Selection of methods for assessing sediment toxicity in California bays and estuaries

Darrin J. Greenstein and Steven M. Bay

Abstract

Toxicity tests are often used in sediment assessment programs. However, the choice of methods has been largely limited to acute tests. Where sublethal methods have been employed, there has been little consistency among programs in the types of the sublethal tests used. The goal of this study was to develop a method for choosing a suite of acute and sublethal tests for use in a California statewide assessment program and to develop a set of method specific thresholds for classifying the degree of toxicity within a multiple line of evidence framework consisting of sediment chemistry, benthic community structure, and sediment toxicity. A group of candidate methods was evaluated using feasibility and performance criteria. Toxicity thresholds were calculated based on test variability and sensitivity. As a result of the evaluation, three acute toxicity methods using amphipods (*Eohaustorius estuarius*, *Rhepoxynius abronius* and *Leptocheirus plumulosus*), and two sublethal methods using a polychaete and mussel embryos (*Neanthes arenaceodentata* growth and *Mytilus galloprovincialis* embryo development at the sediment-water interface) were selected for recommendation. Thresholds for toxicity categories corresponding to Nontoxic, Low Toxicity, Moderate Toxicity and High Toxicity were developed for each test method. While these toxicity categories and thresholds provide a consistent framework for the interpretation of test results among different methods, additional research is needed to determine their effectiveness for predicting impacts to benthic communities.

Introduction

Toxicity tests have been widely used to assess sediment quality and are an integral part of the sediment quality triad used in many marine monitoring and assessment programs (Long and Chapman 1985, Fairey *et al*. 1998, Hunt *et al*. 2001b). The applications have included dredged material characterization, contaminated site assessments, and regional monitoring surveys such as the US EPA's Environmental Monitoring and Assessment Program (EMAP; Strobel *et al*. 1995). Much of the testing has employed acute amphipod survival methods using protocols established by the U.S. Environmental Protection Agency (USEPA 1994). In Europe, the OSPAR Convention allows for the use of several different taxa for sediment toxicity testing protocols (OSPAR 2007). However, there are not sublethal endpoints for any of the protocols, and the preferred method is the amphipod, *Corophium volutator*.

Many sediment quality assessment programs use a suite of two or more toxicity tests that include both acute and sublethal tests. Some EMAP studies have used amphipod acute testing in conjunction with a variety of sublethal methods in different parts of the country (Ringwood *et al*. 1996, Bay *et al*. 1998). For example, as part of the National Status and Trends Program multiple sublethal tests were conducted in concert with an amphipod acute test on sediments from the Biscayne Bay region of Florida (Long *et al*. 1999). In addition, the State of Washington has a program for monitoring and assessing sediments that uses a combination of acute amphipod tests and two sublethal tests with other taxa (PSWQA 1995).

A wide diversity of sublethal toxicity test methods have been used as part of marine sediment assessment programs. Sublethal test methods always include measures of toxicity other than survival (*e.g.,* growth, development), but organism survival may also be an additional endpoint used for data evaluation and interpretation. There is little consistency among programs in the types of the

sublethal tests used; selection is often site-specific and is based on factors such as availability of test organisms, cost, laboratory staff and expertise, local interests, habitat type, 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 test using sea urchin or mussel embryos (Anderson *et al*. 1996).

While test sensitivity is an important issue, the selection of sediment toxicity test methods requires a consideration of many other factors, depending upon the study's objectives and design. Much diversity 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, with 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 responsiveness.

The sediment quality objectives (SQO) program in the State of California provides an example of the many factors to be considered when sediment toxicity tests are used in a regulatory context. This program uses toxicity testing, along with sediment chemistry and benthic infaunal analysis to assess the condition of bays and estuaries throughout the state (CSWRCB 2009). The toxicity test methods for such a statewide regulatory program should be ecologically relevant and protective by using exposure methods with a well-understood linkage to sediment contamination and using test species that respond to contaminants at concentrations that are likely to cause ecological impacts. These tests should also be standardized to ensure consistent application and be feasible for application in a variety of situations. In addition, a consistent and relatively simple method of toxicity data interpretation is needed so that station assessments conducted in one region can be compared to the results from other regions, both spatially and temporally. Larger scale programs, such as EMAP and NOAA Status and Trends, have similar needs as a statewide program, but with the additional need to have data that are comparable over an even larger spatial scale. Comparisons of sediment toxicity test methods have been conducted previously for freshwater species in Canada (Keddy *et al*. 1995) and for use in testing dredged materials in Europe (Nendza 2002). However, these comparisons have not addressed many of the issues of concern for a statewide program, were for use in different habitats, or were for a different subset of test methods.

Selection of the thresholds used for sediment toxicity data interpretation is an important aspect of the test method that is frequently given little consideration with respect to application in large-scale regulatory programs. Most thresholds use some form of statistical evaluation, usually combined with individual judgment regarding the test acceptability or biological significance of the response (Thursby *et al*. 1997). Data on control performance, test variance, and relationship of response to biological significance is needed to guide threshold selection, but such data may not be available for some methods, especially new methods/applications. Consequently, data interpretation thresholds from other methods or programs are often adopted without evaluation as to whether they are appropriate for the current program's objectives and design.

The current study had two principal objectives. The first objective was to evaluate a variety of acute and sublethal sediment toxicity tests in order to identify methods that were best suited for use in California's SQO 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 system to classify the toxicity results of each test method into multiple categories of effect, relative to the control response. 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.

