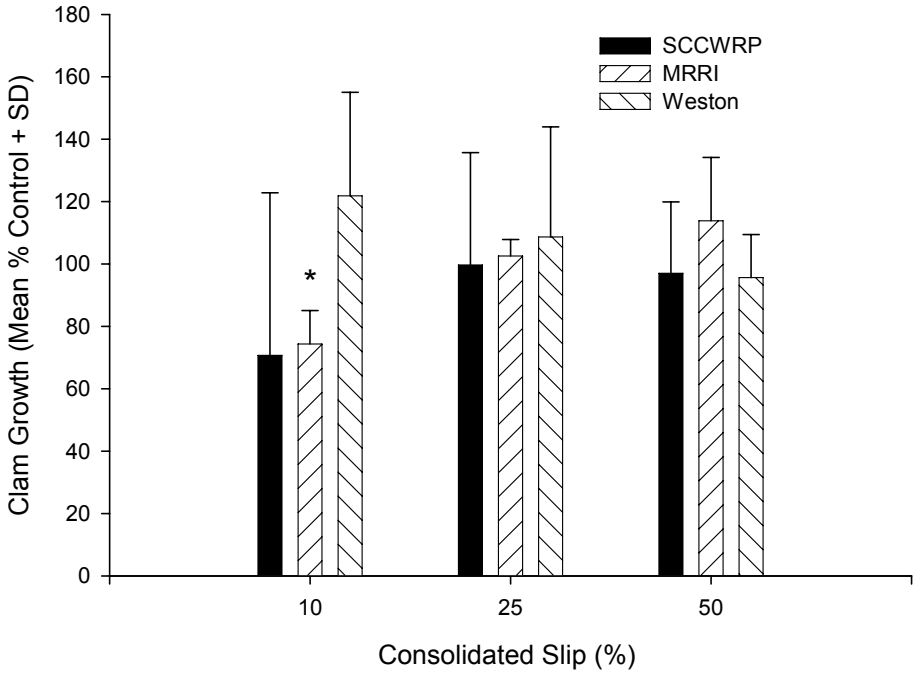

Figure 5. LTH sediment control adjusted juvenile clam 7-day growth test responses to CS dilutions from Marine Resources Research Institute (MRRI), Weston Solutions (Weston), and Southern California Coastal Water Research Project (SCCWRP).. *indicates values significantly different from LTH.

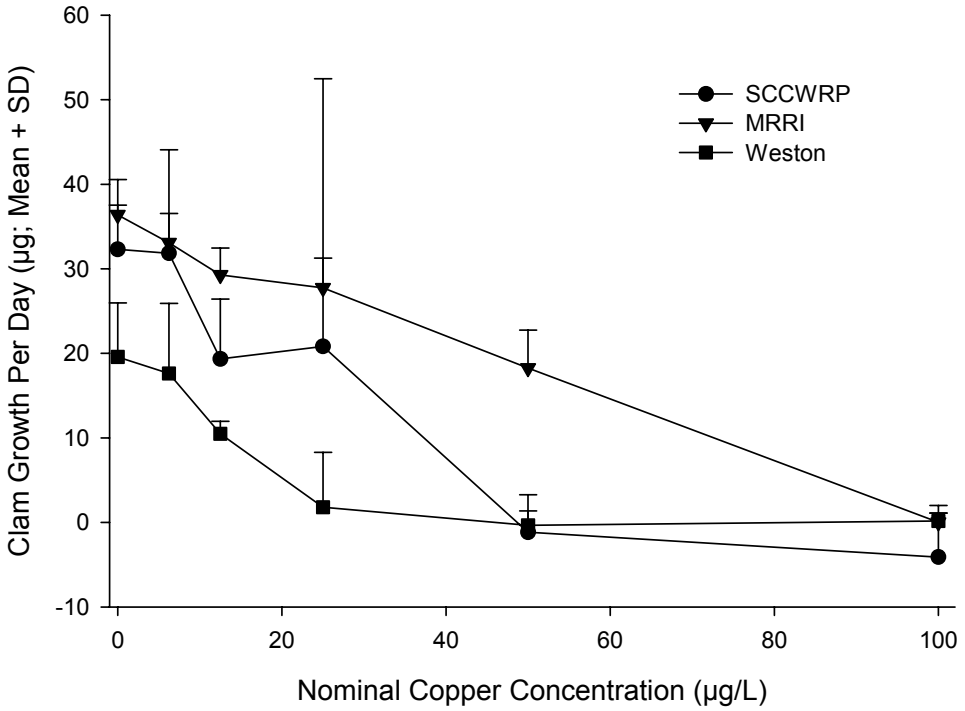

Figure 6. *Mercenaria mercenaria* growth dose-response plot of copper reference toxicant exposures to copper for Marine Resources Research Institute (MRRI), Weston Solutions (Weston), and Southern California Coastal Water Research Project (SCCWRP).

