INTERLABORATORY VARIABILITY OF AMPHIPOD SEDIMENT TOXICITY TESTS IN A COOPERATIVE REGIONAL MONITORING PROGRAM

STEVEN M. BAY[1*], ANDREW JIRIK[2], AND STANFORD ASATO[3]

[1]Southern California Coastal Water Research Project, 7171 Fenwick Lane, Westminster, CA, U.S.A.;
[2]Port of Los Angeles, Environmental Management Dept., 425 S. Palos Verdes St., San Pedro, CA, U.S.A;
[3]City of Los Angeles, Environmental Monitoring Div., 12000 Vista del Mar, Playa del Rey, CA, U.S.A.;

[*](author for correspondence, e-mail: steveb@sccwrp.org)

Abstract. Marine sediment toxicity tests are widely applied in monitoring programs, yet relatively little is known about the comparability of data from different laboratories. The need for comparability information is increased in cooperative monitoring programs, where multiple laboratories (often with variable skill levels) perform toxicity tests. An interlaboratory comparison exercise was conducted among seven laboratories in order to document the comparability of sediment toxicity measurements during the Bight'98 regional sediment survey in southern California. Sediments from four stations in Los Angeles and Long Beach Harbors were tested using a 10-day survival test of the amphipod *Eohaustorius estuarius*. All laboratories successfully performed the sediment test and associated reference toxicant test. Statistically significant differences were found in mean amphipod survival rates among some laboratories for the field-collected sediments, but there was little evidence of a consistent bias among laboratories. Although the reference toxicant test indicated a five-fold variation in test sensitivity among laboratories, these results were not accurate predictors of interlaboratory performance for the sediment tests. The laboratories demonstrated excellent concordance (Kendall's W = 0.91) in ranking the field-collected sediments by toxicity. Agreement on classifying the sediments into categories (nontoxic, moderately toxic, and highly toxic) based upon the percent of survival was best for highly toxic sediments. An analysis of test precision based upon the variance among replicates within a test indicated that the measured survival rate for a sample may vary by up to 12 percentage points from the actual response.

Keywords: sediment toxicity, southern California, variability, amphipod, regional monitoring

1. Introduction

Laboratory tests that measure the toxic effects of sediments on benthic organisms, such as amphipod crustaceans, are integral components of national programs to assess sediment quality in coastal areas of the United States (Long, 2000). Amphipod toxicity tests have been used in southern California to examine temporal changes in sediment quality associated with pollution reduction (Swartz *et al*., 1986; Bay, 1992), identify toxic hot spots (Fairey *et al*., 1996), and estimate the spatial extent of sediment toxicity (Anderson *et al*., 1997; Bay *et al*., 1998). Amphipod toxicity tests were an integral part of the Bight'98 regional survey, a cooperative regional survey that included the testing of 241 sediment samples from the Southern California Bight (Figure 1, Bight '98 Steering Committee, 1998).

Cooperative integrated regional monitoring studies conducted in southern California are distinguished by the extensive participation of local agencies in the collection and analysis of samples. Over 60 stakeholder agencies participated in the

planning and sampling activities during the Bight'98 regional survey, and seven local laboratories analyzed the sediment samples for toxicity to amphipods. Such a high level of cooperation in these studies provides great advantages in terms of cost sharing, methods standardization, and communication of the results. The active involvement of diverse agencies also presents a concern regarding the integrity of the data collected, however. Stakeholder agencies may have varying degrees of expertise with the range of sample types or analyses utilized in a regional survey, resulting in a greater chance for technical errors or noncomparable data. Consequently, there is an increased need for method standardization and quality assurance programs in cooperative regional surveys. Similar concerns regarding data comparability also arise when the results of environmental studies from diverse programs are combined for higher level analyses to describe spatial patterns or derive sediment quality guidelines (e.g. Long *et al.*, 1995; 1998).