Methods

Protocol Comparisons

A set of candidate acute and sublethal sediment toxicity test methods was selected for evaluation. The methods that were selected included direct exposure to sediment, 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 US Army Corps of 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.

The general strategy for evaluation of the test methods was to assess each one using a set of common parameters (Table 1). Each test method 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). Evaluation within each parameter was either on a binary (+/-) basis or a categorical determination. The three feasibility characteristics (organism availability, method description, and technical difficulty) were evaluated using the binary, pass/fail, scoring system. These characteristics were deemed to be so important that a test was classified as not feasible if minimum criteria for these were not met (Table 1). A similar strategy has been used in another study where certain evaluation categories were deemed essential (Keddy *et al*. 1995). The performance and cost parameters were summarized into categories that reflected the relative level of attainment (e.g., poor, fair, good).

The following test characteristics were evaluated with associated information coming from literature sources, as well as from contact with scientists having direct experience with the various methods, including the authors: **1) Organism availability**. Ideally, test organisms should be available from multiple suppliers on a year-round basis with no seasonal variation in test sensitivity. However, the minimum requirement was a single commercial supplier. **2) Method description.** Methods had to have a detailed published protocol, with the availability of control acceptability criteria and quality assurance standards for parameters such as water quality in order to receive a "+" rating. **3) Technical difficulty**. 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 that cannot be obtained with relative ease, and complexity of the exposure system. **4) Concordance of sublethal responses**. For the sublethal methods, there was an expectation that if a site was acutely toxic to a test organism, then an effect would also be seen for the sublethal test. Conversely, if a site was considered to be in "reference condition" (i.e., low chemical concentrations and/or unaffected benthic community) then there would be an expectation that no toxicity would be observed. To evaluate concordance, the response of an acute amphipod test was used as the basis for comparison. **5) Relative sensitivity**. Test sensitivity was evaluated relative to the acute amphipod test species most commonly used in California, *Eohaustorius estuarius*. 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. For many of the methods, no data were available, so a study was conducted to help fill this information gap (Greenstein *et al*. 2008). **6) Variability among laboratories**. In addition to literature sources which compared interlaboratory variability, supplemental testing was conducted for the *Mercenaria mercenaria* growth test and the sediment-water interface test using mussel embryos (Bay *et al*. 2007). **7) Variability within laboratories.** Reference toxicant exposure data supplied by laboratories that routinely use the methods were applied to evaluate intralaboratory variability. **8) Precision.** This category was assessed by comparing between replicate variability among the methods from data in the literature or as supplied by laboratories. **9) Documentation of confounding factors**. The methods were evaluated for the presence of toxicity information on non-contaminant factors, such as ammonia or sediment grain size. **10) Cost.** The unit cost of each test was evaluated relative to the standard 10-day amphipod survival test, assumed to be approximately $800 per sample.

A weighting factor was established for each category based on its relative importance. The

Table 1. Description of toxicity test method evaluation parameters.

Category	Description	Evaluation Type	Criteria
Organism Availability	Year-round availability of test organism from commercial supplies	Pass/Fail (+/-)	At least one commercial source of organisms available
Method description	Availability of a standardized, reviewed protocol	Pass/Fail (+/-)	A published document with complete description; test acceptability criteria available.
Technical difficulty	Likelihood of laboratories to be able to perform the test in a consistent manner	Pass/Fail (+/-)	Not require any specialized skills or equipment that most laboratories would not already possess
Concordance of sublethal responses	Ability of a test method to identify sites that are clearly contaminated or reference, compared to acute tests	Categorical (1-3)	Concordance with acute amphipod test: Good=>75%; Fair=<75%>50%; Poor<50%
Relative sensitivity	Comparison of sensitivity to that of standard acute amphipod test	Categorical (0-3)	Frequency of finding a station toxic relative to the amphipod test: Often=>50% of stations; Sometimes=<50%>20, Rarely<20%; Never=0%
Variability among laboratories	Expectation of consistency of results between laboratories	Categorical (1-3)	Good = CV<50%; Fair= CV >50% <75%; Poor=CV>75% (CV= coefficient of variation; mean/standard deviation)
Variability within laboratories	Expectation of consistency of results within laboratories	Categorical (1-3)	Based on the range of median acute amphipod standard deviations. High=below range; Similar=within range; Low=above range
Precision	Expectation of between replicate variability	Categorical (1-3)	Based on the range of median acute amphipod standard deviations. High=below range; Similar=within range; Low=above range
Documentation of confounding factors	Level to which non-contaminant (e.g. grain size) effects have been tested	Categorical (1-3)	Data available for confounding factors: Good=Four or more factors; Fair= 2 or 3 factors; Poor= Less than 2 factors
Cost	Relative expense of conducting test	Categorical (0-3)	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.

comparative sensitivity category was assigned the highest weight: a factor of 4. The high weight given to this category was based on the importance of having test methods that are more responsive to contaminant-related toxicity than the acute toxicity methods currently in use. The precision parameter was deemed to be the least important since it can be somewhat ameliorated by changing the level of replication and was therefore 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. The acute and sublethal test methods were evaluated separately during the assessment process since some of the comparisons were made relative to the acute tests.