Table 1. Coefficients of variation for Marine Resources Research Institute (MRRI), Weston Solutions (Weston), and Southern California Coastal Water Research Project (SCCWRP) results for the *Mercenaria mercenaria* 7-day growth endpoint.

	Laboratory		
Sample	MRRI	SCCWRP	Weston
LTH	12.8	6.7	13.9
OC	41.8	8.7	97.2
0.1 mg/L nonylphenol	37.8	80.2	41.3
1.0 mg/L nonylphenol	40.1	15.0	15.8
10.0 mg/L nonylphenol	42.8	38.9	106.5
10% CS	14.5	73.8	27.2
25% CS	5.2	36.1	32.5
50% CS	17.9	23.5	8.6
Mean	26.6	35.4	42.9

DISCUSSION

An important attribute of any toxicity test is that the results are comparable between laboratories using the method. There must be confidence that similar results can be achieved when any given test is used by a reputable laboratory. In the current study comparisons were made for the SWI test using mussel embryos and the *M. mercenaria* test. For each intercalibration, the results of a laboratory highly experienced in the use of the method was compared to laboratories with much less experience.

The SWI test has been used previously by a number of laboratories for field studies. However, no previous intercalibration testing has been conducted. For this study, only two laboratories performed the intercalibration. All but two of the samples tested had a significant difference between the laboratories. In all of the cases where there was a difference, the more experienced laboratory had more sensitive results. No other clear explanation for the differences between the laboratories is apparent. Possible explanations are differences in toxic exposure due to differences in sample handling, differences in interpretation of the microscopic endpoint and differences in animal sensitivity. Given the simplicity of the endpoint determination and the similarity in the EC50 values of the reference toxicant between the laboratories, the last two reasons seem unlikely. While there was no previous interlaboratory testing for the SWI test, there is interlaboratory data for the *M. galloprovincialis* embryo test in aqueous solutions. In that testing, it was found that coefficient of variation between five laboratories was 23.6% for cadmium and 14.4% for lyophilized pulp mill effluent (U.S. EPA 1995). The coefficient of variation from the copper reference toxicant exposure in the current study was 7.9%, which compares favorably with the previous study.

For the *M. mercenaria* test, there was no significant difference in growth among the laboratories for most of the sediment samples. However, the less experienced laboratories encountered a higher degree of between replicate variability than the experienced laboratory. This variability may in part explain the lack of a significant difference among the laboratories. With more familiarity with the procedures, the between replicate variability should decrease, as should the degree of difference in mean growth.

Examining various aspects of the results can help to make an overall assessment of the degree of variability between laboratories in this study. For the SWI testing, the agreement between the laboratories for the nonylphenol spikes was judged to be fair, with one laboratory finding significant toxic response for all three concentrations, while the other found only one. However, both laboratories agreed that there was complete mortality at the highest concentration. There was poor agreement for the CS dilutions with one laboratory finding toxicity in two dilutions and the other finding no toxicity in any. Finally, both laboratories had very good agreement on the reference toxic exposure. Given this mixture of results the overall assessment is that the interlaboratory agreement was assessed as fair.

The *M. mercenaria* results can be judged for interlaboratory agreement using the same method. For the nonylphenol spiked sediments, there was good agreement between two of the laboratories, but poor agreement with the third. For the CS dilutions, there was decent agreement among all three laboratories, however there was very little toxicity associated with the samples. While there was a fairly wide spread in the IC50 data for the reference toxic, data fell within range of variability observed during previous testing. As for the SWI test, it was judged that the overall degree of interlaboratory variability for the *M. mercenaria* test was a rating of fair.

LITERATURE CITED

Anderson, B.S., J.W. Hunt, M. Hester and B.M. Phillips. 1996. Assessment of sediment toxicity at the sediment-water interface. pp. 609-624 *in*: G.K. Ostrander (ed.), Techniques in aquatic toxicology. CRC Press Inc. Boca Raton, FL.

Ditsworth, G.R., D.W. Schults and J.K.P. Jones. 1990. Preparation of benthic substrates for sediment toxicity testing. *Environmental Toxicology and Chemistry* 9:1523-1529.

Keppler, C.J. and A.H. Ringwood. 2002. Effects of metal exposures on juvenile clams, *Mercenaria mercenaria*. *Bulletin of Environmental Contamination and Toxicology* 68:43-48.

Lamberson, 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, Inc. Boca Raton, FL.

Ringwood, A.H. and C.J. Keppler. 1998. Seed clam growth: An alternative sediment bioassay developed during EMAP in the Carolinian Province. *Environmental Monitoring and Assessment* 51:247-257.

U.S. Environmental Protection Agency (USEPA). 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA/600/R-94/025. Office of Research and Development, U.S. Environmental Protection Agency. Narragansett, RI.

USEPA. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. Office of Research and Development. Cincinnati, OH.