Evaluating consistency of best professional judgment in the application of a multiple lines of evidence sediment quality triad

Steven Bay, Walter Berry[1], Peter Chapman[2],
Russell Fairey[3], Tom Gries[4], Edward Long[5],
Don MacDonald[6] and Stephen B. Weisberg

Abstract

The bioavailability of sediment-associated contaminants is poorly understood. Often, a triad of chemical concentration measurements, laboratory sediment toxicity tests, and benthic infaunal community condition is used to assess whether contaminants are present at levels of ecological concern. Integration of these three lines of evidence is typically based upon best professional judgment by experts; however, the level of consistency among expert approach and interpretation has not been determined. In this study, we compared the assessments of 6 experts who were independently provided data from 25 California embayment sites and asked to rank the relative condition of each site from best to worst. The experts were also asked to place each site into one of six predetermined categories of absolute condition. We provided no guidance regarding assessment approach or interpretation of supplied data. The relative ranking of the sites was highly correlated among the experts, with an average correlation coefficient of 0.92. Although the experts' relative rankings were highly correlated, the categorical assessments were much less consistent, with only 1 site out of 25 assigned to the same absolute condition category by all 6 experts. Most of the observed categorical differences were small in magnitude and involved the weighting of different lines of evidence in individual assessment approaches, rather than interpretation of signals within a line of evidence. We attribute categorical differences to the experts' use of individual best professional judgment and consider these differences to be indicative of potential uncertainty in the evaluation of sediment quality. The results of our study suggest that specifying key aspects of the assessment approach *a priori* and aligning the approach to the study objectives can reduce this uncertainty.

Introduction

Sediment is composed of a complex matrix of constituents that makes interpretation of chemical contamination data challenging. Bulk measures of chemical concentration fail to differentiate between the fraction that is tightly bound to sediment and the fraction available for transport across biological membranes via interstitial water. Furthermore, some benthic organisms ingest sediment and can assimilate chemicals bound onto particles. Thus, even measurement approaches that differentiate interstitial water chemical concentrations, such as equilibrium partitioning models or direct measurement of pore water chemistry, do not fully describe chemical bioavailability in the sediment (Wenning *et al.* 2005).

Consequently, assessments of sediment quality conditions are often conducted by augmenting chemical measurements with toxicity tests and/or measures of benthic infaunal condition. Chemical measurements can be enhanced by toxicity tests that integrate the effects of multiple contaminants. However,

[1] *United States Environmental Protection Agency, Narragansett, RI*
[2] *Golder Associates, North Vancouver, BC, Canada*
[3] *Moss Landing Marine Laboratories, Moss Landing, CA*
[4] *Washington Department of Ecology, Lacey, WA*
[5] *ERL Environmental, Salem, OR*
[6] *MESL, Nanaimo, BC, Canada*

toxicity tests are typically conducted under laboratory conditions and use species that may not occur naturally at the test site, making it difficult to interpret ecological significance of the results when used alone. Benthic community condition is a good ecological indicator because benthic animals readily exhibit impacts of sediment contamination.

Conversely, the use of benthic community condition alone is problematic because the benthos is potentially affected by a diverse battery of noncontaminant variables including: depth, texture, organic carbon content, salinity, dissolved oxygen concentration, currents, tides, and physical habitat disturbances. Benthic conditions are also affected by biotic interactions, such as predation and competition. For these reasons, benthic communities are naturally highly variable.

Habitat measures are often combined into a multiple lines of evidence (MLOE) triad that integrates exposure and effect to assess chemical concentration levels in terms of biological concern (Long and Chapman 1985). Presently, no single, universally accepted method for interpreting triad data and classification of sediments based on an MLOE approach exists (Chapman *et al.* 2002, Wenning *et al.* 2005, Long and Sloane 2005). Each regulatory or monitoring program uses an approach developed through their unique experience. Multiple approaches for integrating triad data have been developed, including: simple logic systems based on presence/absence, statistical summarization, and best professional judgment (Burton *et al.* 2002, Chapman *et al.* 2002, Wenning *et al.* 2005). Regardless of the specific integration approach, most MLOE assessment approaches use some form of BPJ to address uncertainties or conflicts in the data (Chapman and Anderson 2005, Forbes and Calow 2004, Long *et al.* 2005). Expert systems based on fuzzy logic methods have also been used to integrate complex data sets and interpret uncertain results in a consistent fashion (Hollert *et al.* 2002), yet even these approaches must rely on BPJ for the development of classification rules.