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 (3 equaling the best performance within a category) assigned based on the data review. For most of the categories, scores ranged from 1 to 3 unless no data could be located, in which case a zero was assigned. For the sensitivity and cost categories, a four level classification was made with zero being the lowest possible score. This was done due to the wide range of data in these categories and the judgment that the poorest performance for them was highly undesirable. Each value was multiplied by its respective weighting factor to produce a score for the characteristic. The scores were then summed to obtain a final score for each candidate test method. The test methods that both met the feasibility criteria and had the highest scores were recommended for use in the SQO program.

Toxicity Threshold Calculations

An ordinal scoring system consisting of four categories of response was developed for each of the recommended tests. 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. 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. 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 response: **1) Nontoxic:** Response not substantially different from that expected in sediments that are uncontaminated and have optimum characteristics for the test species; **2) Low Toxicity**: A response that is of low magnitude; the response may not be greater than between test variability; **3) Moderate Toxicity**: High confidence that a statistically significant effect is present; **4) High Toxicity**: Highest confidence that a toxic effect is present and the magnitude of response is among the strongest effects observed for the test. The four toxicity categories were intended to classify the response relative to the range of response observed for each method. These categories were not intended to represent levels of biological response expected in the field (e.g., change in benthic community condition), because the biological response data needed to calibrate such relationships were not available for some methods.

Classification of the test response was accomplished by applying a series of numeric thresholds (based on magnitude of response) and statistical criteria in a stepwise process (Figure 1). This approach relies on the comparison of the test result (e.g., % survival) to 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 the distribution of historical toxicity response magnitudes. A statistical criterion was also used in the classification scheme (Figure 1). 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 different relative to the control (t test, $p \leq 0.05$). A statistical significance criterion was not applied to the High Toxicity category because the derivation of this threshold already incorporated a high degree of statistical confidence.

The threshold separating the Nontoxic and Low categories was defined as the lowest acceptable control response value (e.g., survival or growth) for the given test. 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 response value is defined as the mean value for the endpoint for a given test method (i.e., survival, growth). Any sample having a

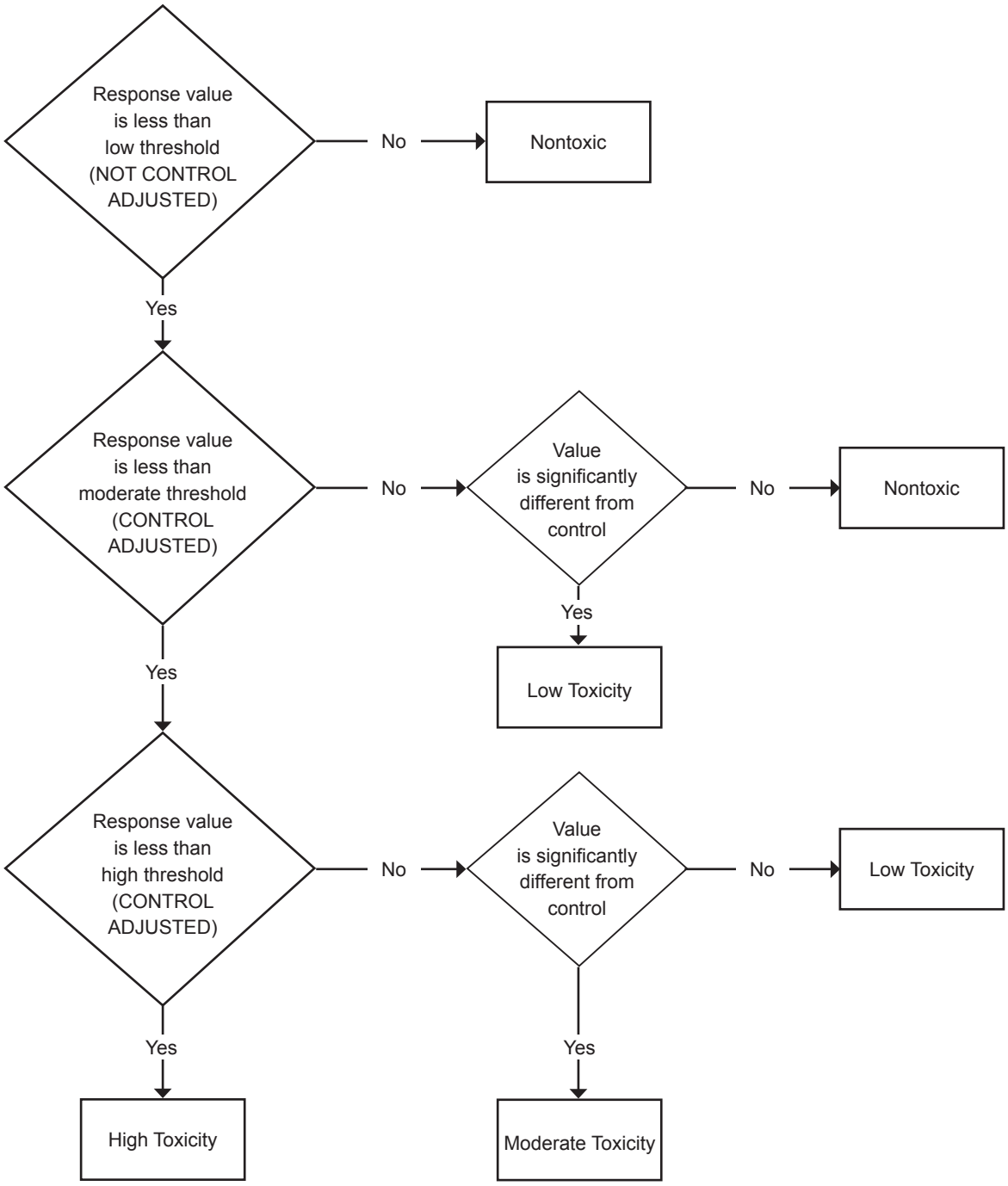

Figure 1. Conceptual approach for assigning the category of toxic effect from exposure response (e.g., percentage survival or growth) data.