Information regarding the comparability of sediment toxicity data is especially lacking. Although standard amphipod test methods are available (ASTM, 1992; USEPA, 1994a), few interlaboratory comparison studies have been conducted (Mearns *et al.*, 1986; Schlekat *et al.*, 1995; USEPA, 2001). While these studies demonstrate that amphipod toxicity tests are reliable and reproducible, they may not be applicable to data produced by a cooperative regional study for two reasons. First, highly experienced laboratories participated in the previous studies, which may not be the case in a cooperative study. For example, of the seven Bight'98 toxicology laboratories, only one had previously worked with the specific test species used (*Eohaustorius estuarius*),
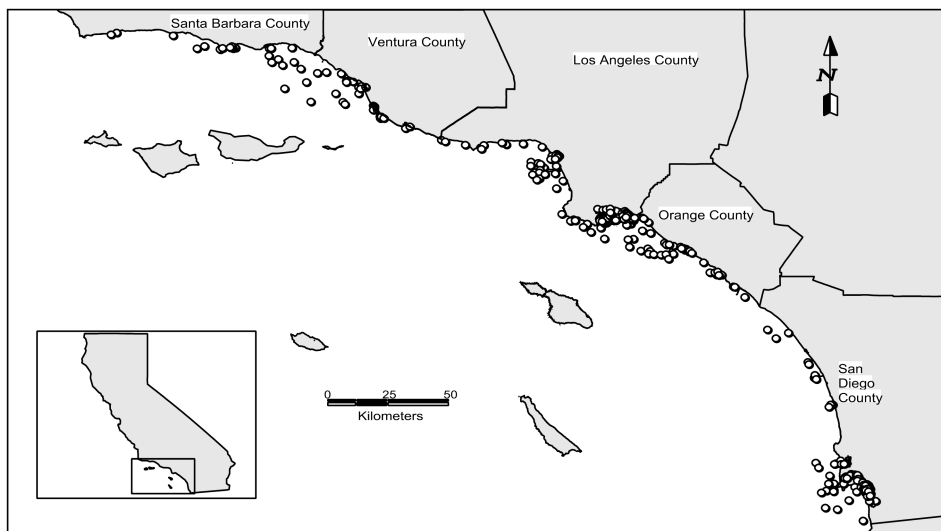
Figure 1. Location of stations analyzed for sediment toxicity during the Bight 98 regional survey of southern California.

and two of the laboratories had not previously conducted sediment toxicity tests with amphipods. Second, most of the existing interlaboratory test data were generated using either spiked or diluted sediment samples. The use of such highly manipulated samples may produce results different from those obtained with field-collected sediments that are handled differently.

This article presents the results of an interlaboratory comparison study that was conducted to accomplish three objectives. The first objective was to determine whether laboratories having limited sediment testing experience were able to perform the 10-day *Eohaustorius* test in accordance with pre-established test acceptability criteria. The second objective was to assess the degree of agreement among laboratories for sediment toxicity samples typical of those encountered in a cooperative regional survey. The third objective was to evaluate whether the participation of multiple laboratories introduced greater variability in test results compared to analyses by a single laboratory.

2. Methods

2.1 SAMPLE COLLECTION AND HANDLING

Sediment samples were collected on May 7, 1998 from four locations in Los Angeles/Long Beach Harbor, California (Figure 2). Station locations were selected based upon prior data to represent a gradient of toxicity ranging from highly toxic ($<50\%$ survival) to moderately toxic (approximately 70% survival). Elevated concentrations of metal and organic contaminants were present at each of the sites,
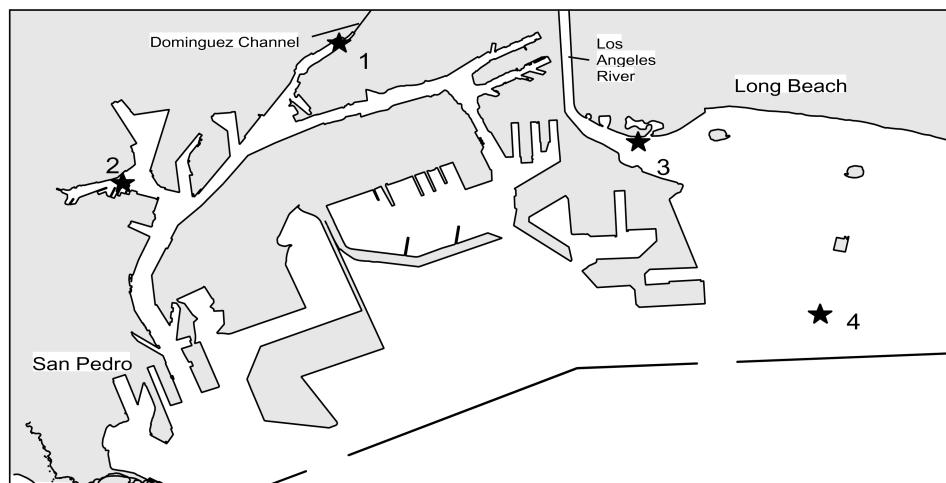
Figure 2. Location of sites sampled in Los Angeles/Long Beach harbor for interlaboratory toxicity comparison.