While general constructs for interpretation and integration of triad data exist, experts often disagree about the importance of different construct elements, leading to uncertainty about the application of BPJ for sediment quality assessment. In this study, we attempt to quantify this uncertainty by comparing the assessments of six experts who were provided data from triad lines of evidence and asked to classify conditions for a common set of sites, using individually selected approaches and 6 predetermined condition categories.

Methods

We distributed sediment chemistry, sediment toxicity, and benthic infaunal community condition data for 25 sites to 6 experts and asked them to rank the sites from best to worst condition. We also asked the experts to rate them categorically with respect to absolute condition. The experts were selected to represent a diverse range of perspectives and experience. One expert was affiliated with an academic institution, one with a state government that has a sediment quality assessment program in place, two with federal agencies (one retired), and two with private consulting firms that are frequently asked to conduct BPJ assessments. Each of the experts had at least 15 years of experience in conducting assessments of sediment quality, including advising national, state, and local agencies with regards to management and remediation decisions. The experts had also authored numerous reports and peer-reviewed publications regarding sediment quality assessment.

We selected the 25 sites from a California database created for the establishment of standardized sediment quality objectives. Sites were selected from the database by rank ordering them according to overall chemical concentrations based on the respective mean effects range median quotient (ERMq; Long *et al.* 1995, 1998, 2000, and 2006) and then randomly selecting from quartile groups, so that a range of exposure conditions were represented. Twenty-one of the sites were located in euryhaline coastal bays in southern California; four sites were located in polyhaline areas of the San Francisco Bay.

The data provided to the experts for each site included depth, percent sediment fines, percent total organic carbon, chemical concentrations, toxicity, and benthic infaunal condition. We provided chemical concentration data for 11 metals, 21 polycyclic aromatic hydrocarbons (PAHs), chlorinated pesticides (DDTs and chlordanes), and total PCBs (sum of congeners). Summary results from three types of sediment quality guidelines were also provided: mean ERMq, mean SQGQ1 (Fairey *et al.* 2001), and the sum of acute toxic units calculated using the USEPA equilibrium partitioning approach for PAHs and pesticides (USEPA 2003 and USEPA 2004). The toxicity data were from a ten-day *Eohaustorius estuarius* mortality test conducted according to standard methods (USEPA 1994). Because not all of the MLOE experts had familiarity with California benthos, we provided benthic infauna data as a four-category condition assessment developed by consensus of benthic experts (Weisberg *et al.* In press). In addition, benthic species abundance data were made available to the experts upon request.

We asked the experts to rank the relative sediment quality of each site from best to worst and to assign each site to one of six absolute condition categories, using any method of their choice. Although each expert used an approach based on individual experience and BPJ, the absolute condition categories were based on categories under consideration by the State of California for use in statewide sediment quality objectives. Absolute condition categories included:

- Unimpacted. Confident that any sediment contamination at the site is not causing significant adverse direct impacts to aquatic life. The sediment conditions support a benthic community composition that is similar to that attained in reference areas representing the best available conditions in the region. Agreement among the LOE is high.
- Likely Unimpacted. Sediment contamination at the site is not expected to cause significant adverse direct impacts to aquatic life. Some disagreement among the LOE exists, indicating uncertainty in the classification.
- Possibly Impacted. Sediment contamination at the site may be causing adverse impacts to aquatic life, but these impacts may be moderate or variable in nature. LOE agreement with respect to minor levels of effect may exist, or substantial disagreement among the LOE may be present.
- Likely Impacted. Sediment contamination at the site is causing significant adverse direct impacts to aquatic life. Disagreement among the LOE may exist, but the evidence for adverse contaminant-related impact is persuasive.
- Clearly Impacted. Sediment contamination at the site is causing severe adverse direct impacts to aquatic life.
- Inconclusive. Unable to classify the site. Extreme disagreement among the LOE indicates that either the data are suspect or additional information is needed before a classification can be made.

The absolute condition assessments were analyzed in terms of overall disagreement and bias. First, for each expert, we identified the total number of categories for which the expert's categorical assessment of a site differed from the median categorical assessment of all other experts for that site. The number of differences was then summed for all sites to indicate the overall rate of disagreement. Second, we calculated the bias of each expert by incorporating a sign into the sum of the category differences from the median, with a positive sign indicating a more impacted assessment than the median. The bias of each expert was determined as the respective net of positive and negative differences. Sites identified as inconclusive by an expert were excluded from the disagreement and bias calculations for that expert.