response value that is greater than or equal to the low threshold will be classified as nontoxic, regardless of whether a statistical difference from the control exists (alpha=0.05, Figure 1). Since this threshold is not based on comparison to the control response within a sample batch, it is the only threshold for which the data are not control normalized before comparison.

The intent of the Moderate Threshold is to distinguish between samples producing a small response of uncertain statistical significance and larger responses representing a reliably significant difference relative to the control. This threshold was based on the MSD, which was specific to each test method. The MSD represents the minimum

difference between the control and sample response that is necessary to be statistically different at $p \leq 0.05$ level with $\alpha = 0.05$. 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 demonstrating a high degree of confidence that the sample is different from the nontoxic condition.

The MSD values were calculated using the replicate control and sample data from many toxicity tests. Details of this calculation can be found in Phillips *et al*. (2001). For each combination of a control and a sample, the variance of the replicates, number of replicates, and the one tailed t-critical value for the pair were used to calculate a single MSD value. 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 1). Sample response values that are less than the moderate threshold (i.e., lower survival) and are significantly different from the control are categorized as Moderate Toxicity.

The intent of the High Toxicity threshold is to identify samples producing a large and highly significant effect from those samples producing lesser effects. This threshold was based on a combination of two characteristics: test variability and cumulative percent distribution of response values for toxic samples. The test variability portion of the threshold ensured high likelihood of statistical significance of the high threshold. Incorporation of the distribution of the toxic sample data ensured that the high toxicity threshold represented a magnitude of response expected to occur infrequently and only in the most contaminated samples.

The 99th percentile MSD value was used to link the high threshold to test variability. A sample having a response value (i.e., survival or growth) that falls below this limit 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 in the database were ranked in descending order based on the control-adjusted mean response. The 75th percentile of the data was selected for the response distribution component. The value obtained from this calculation represents the response associated with the most strongly affected 25% of the toxic samples found in California. Data for this calculation were based on stations within California in order to obtain a response value that was relevant to the characteristics of sediments likely to be evaluated with the test.

Both the variability and data distribution response values represented important, but different, aspects of the high threshold. Therefore, the mean of the two values was used as the high threshold.

Results and Discussion

The initial process of method selection led to the identification of six candidate sublethal methods for evaluation (Table 2). In addition, four amphipod species recommended by the USEPA for testing acute sediment toxicity were also included on the list (USEPA 1994). The methods were a mixture of commonly used protocols, such as the amphipod tests, and some sub-lethal methods that had been used sparsely, such as the lysosomal stability test. However, each method had published results that showed their promise for use in sediment toxicity assessments.

The evaluation of the candidate acute and sublethal tests identified five methods that had the best overall combination of technical feasibility and relative performance. These methods include three acute amphipod and two sublethal test methods (Table 3). The evaluation process identified several key differences among the test methods that were important in the final selection.

Table 2. Characteristics of candidate sediment toxicity tests selected for evaluation.

Species	Taxonomic Group	Duration (days)	Matrix	Endpoint(s)	Literature Level[1]	Citations	QA criteria[2]	State/National Programs Use[3]
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
Leptocheirus plumulosus	Amphipod	28	Whole sediment	Growth, reproduction, survival	Well established	USEPA 2001	Yes	USACE
Neanthes arenaceodentata	Polychaete	28	Whole sediment	Growth, survival	Published	ASTM 2002b; Farrar and Bridges 2011	Yes	USACE [4] 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	Published	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]Indication of how widely the method protocol has been published and rigor with which it has been reviewed.

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

[3]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.

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

Acute Tests

Analysis of the acute amphipod methods found that *E. estuarius*, *L. plumulosus* or *R. abronius* were the best choices to recommend for use in California sediment assessments. 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). All of the amphipod species were scored as having met the feasibility criteria. Both *E. estuarius* and *R. abronius* consistently received the most favorable category classifications for reproducibility, documentation of confounding factors, and cost. *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 shown to have good reproducibility between laboratories (Bay *et al*. 2003b). The amphipod *R. abronius* has also been 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 when tested simultaneously. An interlaboratory comparison exercise using this method found good agreement amongst the testing laboratories (Mearns *et al*. 1986). Sediments with a silt-clay content of ≥ 80% have been shown to be a confounding factor for *R. abronius* (DeWitt *et al*. 1988), which might make this species less desirable for use in embayments where finer grained sediments are common. A slightly lower total score was obtained for the acute test with *L. plumulosus* (Table 3), which was due to lower reproducibility within and among laboratories. The *L. plumulosus* 10-day test has been conducted in California on a much more limited basis. However, it has long been used in other parts of the country, especially on the

Table 3. Ranking matrix of acute and sublethal sediment toxicity methods. Final score is sum of ratings (maximum score = 45). Weighting factors for performance and cost are in parentheses.