although the greatest contamination was expected at stations 1 and 3, which receive inputs from large urban stormwater discharges and also from port activities. Station 1 is downstream of a USEPA Superfund site for DDT contamination and also contains elevated concentrations of metals (copper), chlordane, PAHs, and total organic carbon. Station 2 was expected to contain higher concentrations of PCBs (from shipyard operations), relative to the other stations. The principal contaminants at station 3 were metals (lead and zinc) and PAHs. Station 4 was located distant from specific sources of contaminants and was expected to contain the lowest concentrations of contaminants, relative to the other stations. The samples collected for this study were not chemically analyzed.

Sediment samples were collected with a $0.1 \, m^2$ modified Van Veen grab. Multiple grab samples were taken at each station to provide 12 L of surface sediment (top 4 cm of grab). The sediment was placed in polycarbonate containers, chilled on ice, and transported to the laboratory for storage at 4°C in the dark. Each sediment sample was homogenized using an overhead mixer on the following day, split into subsamples, and then distributed to the testing laboratories within 48 hours. The test sediments were kept chilled during transport and storage at the testing laboratories.

2.2 Test Procedures

Sediment toxicity was determined using a 10-day amphipod survival test (USEPA, 1994a; ASTM, 1992). Test organisms, *Eohaustorius estuarius*, were obtained from Beaver Creek, Oregon. The animals were shipped by overnight courier to each laboratory, where they were acclimated under conditions of 20 g/kg salinity, aeration, constant illumination, and 15°C until the initiation of the test on May 12, 1998. Each laboratory received animals from the same collection batch and sorted the specimens to exclude gravid females and specimens outside of the recommended 3-5 mm length range from use. Sediment toxicity tests were conducted in 1 L glass test containers containing a 2 cm layer of sediment. The overlying water was filtered coastal seawater collected by each laboratory. Five replicates were tested for each sediment sample. A negative control (consisting of test animal collection site sediment) was included in each batch of samples tested. At the end of the 10 day exposure period, the sediment was screened through a 0.5 mm-mesh screen and the number of surviving amphipods was recorded.

A cadmium reference toxicity test was conducted concurrently with each sediment toxicity test. The aqueous phase reference toxicant test consisted of three replicates of five concentrations (0.3, 1.0, 3.2, 5.6, and 10.0 mg Cd/L) plus a control sample. The test concentrations were prepared from a common stock solution that was provided to each laboratory. The number of surviving animals was recorded at the end of 4 d and the LC_{50} (Lethal Concentration, 50% mortality) was calculated. Statistical analysis was conducted using the procedures recommended by USEPA (1994a).
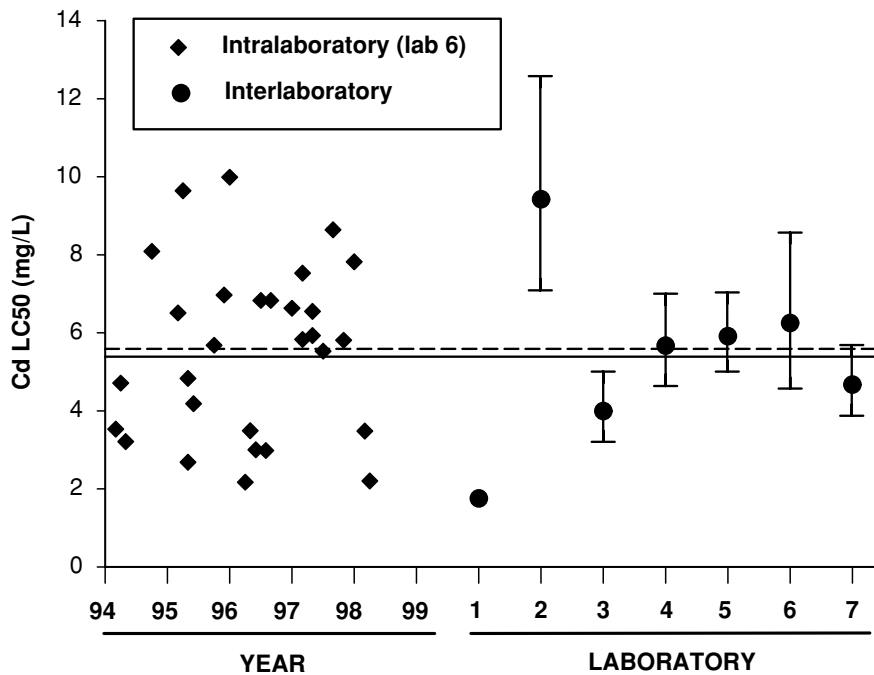
3. Results