Results

We found the relative site rankings to be highly correlated among all the experts, with an average Spearman rank correlation coefficient of 0.92 between experts (Table 1). We also found that there were no experts who deviated notably from their peers; the range in correlation coefficients among the experts was 0.83 - 0.97.

Notably, although the experts were highly correlated with respect to ordinal site rankings, considerable differences in how the experts rated the sites categorically were present. As such, we found that experts disagreed by more than one category for 33% of the sites, categorical agreement among five of the six experts was observed for only 24% of the sites, and complete agreement was obtained for only one site (Table 2).

We attributed this inconsistency to bias among experts, rather than random error (Table 3). For example, Experts 2 and 5 interpreted the toxicity,

Table 1. Spearman rank correlation of site rank among experts. N = 25, all correlations are significant at p <0.001.

	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
Expert 1	0.93	0.97	0.92	0.96	0.94
Expert 2		0.95	0.85	0.94	0.89
Expert 3			0.93	0.93	0.92
Expert 4				0.83	0.87
Expert 5					0.91

Table 2. Categorical site assessment by expert. U = Unimpacted, LU = Likely Unimpacted, PI = Possibly Impacted, LI = Likely Impacted, CI = Clearly Impacted, I = Inconclusive. Shaded boxes indicate sites assigned to impacted categories.

Site	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Median
1	U	U	LU	U	U	U	U
2	LU	PI	PI	PI	LU	LU	PI
3	LU	LU	PI	PI	LU	LU	LU
4	U	LU	LU	LU	U	U	LU
5	LI	PI	LI	PI	LU	LI	LI
6	U	U	LU	U	U	U	U
7	LU	I	PI	I	LU	I	LU
8	LI	I	LI	I	PI	I	LI
9	PI	I	LI	PI	LU	I	PI
10	LI	PI	LI	CI	PI	LI	LI
11	CI	PI	LI	CI	LI	CI	CI
12	PI	LU	PI	PI	LU	LU	PI
13	PI	PI	PI	PI	LU	PI	PI
14	LI	PI	LI	PI	LI	LI	LI
15	CI	PI	LI	PI	LI	LI	LI
16	PI	LU	PI	I	U	PI	PI
17	PI	LU	PI	LI	U	PI	PI
18	U	U	U	U	U	U	U
19	CI	PI	CI	CI	CI	CI	CI
20	CI	LI	CI	CI	CI	CI	CI
21	CI	LI	CI	CI	CI	CI	CI
22	CI	LI	CI	CI	CI	CI	CI
23	U	U	LU	I	U	U	U
24	U	U	LU	I	U	U	U
25	U	U	LU	I	U	U	U

chemistry, and benthic community data more leniently, ranking many sites as less impacted than their peers. In contrast, Expert 3 consistently interpreted these indicators more severely than the other experts. Large differences in classification among the experts were infrequent. Among all possible pairwise comparisons, the experts' assessments differed by more than one category in only 7% of the site pairs.

In addition to disagreement regarding categorical ranking, the experts also disagreed about which and how many sites should be classified as inconclusive. Seven of the sites were classified as inconclusive by at least one expert, whereas only two sites were classified as inconclusive by at least three experts. Three experts listed no sites as inconclusive, and Expert 4 assigned 24% of the sites to this category (Table 2).

Discussion

Subsequent to receiving their site ratings, we interviewed the experts to understand individual assessment processes. While all of the experts integrated data from MLOE to rank and classify the sites, each expert used a different specific approach based on respective philosophy and experience in relation to the constraints of the data set. Some of the experts used a numeric approach that integrated scores or ranks based on levels of response within an LOE, while others based their classifications on more subjective comparisons of concordance and relative magnitude among the LOEs (Table 4). Despite these considerable differences in approach, we observed substantial similarity in outcomes.

Table 3. Summary of categorical assessments for each expert. Differences in the number of sites are due to the exclusion of sites classified as inconclusive. Disagreement values represent the total number of category differences between the expert's assessment and the median of all other experts' assessments. Bias values reflect the net of positive or negative assessment differences.