	Feasibility				Performance and Cost								
	Organisms Availability	Method Description	Technical Difficulty	Overall Feasibility	Concordance of Sublethal Responses (2)	Relative Sensitivity (4)	Variability among Laboratories (2)	Variability within Laboratories (2)	Precision (1)	Documentation of Confounding Factors (2)	Cost (2)	Total Score	Recommended for Use in the California SQO Program
Amphipod Acute													
E. estuarius	+	+	+	Yes	NA	8	6	6	2	6	6	34	Yes
R. abronius	+	+	+	Yes	NA	8	6	6	2	6	6	34	Yes
L. plumulosus	+	+	+	Yes	NA	12	4	2	2	4	6	30	Yes
A. abdita	+	+	+	Yes	NA	4	2	6	2	4	6	24	No
Sublethal Methods													
Mercenaria growth	+	-	+	No	4	8	4	4	2	6	6	34	No
Neanthes survival and growth	+	+	+	Yes	4	8	6	6	1	6	2	33	Yes
SWI *M. galloprovincialis*	+	+	+	Yes	4	8	4	6	1	4	6	33	Yes
SWI *S. purpuratus*	+	+	+	Yes	4	4	0	6	1	6	6	27	No
L. plumulosus-28 Day	+	+	+	Yes	4	8	4	6	1	6	2	31	No
A. tenuiremis lifecycle	-	+	-	No	6	12	0	6	3	4	0	31	No
C. virginica lysosomal stability	+	-	-	No	2	4	0	0	1	2	4	13	No

NA= Not analyzed

SWI= Sediment-water interface exposure

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). A desirable attribute of *L. plumulosus* is that it is easily cultured in the laboratory and therefore available year round from commercial suppliers.

The final amphipod species, *A. abdita*, was assigned the lowest total score among the four acute test species. This species has been used in several monitoring and assessment studies within California (Bay *et al*. 1998, Hunt *et al*. 2001a, USEPA 2006). The low score was driven by relatively low responsiveness compared to *E. estuarius* and a lower reproducibility among laboratories (Table 3). Specifically, paired tests of California sediments frequently found a lack of toxic response in *A. abdita* for samples causing greater than 20% mortality to *E. estuarius* (Figure 2). This may be due to the fact that *A. abdita* does not burrow in sediment, but lives in a tube-like structure and does not ingest sediment (Anderson *et al*. 2008). The *A. abdita* test was also rated the most technically difficult of the 10-day tests based on the experience of some California laboratories that have encountered a higher test failure rate, based on control survival, than for other amphipod species (B. Phillips, personal communication). *A. abdita* is widely used as a sediment toxicity indicator in many monitoring programs and the data have been used to characterize sediment quality on a national scale (Long 2000, USEPA 2006). Laboratories outside of California have had a high rate of success in conducting tests with *A. abdita* and technical difficulties reported in California should not preclude the use of the test in other regions.

Sublethal Tests

The candidate sublethal tests were more variable in regards to feasibility, performance, and cost than the acute methods (Table 3). Three methods met all feasibility criteria and the highest and similar ranking scores, covering a narrow range of 31-33 out of maximum total of 45 (Table 3). However, based on the results of this analysis, two sublethal methods were chosen to be recommended for use in the California SQO program, the *N. arenaceodentata* 28-day growth test and the sediment-water interface test with mussel, *M. galloprovincialis*. The third method, 28-day *L. plumulosus* test, is also a feasible test, but was judged to be less desirable because the

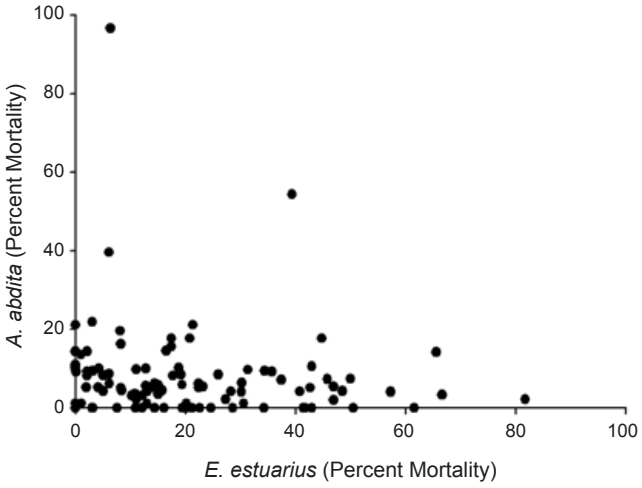

Figure 2. Comparison of mortality data between *A. abdita* and *E. estuarius* on split samples.

test is more costly to perform, provides no increase in taxonomic diversity and its sensitivity relative to the 10-day survival test with the same species is uncertain.

The *N. arenaceodentata* growth and survival test was tied for the highest ranking of the feasible sublethal tests (Table 3). It is fairly well established with an ASTM method (ASTM 2002b), 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, Kennedy *et al*. 2009). 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 (Figure 3) and more sensitive than the survival test with *E. estuarius*. However, other studies have reported lower sensitivity compared to *E. estuarius*. While the *N. arenaceodentata* test is one of the more expensive to conduct, it has relatively high sensitivity (among the sublethal tests), reliability, and technical feasibility.