All of the participating laboratories met the test acceptability criteria for the sediment and reference toxicant tests, which included (USEPA, 1994a): $\geq 90\%$ survival in controls, water quality parameters within the tolerance range of the test species, and reference toxicant LC_{50} within 2 SD of prior tests. Each laboratory obtained nearly 100% survival in the collection site control sediment and recorded at least 90% survival for the reference toxicant seawater control. The laboratories reported that all of the experiments were conducted within the parameters of the test protocol (including water quality) and that no tests had to be repeated.

3.1 INTERLABORATORY COMPARABILITY

3.1.1 Reference Toxicant

Results of the reference toxicant tests varied approximately five-fold among laboratories. The cadmium LC_{50} for each laboratory (calculated using the Spearman-Karber method) ranged from 1.8 to 9.4 mg/L (Figure 3). The most sensitive test result (lowest LC_{50}) was reported by Laboratory 1; its LC_{50} was significantly less


Figure 3. Cadmium reference toxicant test results from the interlaboratory comparison (circles) and from multiple tests conducted by laboratory 6. Error bars indicate the 95% confidence limits. The 95% confidence limits for laboratory 1 are 1.6–1.9 mg/L. Means for the interlaboratory and intralaboratory data are shown by the solid and dashed lines, respectively.

than the values reported by the other laboratories. The LC_{50} for Laboratory 2 was significantly greater than five of the other laboratories. The 95% confidence limits for the results from the other laboratories overlapped one another, indicating that the results were similar. None of the results were classified as an outlier, since all were within two standard deviations of the mean value (USEPA, 1994a). However, inclusion of the data for Laboratory 1 inflated the standard deviation to 2.4. Laboratory 1 would have been classified as an outlier if the standard deviation based on laboratories 2–7 (1.9) was used to establish the control limits.

The reference toxicant test results reported by the interlaboratory study participants were similar to the results of multiple tests conducted by a single laboratory (Laboratory 6) over a period of several years (Figure 3). Both the intralaboratory and interlaboratory LC_{50} data sets had similar means (5.6 mg/L and 5.4 mg/L, respectively) and ranges (1.8–9.4 and 2.2–10, respectively), indicating that testing by multiple laboratories had little effect on the accuracy or precision of the results.

3.1.2 Sediments

Each of the four sediments was consistently identified as toxic by the laboratories. A statistically significant difference in amphipod survival relative to the control sample (t-test, $p < 0.05$) was obtained for all but one of the 28 samples evaluated in the study (4 sediments x 7 laboratories). High mortality (< 35% mean survival among laboratories) was produced by sediments from two stations, 1 and 3 (Figure 4, Table 1). Sediment from station 1 was most toxic; a mean survival of 11%
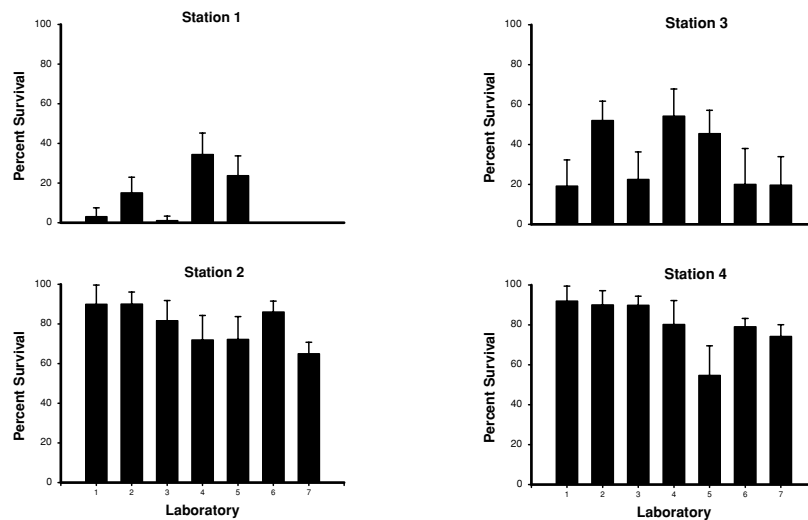
Figure 4. Survival results for *Eohaustorius estuarius* exposed to field sediments. Bars represent the mean (+ 1 standard deviation) of five replicates tested at each laboratory.