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
# Sites	25	22	25	19	25	22
Disagreement	7	16	13	10	15	5
Bias	1	-12	11	4	-15	-1

Most of the differences in site classification were due to differing philosophies with respect to the weighting of the three lines of evidence. Most of the experts placed the greatest emphasis on the benthic community condition, but differed in application of the data. Expert 4 used benthic condition as a low trigger threshold, a way to eliminate sites for which no effect was observed, and placed greater emphasis on the toxicity and chemistry data to assess the biological impact of chemical exposure. Others used benthos as the primary means for assessment, considering it the endpoint of ultimate interest, and then used the chemistry and toxicity data as a means for assessing the likelihood that an effect had been chemically mediated. Expert 2 chose not to use the chemistry data at all, considering toxicity to be a better measure of exposure than chemistry data given the number of potentially unmeasured chemicals in a typical sediment screen and the inability of routine sediment chemical analysis to describe variations in contaminant bioavailability.

All of the experts agreed that it was critical to demonstrate a linkage between chemical exposure and biological effects in order to classify a station as

Table 4. Key attributes of the assessment approaches used by experts to classify study sites.

Expert	LOE Used	Chemistry SQG Use[1]	Toxicity Evaluation	Relative Benthos Weight	Assessment Approach
1	All	ERMq, SQGQ1	MSD[2]	Greater	LOE Concordance, Response Magnitude, Sum of LOE Scores
2	Benthos Toxicity	None	MSD	Greater	LOE Concordance, Response Magnitude, Limitations in Toxicity Data
3	All	LRM, EqP	Magnitude[3]	Equal	Average of LOE Scores
4	All	ERMq, SQGQ1	MSD	Greater	LOE Concordance, Response Magnitude
5	All	SQGQ1	MSD	Equal	LOE Concordance, Response Magnitude
6	All	ERMq, SQGQ1, EqP, AET	MSD	Greater	LOE Concordance, Response Magnitude, Potential for Physical Effects on Benthos

[1] ERMq: mean quotient of Effects Range-Median values; SQGQ1: mean quotient of various SQGs; LRM: probability of toxicity from logistic regression models; EqP: acute toxic units from PAHs and chlorinated pesticides; and AET: apparent effects threshold.

[2] Minimum significant difference of 20% survival relative to control.

[3] Evaluation based on magnitude of 90-100%, 80-90%, 65-80%, 50-65%, and < 50% survival relative to control.

impacted due to contamination. Making this link is important for two reasons: (1) to distinguish between chemical and other causes of any effects and (2) to provide information on specific chemical causes that can be used to determine sources and management actions (Chapman 2007). The assessment conducted by the experts in this study was intended to address the first application (i.e., determining if impacts related to contamination were present). This is the primary information needed in the first steps of a site assessment. Determining the specific chemicals responsible for impacts often requires a larger data set and different analytical approaches (e.g., toxicity identification evaluations, contaminant bioavailability analyses) in order to discriminate among the complex mixture of contaminants present in most locations and determine those most likely responsible for effects.

Few classification differences were due to the assessment approach for individual LOE. Each expert based assessments of toxicity on the magnitude of survival, with most using a minimum significant difference criterion of 80%, which was based on the statistical power of the amphipod test to detect differences from the control (Phillips *et al.* 2001 and Thursby *et al.* 1997). For chemistry, where there are numerous methods for assessing magnitude of chemical exposure, reasonable agreement among five of the experts regarding classification of data from this LOE was noted. The five experts relied on a type of collective empirically-based approach, such as mean ERMq or logistic regression (Field *et al.* 2002), as a primary means of assessment, although the experts varied in the particular approaches used (Table 4). Three of the five experts augmented this approach with examination of individual chemical concentrations based on empirically derived thresholds.

The greatest difference among the experts was in the use of the organics equilibrium partitioning (EqP) results as supplemental information (Table 4), particularly as EqP assessments sometimes conflicted with those conducted using empirical guidelines. These differences prompted one expert to classify some of the sites as inconclusive due to uncertainty regarding the biological significance of the chemical exposure. Most of the experts tended to downplay EqP values because the values were more likely to conflict with other LOE than empirical approaches. However, a few experts indicated that they downplayed the use of EqP because of inconsistencies in the data (i.e., not all EqP constituents were analyzed at all sites). Although complete data would have been preferable, similarly minor inconsistencies were observed for the empirical threshold assessments. Consequently, the data used in this study are representative of data collated from multiple studies for the development of an integrated regional assessment.