The sediment-water interface (SWI) test using mussel embryos received the same total score as the *N. arenaceodentata* test (Table 3). The exposure protocol for this procedure is published in a well respected compendium of toxicity test methods (Anderson *et al*. 1996) 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*. 2001b, Bay *et al*. 2004, Brown and Bay 2011). 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 (Figure 3; Bay *et al*. 2004).

The *L. plumulosus* 28-day test received a relatively high total score that was only two points below the *N. arenaceodentata* and SWI test methods. 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, Kennedy *et al*. 2009). During a study of California sediments, the *L. plumulosus* 28-test experienced a test failure and variability in the reproduction endpoint caused that data to be unreliable (Greenstein *et al*. 2008). 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).

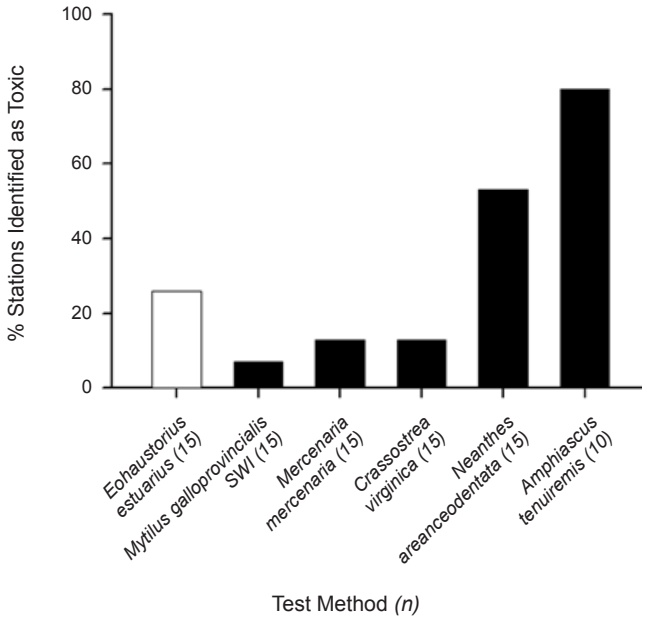

Figure 3. Percentage of stations that were identified as being toxic by the standard amphipod survival test and sublethal sediment toxicity methods. Number of samples tested is in parentheses. SWI = sediment-water interface (Greenstein *et al.* 2008).

The remaining three sublethal test methods had limited method documentation, organism availability, or a high degree of technical difficulty that resulted in an overall rating of not feasible for use in a statewide assessment program at this time. These methods were the bivalve *M. mercenaria* growth test, the copepod *A. tenuiremis* life-cycle test, and the lysosomal destabilization test using the oyster *C virginica*.

The *M. mercenaria* growth test received the highest total performance score of any of the methods, based on average to slightly above ratings in all of the categories (Table 3). However, there is not a single, cohesive document that completely details the protocol and there are no test acceptability criteria. The test is economical and is not technically difficult to perform. The method exhibited fair reproducibility between laboratories in a round-robin study (Bay *et al*. 2007). In a previous study in the EMAP Carolinian Province, the clam growth test found no toxicity in reference areas, but was able to identify areas that were clearly degraded as toxic (Hyland *et al*. 1999) and was the best of the toxicity tests conducted at predicting expected bioeffects (Van Dolah *et al*. 1999). 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 (Figure 3).

The life-cycle test with the copepod *A. tenuiremis* was by far the most sensitive of the sublethal methods compared to amphipod acute tests (Figure 3 and Table 3). This method was also shown to be very sensitive compared to an amphipod acute test in a previous study in Florida (Long *et al*. 1999). Nevertheless, the *A. tenuiremis* test did not pass either the animal availability or technical difficulty feasibility criteria. 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 increased technical difficulty. In order to conduct this toxicity test, a laboratory must maintain the copepod culture, and cultures of three algal species used to feed them. The *A. tenuiremis* life cycle test is also approximately three times

more expensive than other tests and has received no interlaboratory testing to document reproducibility.

The oyster lysosomal destabilization test had the lowest total score of any of the test methods (Table 3). Besides the low ranking, this test method does not have a complete protocol that is published (Table 2). In preliminary testing 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 had not been used in field studies alongside other test methods. In a multi-species comparison, the oyster lysosomal destabilization test was less sensitive compared to an acute amphipod test (Figure 3).

Threshold Calculations

The thresholds derived in this study represent a 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 confidence, these categories represent the two factors that are essential to describing a toxicity test response for these assessments.

The low threshold for each of the toxicity test methods was based on the control acceptability criteria for the given protocol (Table 4). For the *N. arenaceodentata* growth endpoint, the threshold is based on animal weight data, according to a revised ASTM protocol that is in preparation (J. D. Farrar, personal communication). For the sediment-water interface test using *M. galloprovincialis* embryos, the low threshold for normal-alive was based on the control criterion established by the Marine Pollution Studies Laboratory, Granite Canyon (B. Phillips, personal communication).

The moderate thresholds were all based on MSD calculations using various datasets (Table 4). For the *E. estuarius* and *R. abronius* 10-day survival test and the sediment-water interface test with *M. galloprovincialis* embryos the moderate threshold was calculated using data from the California Sediment Quality Objectives database (ftp://ftp.sccwrp.org/pub/download/TOOLS/SQO/sqo.zip), which included 876, 264, and 118 samples, respectively. The threshold for the *L. plumulosus* 10-day survival test was calculated using data from tests on sediment from throughout the U.S., provided by multiple laboratories. Few of the 199 samples in the *L. plumulosus* dataset were from stations located in California. The threshold of the *N. arenaceodentata* growth test was calculated from tests of 92 samples from throughout the United States, as relatively few data from tests on sediments from California were available.