Table 1. Summary of toxicity test results for harbor sediment samples.

Station	Percent Survival			Variance	
	Mean	Range	95% CI	Intralab	Interlab
1	11	0–34	±13	41	188
2	79	65–90	±9	80	98
3	33	19–54	±15	180	269
4	80	55–92	±12	74	169

was obtained for this sample and two laboratories reported 0% survival. Low-moderate toxicity was detected in the remaining stations (2 and 4). The laboratories reported 55–92% survival in tests with these samples (Figure 4).

There was little correspondence between the reference toxicant results and the relative response to sediment reported by each laboratory. For example, the reference toxicant results indicated that the amphipod test conducted by Laboratory 1 was significantly more sensitive than the other laboratories' tests, yet the sediment test results for Laboratory 1 were similar to those of most of the other laboratories (Figure 4).

A similar amount of variation among laboratories was present for each sample. Standard deviations for stations 1, 2, 3, and 4 were 14, 10, 16, and 13, respectively. The 95% confidence interval for survival ranged from 9–15% among the stations (Table 1), while an overall confidence interval of ± 12% was calculated based on the grand mean variance of the means for all samples. Thus, the percentage of survival measured by a laboratory is expected to lie within 12 percentage points of the true value 95% of the time.

The variation in the sediment test data attributable solely to interlaboratory variability could not be calculated because repeated analyses of the same sample by a single laboratory were not conducted. As an alternative, the pooled variance among replicates within each laboratory was compared to the pooled variance in mean survival among laboratories for each station. The interlaboratory variance was higher than the replicate variance for each of the four sediment samples (Table 1). Station 1 (the most toxic sample) showed the greatest difference, with the interlaboratory variance being 4.5 times greater than the within-replicate variance. The interlaboratory variance was 1.2 to 2.3 times greater than the replicate variance for the other three stations.

The individual sample results among laboratories were compared using three approaches, each reflecting a different method of data interpretation. For the first approach, the percent of survival results among laboratories was compared. Examination of the survival results indicated that some bias may have been present in the data from two laboratories. Laboratory 7 reported the lowest or second lowest survival rate for each of the sediment samples and the results for Laboratory 2 were within the top three for each sample (Table 2, Figure 4). These differences were not statistically

Table 2. Laboratories arranged by order of survival results. Laboratories not significantly different from one another (Tukey pairwise comparison, p > 0.05) are underscored.

Station	Laboratory Number						
	Highest Survival						*Lowest Survival*
1	4	5	2	1	3	6	7
2	2	1	6	3	5	4	7
3	4	2	5	3	6	7	1
4	1	2	3	4	6	7	5

significant, however. Analysis of variance followed by Tukey pairwise comparisons indicated that the results for Laboratories 2 and 7 were similar to those for most of the other laboratories (Table 2). Significant differences in the percent of survival were detected between laboratories in 25% of the 84 possible pairwise comparisons, but no consistent pattern was observed among laboratories in these differences. Only one case (Laboratory 5 for station 4) was encountered where the results for a laboratory were different from all other laboratories (Table 2).

The second evaluation approach examined the ability of the laboratories to assess the relative toxicity of the four samples. The sediments were assigned ranks based upon the survival results. Each of the seven laboratories ranked stations 1 and 3 as the most toxic and next most toxic samples, respectively (Table 3). Four of the laboratories ranked the sediments in exactly the same order and a fifth only differed in that stations 2 and 4 were tied in the rankings. The Kendall coefficient of concordance based upon these data was 0.91, indicating a high level of agreement (p < 0.01) among laboratories.