Several of the experts indicated that a more complete chemistry dataset would have aided the assessment in two respects. First, the inclusion of information on sediment factors affecting contaminant bioavailability, such as sediment acid volatile sulfides (AVS) and the concentration of black carbon would have enabled Expert 2 to include the chemistry data in the assessment and may have assisted the other experts in resolving some of inconsistencies in the data that led to inconclusive results. Second, the inclusion of data for additional analytes of concern (e.g., current use pesticides) would have given some experts greater confidence in their assessment decisions with respect to the chemistry and benthos lines of evidence by indicating whether additional potential toxicants were present. The lack of established assessment guidelines for these additional chemicals is problematic, but mechanistic models and effects data from the literature can be used to assist in interpretation when using best professional judgment.

The data limitation that concerned the experts most was the availability of only a single toxicity test. Many of the inconclusive site classifications were based on inconsistencies between the toxicity and benthos responses, raising uncertainty regarding the contribution of physical and chemical factors in the assessment of benthic community condition. Several experts indicated that they would have relied on toxicity testing more to resolve these inconclusive findings if data from additional toxicity tests had been available. Notably, at least two contrasting concerns were expressed about use of a single test. On one level, the experts were concerned about false positives (a toxicity test response incorrectly assumed to be caused by chemical contamination) due to effects associated with sediment handling. On an entirely different level, experts expressed concern about differential sensitivity among test species and the possibility that a lack of sublethal endpoints could lead to false negatives.

The inclusion of additional toxicity tests would have improved the experts' confidence in their assessments, provided that the additional tests incorporated additional pathways of exposure (e.g., interstitial water), sublethal endpoints (e.g., growth and fecundity), or longer exposure durations. The additional results would have verified that toxicity was not due to confounding factors (e.g., sediment particle size or ammonia), and provided greater assurance that the presence of toxicity was not overlooked due to the choice of a single test method that was not responsive to the contaminants or pathways of exposure present at the site.

The benthic ecology data played an important role in determining the site assessments in this study, but most experts also identified uncertainties with interpreting the data. The greatest source of uncertainty was related to the potential for benthic infaunal community composition to be affected by habitat or physical factors and the inability to distinguish such alterations from contaminant effects. Additional sources of uncertainty are seasonal changes due to reproduction, changes in natural assemblage characteristics among habitats or geographic regions, and the lack of consistent methods of interpreting species abundance data. We reduced these uncertainties in this study by providing each expert by with a benthic community assessment based on the consensus of benthic experts, which controlled for many of these factors. This approach may not be feasible in other studies, and additional steps may be needed to reduce uncertainty such as identifying the distribution of major benthic community assemblages, restricting analyses to a specific time of year, and developing indices or other statistical approaches to interpret the data.

The experts also indicated that the ambiguity of the predetermined category definitions, leading to differences in interpretation, was another potential source of disagreement. Specifically, they expressed concern that the assessment categories confounded several factors: confidence that there is an effect, magnitude of the effect, and likelihood that the effect is chemically mediated. This uncertainty is particularly evident where their disagreements occurred along the classification gradient. Two of the classification categories represented "unimpacted" conditions and three represented "impacted" conditions. The rate of disagreement among the experts across this condition boundary was less than disagreement for classification gradations on either side of the boundary. Notably, we observed complete agreement among the experts with respect to this boundary for 16 of the 25 sites (Table 2). Thus, although the experts often disagreed about the magnitude or certainty assigned to an individual site classification, they rarely disagreed about whether a site was impacted or unimpacted.

Overall, we found the use of BPJ in the integration of a MLOE triad to be a significant source of variation in the evaluation of sediment contaminant exposure and its environmental impacts. Differences among the experts regarding assessment approach, LOE weighting, and indicator interpretation reveals an important source of uncertainty that should be considered in conducting ecological risk assessments. The significance of these results for making management decisions depends upon the nature of the question. The impact on large scale assessments where the objective is to identify the worst locations or describe the relative condition of sites is likely to be small, as there was good agreement among the experts in terms of overall condition classification and relative site ranking. The impact will be more significant with respect to making management decisions for specific sites, particularly those with intermediate levels of contamination, toxicity, or biological alteration, as these sites may be variously classified as Likely Unimpacted (no remediation needed), Inconclusive (more data needed), or Likely Impacted (potential remediation).