The *M. galloprovincialis* sediment-water interface test low and moderate thresholds appear to represent a very narrow range of response (Table 4). However, 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

Table 4. Response threshold values for the recommended sediment toxicity test methods.

Species	Low (%)	Moderate[2] (% Control)	High[3] (% Control)
E. estuarius (survival)	90	82	59
R. abronius (survival)	90	83	70
L. plumulosus (survival)	90	78	56
N. arenaceodentata (growth)	90[1]	68	46
M. galloprovincialis (development)	80	77	42

[1] % of control growth.

[2] Number of data used in calculation is same as that used for 99th percentile MSD (Table 5).

[3] Number of data used in calculation is given in Table 5.

a noncontrol-adjusted value of 65% (77% of 85% is 65%), representing a response window of about 15% for the low toxicity category.

The high threshold calculation produced a greater range of values (42% for *M. galloprovincialis* to 70% for *R. abronius)* between test methods than did the low and moderate thresholds (Table 4). This greater range is indicative of the range of sensitivities of the various test methods. The MSD_{99} values (expressed as the control normalized percentage response) ranged from 46% for *N. arenaceodentata* to 73% for *R. abronius* (Table 5). The 75th percentile ranged from 24% for *M. galloprovincialis* to 66% for *R. abronius*. 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.

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 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 moderate threshold 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 sediment-water interface 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 analyses described here were used to select a suite of toxicity test methods for use in a multiple line of evidence sediment assessment framework in California. This suite represents test methods that had the best relative 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. These limitations included a shortage of test data for the *L. plumulosus* and *N. arenaceodentata* sublethal methods and a narrow range of available sublethal test methods from which to choose. Further development of the methods for the *M. mercenaria* and *A. tenuiremis*, as well as other sublethal methods, is recommended for future evaluation for use in large scale assessments.

The response classification thresholds developed in this study were developed for application within a multiple line of evidence framework and should not be used as stand-alone predictors of biological impairment. These thresholds were based on statistical parameters unrelated to ecological responses, and the categories of Low, Moderate, and High Toxicity describe response relative to the test control, not

Table 5. Data used in calculation of high threshold values for selected acute and sublethal sediment toxicity test methods. The high threshold is the mean of the two response values shown in the table.

Species	99th Percentile MSD	Number of Data Points	75th Percentile Response	Number of Data Points
E. estuarius	61	876	57	333
R. abronius	73	264	66	114
L. plumulosus	54	199	571	_[1]
N. arenaceodentata				
Growth	46	92	_[2]	_[2]
M. galloprovincialis	60	118	24	28

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

[2]No California data available.

benthic infauna in the field. Toxicity tests and benthic community assessment measure different aspects of sediment quality and there is no *a priori* expectation that toxicity test results will correspond to effects on biota in the field, although such a relationship may be incorrectly assumed (Chapman and Wang 2001, Hose *et al*. 2006). Investigations of the relationship between acute toxicity to marine amphipods and benthic community effects demonstrate that such toxicity tests are ecologically relevant, as responses characteristic of the Moderate and High Toxicity categories correspond with adverse effects on benthic infauna, but the relationship is highly variable (McGee *et al*. 1999, Anderson *et al*. 2001, Long *et al*. 2001, Bay *et al*. 2003a). Similar comparisons cannot be made for marine sublethal tests due to limited data. As more data become available, the thresholds developed in this study should be reevaluated in order to confirm that they perform as intended and to document their correspondence with impacts to benthic infauna.

It is unlikely that any one test method is going to be sensitive to all of the individual contaminants that may be present at a given location. For example, *E. estuarius* is insensitive to copper (McPherson and Chapman 2000), while *N. arenaceodentata* is quite sensitive (Pesch and Morgan 1978). Conversely, *E. estuarius* is more sensitive to acenaphthene than *N. arenaceodentata* (Horne *et al*. 1983, Swartz *et al*. 1995). It is therefore recommended, that at least two of the tests be used together, preferably one acute and one sublethal. By combining the results of two methods, the reliability of the toxicity line of evidence is increased.

Literature Cited

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.

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.

Anderson, B.S., S. Lowe, B.M. Phillips, J.W. Hunt, J. Vorhees, S. Clark and R.S. Tjeerdema. 2008. Relative sensitivities of toxicity test protocols with the amphipods *Eohaustorius estuarius* and *Ampelisca abdita*. *Ecotoxicology and Environmental Safety* 69:24-31.

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.