The final assessment approach examined the ability of the laboratories to classify the degree of sediment toxicity using response thresholds (nontoxic, moderately toxic, and highly toxic) typical of regional monitoring programs. The agreement among laboratories in classifying the samples varied and was dependent upon the magnitude of toxicity. All seven laboratories classified sediment from station 1 (the most toxic sediment) as highly toxic (Table 3). Relatively consistent results were also obtained for station 3; all seven laboratories classified this station as toxic, with five classifying it as highly toxic and two as moderately toxic. The classification results were more variable for stations 2 and 4; approximately half of the laboratories placed these sediments into the nontoxic category (Table 3). The mean survival rate (among laboratories) for these two stations was 80%, the same as the response threshold used to distinguish between nontoxic and moderately toxic samples.

Table 3. Classification of the test sediments based on either relative toxicity (rank) or response-based categories.

Station	Category by Laboratory[a]							Category by Laboratory[b]						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	4	4	4	4	4	4	4	H	H	H	H	H	H	H
2	2	1.5	2	2	1	1	2	N	N	N	M	M	N	M
3	3	3	3	3	3	3	3	H	M	H	M	H	H	H
4	1	1.5	1	1	2	2	1	N	N	N	M	M	M	M

[a] Rank is based on the relative survival percentage for the four stations tested by the laboratory, with a rank of 4 corresponding to the lowest survival (most toxic). In the case of ties (same survival for two stations) the mean of the ranks involved is assigned to each station (e.g., stations 2 and 4 for laboratory 2).

[b] N = nontoxic, M = moderately toxic (survival = 79–50% of control and significantly different from control), H = highly toxic (< 50% of control survival).

4. Discussion

This study has shown that the *E. estuarius* amphipod toxicity test can be conducted with a high degree of success and reproducibility, even among laboratories with varying levels of experience with the method. Each laboratory met the performance criteria specified by the protocol (USEPA, 1994a) and was able to discriminate statistically between the control sediment and moderately toxic field-collected sediments.

Variable results were obtained among some of the laboratories, as shown by statistically significant differences in the results (percent of amphipod survival) for the same sediment sample. However, the variability measured in this study was similar to that found in previous studies (Mearns *et al.*, 1986; Schlekat *et al.*, 1995). This finding is especially noteworthy considering that several of the laboratories participating in this study had limited or no experience conducting the test procedure. Laboratory 5 had no prior sediment testing experience, Laboratory 3 had no amphipod test experience, and Laboratory 1 had less than one year of amphipod test experience. The results for these laboratories were comparable to those from the remaining laboratories, which had at least five years of amphipod test experience. All of the laboratories had multiple years of aquatic toxicity test experience.

No specific cause was identified for the variability in sediment test results observed among laboratories. The lack of a strong bias in the interlaboratory results, indicated by the similarity of station rankings (Table 3), and the similarity of intralaboratory and interlaboratory reference toxicant results, indicates that laboratory-specific differences in organism sensitivity or test methods were not a major factor. The variability may have been due to factors such as changes in contaminant bioavailability due to sediment storage and handling or may reflect the inherent variability of the test organism response.

While these results indicate that precision within one test is greater than the precision within multiple tests conducted by different laboratories, the data cannot be used to determine whether this variability is introduced by multiple tests or by multiple laboratories. Multiple tests within one laboratory may produce a similar level of variability. A measurement of intralaboratory variability, obtained through the repeated measurement of the same sample, is needed to provide a comparison with the interlaboratory variability measured in this study. Such a study is difficult to conduct with sediments because toxicity may be altered by long-term sediment storage. An alternative approach for estimating intralaboratory variability is to simultaneously test splits of the same sample, using different personnel and batches of test organisms (USEPA, 1994b).

A degree of uncertainty must be expected for any study that employs tests that have variability (Mearns *et al.*, 1986). The significance of this variability depends on the way in which the data are used. If the results are used to rank or otherwise describe the relative toxicity of multiple stations, then there is good agreement among laboratories, especially regarding the identification of the most toxic stations ("hot spots"). For regional assessments, thresholds based upon percentage of survival are often used to identify one or more levels of toxic response (Fairey *et al.*, 1996; Long *et al.*, 1998). Our data, representing contaminated field sediments spanning the range of toxicity typically encountered, indicate that survival measurements are likely to vary among laboratories by approximately 12 percentage points from the mean value. This variability may alter the classification of a sediment sample whose true level of toxicity is near a threshold value. This uncertainty is not a unique problem to toxicity tests and can be minimized by stratified sampling designs that utilize the information from multiple sediment samples to characterize the extent and magnitude of toxicity within a region.