This study was limited in scope in that only six experts were involved and sediments represented primarily marine locations within California. While we feel that the results are generally applicable to other habitats and regions, the specific amounts of disagreement and bias reported here may change as a function of the number of participants and their level of expertise. Conducting a follow up study that included a more complete data set from a greater diversity of habitats would strengthen the conclusions from this exercise.

Several steps are recommended in order to reduce the uncertainty associated with the integration and interpretation of Sediment Quality Triad data. First, key elements of the assessment strategy, such as the relative weight of each LOE, how multiple LOE will be combined (e.g., scores, ranks, logic frameworks), and the criteria for determining the assessment conclusion should be determined during the design of the study. Second, comparability among studies can be improved by providing guidance on specific methods for measuring sediment chemistry (e.g., analyte list, detection limits, how sediment quality guidelines are used), sediment toxicity (e.g., test methods, toxicity classification thresholds), and benthic community condition (e.g., which metrics or indices to use, criteria for determining impacts). Finally, uncertainty in sediment quality assessment can be reduced through improved training of the individuals interpreting the data. While each expert participating in this study had extensive experience with interpreting sediment quality data, the expertise of personnel at state and local agencies responsible for conducting or interpreting sediment quality assessments is highly variable and can lead to different interpretations of the same data set. This situation can be remedied through enhanced technology transfer and training activities, such as the sponsorship of short courses in sediment quality assessment and the preparation of guidance documents by international scientific organizations such as the Society of Environmental Toxicology and Chemistry.

Literature Cited

Burton, Jr., G.A., PM. Chapman and EP. Smith. 2002. Weight of evidence approaches for assessing ecosystem impairment. *Human and Ecological Risk Assessment* 8:1657-1673.

Chapman, P.M. 2007. Do not disregard the benthos in sediment quality assessment! *Marine Pollution Bulletin* 54:633-635.

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

Chapman, P.M., B.G. McDonald and G.S. Lawrence. 2002. Weight-of-evidence issues and frameworks for sediment quality (and other) assessments. *Human and Ecological Risk Assessment* 8:1489-1515.

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

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

Forbes, V.E. and P. Calow. 2004. Systematic approach to weight of evidence in sediment quality assessment: Challenges and opportunities. *Aquatic Ecosystem Health and Management* 7:339-350.

Hollert, H., S. Heise, S. Pudenz, R. Brüggemann, W. Ahlf and T. Braunbeck. 2002. Application of a sediment quality triad and different statistical approaches (Hasse diagrams and fuzzy logic) for the comparative evaluation of small streams. *Ecotoxicology* 11:311-321.

Long, E.R. and P.M. Chapman. 1985. A sediment quality triad - measures of sediment contamination, toxicity and infaunal community composition in Puget-Sound. *Marine Pollution Bulletin* 16:405-415.

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

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

Long, E.R., C.G. Ingersoll and D.D. MacDonald. 2006. Calculation and uses of mean sediment quality guideline quotients: A critical review. *Environmental Science and Technology* 40:1726-1736.

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

Long, E.R. and G.M. Sloane. 2005. Development and use of assessment techniques for coastal sediments. pp. 63-78 *in*: S.A. Bortone (ed.); Estuarine Indicators. CRC Press. Boca Raton, FL.

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

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. United States Environmental Protection, Office of Research and Development Agency. Narragansett, RI.

USEPA. 2003. Procedures for the derivation of equilibrium partitioning sediment benchmarks (ESBs) for the protection of benthic organisms: PAH mixtures. EPA-600-R-02-013. United States Environmental Protection Agency. Washington, DC.

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

Weisberg S.B., B.E. Thompson, J.A. Ranasinghe, D.E. Montagne and D.B. Cadien. In press. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. *Ecological Indicators*.

Wenning R., G. Batley, C. Ingersoll, M. Moore and editors. 2005. Use of sediment quality guidelines and related tools for the assessment of contaminated sediments. Society of Environmental Toxicology and Chemistry (SETAC). Pensacola, FL.

Acknowledgments

The authors thank Doris Vidal-Dorsch, Jeff Brown, and Ananada Ranasinghe of the Southern California Coastal Water Research Project for assistance with data compilation and statistical analysis. Work on this project was funded by the California State Water Resources Control Board under agreement 01-274-250-0.