American Society for Testing and Materials ASTM. 2002c. Standard guide for conducting 10-day static sediment toxicity tests with marine and estuarine amphipods (E1367). pp. 693-719, 2002 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., D. Greenstein and D. 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., M.J. Allen, J. Anderson, D.E. Montagne, J.A. Noblet, J.A. Ranasinghe and S.B. Weisberg. 2003a. Integration of the coastal ecology indications. pp. 91 + 99 Appendices *in*: J.A. Ranasinghe, D.E. Montagne, R.W. Smith, T.K. Mikel, S.B. Weisberg, D. Cadien, R. Velarde and A. Dalkey (eds.), Southern California Bight 1998 regional monitoring program: VII. Benthic macrofauna. 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. 2003b. 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. 2011. Temporal assessment of chemistry, toxicity and benthic communities in sediments at the mouths of Chollas Creek and Paleta Creek, San Diego Bay. Technical Report 668. 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. and F. Wang. 2001. Assessing sediment contamination in estuaries. *Environmental Toxicology and Chemistry* 20:3-22.

California State Water Resources Control Board (CSWRCB). 2009. Water quality control plan for enclosed bays and estuaries-Part 1 sediment quality. California State Water Resources Control Board. Sacramento, CA.

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. and T.S. Bridges. 2011. 28-day chronic sublethal test method for evaluating whole sediments using an early life stage of the marine polychaete Neanthes arenaceodentata. ERDC TN-DOER-R14. US Army Corps of Engineers. Vicksburg, MS.

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.

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.

Greenstein, D., S. Bay, B. Anderson, G.T. Chandler, J.D. Farrar, C.J. Keppler, B. Phillips, A. Ringwood and D. Young. 2008. Comparison of methods for evaluating acute and chronic toxicity in marine sediments. *Environmental Toxicology and Chemistry* 27:933-944.

Horne, J.D., M.A. Swirsky, T.A. Hollister, B.R. Oblad and J.H. Kennedy. 1983. Aquatic toxicity studies of five priority pollutants. Report No. 4389. NUS Corp. Houston, TX.

Hose, G.C., B.R. Murray, M.L. Park, B.P. Kelaher and W.F. Figueira. 2006. A meta-analysis comparing the toxicity of sediments in the laboratory and in situ. *Environmental Toxicology and Chemistry* 25:1148-1152.

Hunt, J.W., B.S. Anderson, B.M. Phillips, J. Newman, R.S. Tjeerdema, R. Fairey, H.M. Puckett, M. Stephensen, R.W. Smith, C.J. Wilson and K.M. Taberski. 2001a. Evaluation and use of sediment toxicity reference sites for statistical comparisons in regional assessments. *Environmental Toxicology and Chemistry* 20:1266-1275.

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. 2001b. 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., R.F. Van Dolah and T.R. Snoots. 1999. Predicting stress in benthic communities of southeastern U.S. estuaries in relation to chemical contamination of sediments. *Environmental Toxicology and Chemistry* 18:2557-2564.

Keddy, C.J., J.C. Greene and M.A. Bonnell. 1995. Review of whole-organism bioassays: Soil, freshwater sediment, and freshwater assessment in Canada. *Ecotoxicology and Environmental Safety* 30:221-251.

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. Society of Environmental Toxicology and Chemistry Annual Meeting. Portland, OR.

Kennedy, A.J., J. Steevens, G.R. Lotufo, J.D. Farrar, M.R. Reiss, R.K. Kropp, J. Doi and T.S. Bridges. 2009. A comparison of acute and chronic toxicity methods for marine sediments. *Marine Environmental Research* 68:118-127.

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. 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., C.B. Hong and C.G. Severn. 2001. Relationships between acute sediment toxicity in laboratory tests and abundance and diversity of benthic infauna in marine sediments: A review. *Environmental Toxicology and Chemistry* 20:46-60.

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.

McPherson, C.A. and P.M. Chapman. 2000. Copper effects on potential sediment test organisms: The importance of appropriate sensitivity. *Marine Pollution Bulletin* 40:656-665.

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 sub-lethal dredged material laboratory bioassays. Society of Environmental Toxicology and Chemistry Annual Meeting. Austin, TX.

Nendza, M. 2002. Inventory of marine biotest methods for the evaluation of dredged material and sediments. *Chemosphere* 48:865-883.

OSPAR. 2007. JAMP guidelines for general biological effects monitoring. Ref. No. 1997-7. OSPAR Commission. London, UK.

Pesch, C.E. and D. Morgan. 1978. Influence of sediment in copper toxicity tests with the polychaete *Neanthes arenaceodentata*. *Water Research* 12:747-751.

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.

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.

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.

Swartz, R.C., D.W. Schults, R.J. Ozretich, J.O. Lamberson, F.A. Cole, T.H. DeWitt, M.S. Redmond, S.P. Ferraro and . 1995. SigmaPAH: A model to predict the toxicity of polynuclear aromatic hydrocarbon mixtures in field-collected sediments. *Environmental Toxicology and Chemistry* 14:1977-1987.

Thursby, G.B., J. Heltshe and K.J. Scott. 1997. Revised approach to toxicity test acceptability criteria using a statistical performance assessment. *Environmental Toxicology and Chemistry* 16:1322-1329.

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

USEPA. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. EPA/600/R-95/136. USEPA 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*. USEPA. Washington, DC.

USEPA. 2006. National estuary program coastal condition report. EPA-842/B-06/001. USEPA Office of Water/Office of Research and Development. Washington, DC.

USEPA and United States 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. USEPA. Washington, DC.

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.

Acknowledgments

The authors would like to thank Arthur Barnett for his guidance on the methodology for this project. They would also like to thank Diana Young, Jeffrey Brown, Doris Vidal-Dorsch and Kerry Ritter for their assistance with data collection and analysis. This project was funded by the California State Water Resources Control Board.