References

ASTM (American Society for Testing and Materials): 1992, Guide for Conducting 10-day Static Sediment Toxicity Tests with Marine and Estuarine Amphipods. E1367-92. ASTM 1992 *Annual Book of Standards*, Vol. 11.04, Philadelphia, PA.

Anderson, B. S., Hunt, J. W., Tudor, S., Newman, J., Tjeerdema, R. S., Fairey, R. S., Oakden, J, Bretz, C., Wilson, C. J., LaCaro, F., Stephenson, M. D., Puckett, H. M., Long, E. R., Fleming, T., Summers, K.: 1997, Chemistry, Toxicity, and Benthic Community Conditions in Selected Sediments of the Southern California Bays and Estuaries. Final Report, State Water Resources Control Board, Sacramento, CA, 157 pp. + app.

Bay, S. M.: 1992, 'Temporal and Spatial Changes in Sediment Toxicity in Santa Monica Bay', in: Cross J. (ed), *Annual Report 1990–91 and 1991–92,* Southern California Coastal Water Research Project, Westminster, CA, pp. 81–87.

Bay, S. M., Greenstein, D. J., Jirik, A. W., Brown, J .S.: 1998, Southern California Bight 1994 Pilot Project: VI. Sediment Toxicity. Southern California Coastal Water Research Project, Westminster, CA, 65 pp.

Bight'98 Steering Committee: 1998, Southern California Bight 1998 regional monitoring survey (Bight'98): Coastal Ecology Workplan. Southern California Coastal Water Research Project, Westminster, CA, 66 pp.

Fairey, R. S., Bretz, C., Lamerdin, S., Hunt, J. W., Anderson, B. S., Tudor, S., Wilson, C. J., LaCaro, F., Stephenson, M .D., Puckett, H. M., Long, E. R.: 1996, Chemistry, Toxicity, and Benthic Community Conditions of Selected Sediments of the San Diego Bay Region. Final Report, State Water Resources Control Board, Sacramento, CA, 179 pp. + app.

Long, E. R.: 2000, 'Degraded Sediment Quality in U.S. Estuaries: A Review of Magnitude and Ecological Implications', *Ecological Applications* **10**, 338–349.

Long, E. R., Field, L. J., MacDonald, D. D.: 1998, 'Predicting Toxicity in Marine Sediments with Numerical Sediment Quality Guidelines', *Environmental Toxicology and Chemistry* **17**, 714–727.

Long, E. R., MacDonald, D., Smith, S., Calder, F.: 1995, 'Incidence of Adverse Biological Effects within Ranges of Chemical Concentrations in Marine and Estuarine Sediments', *Environmental Management* **19**, 81–97.

Mearns, A. J., Swartz, R. C., Cummins, J. M., Dinnel, P. A., Plesha, P., Chapman, P. M.: 1986, 'Interlaboratory Comparison of a Sediment Toxicity Test using the Marine Amphipod *Rhepoxynius abronius*', *Marine Environmental Research* **19**, 13–37.

Schlekat, C. E., Scott, K. J., Swartz, R. C., Albrecht, B., Antrim, L., Doe, K., Douglas, S., Ferretti, J. A., Hansen, D. J., Moore, D. W., Mueller, C., Tang, A.: 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.

Swartz, R.C., Cole, F.A., Schultz, D.W., DeBen, W.A.: 1986, 'Ecological Changes in the Southern California Bight near a Large Sewage Outfall: Benthic Conditions in 1980 and 1983', *Marine Ecology Progress Series* **31**, 1–13.

USEPA (U.S. Environmental Protection Agency): 1994a, Methods for Assessing the Toxicity of Sediment-associated Contaminants with Estuarine and Marine Amphipods. EPA/600/R-94/025. U.S. Environmental Protection Agency, Washington, DC.

USEPA (U.S. Environmental Protection Agency): 1994b, Methods for Measuring the Toxicity and Bioaccumulation of Sediment-Associated Contaminants with Freshwater Invertebrates. EPA/600/R-94/024. U.S. Environmental Protection Agency, Duluth, MN.

USEPA (U.S. Environmental Protection Agency): 2001, Methods for Assessing the Chronic Toxic-
         ity of Marine and Estuarine Sediment-associated Contaminants with the Amphipod
         Leptocheirus plumulosus. EPA/600/R-01/020. U.S. Environmental Protection Agency,
         Washington, DC.