A framework for interpreting sediment quality triad data

Steven M. Bay and Stephen B. Weisberg

ABSTRACT

There are numerous approaches for integrating multiple lines of evidence (MLOE) data in a sediment quality triad assessment, but most rely at least partially on best professional judgment (BPJ), which can be problematic in application to large data sets or in a regulatory setting where the assessment protocol needs to be transparent and consistently reproducible. This study presents an approach for standardizing triad-based assessments and evaluate the extent to which it captures and reproduces the assessments of experts employing BPJ on the same data. The framework is based on integrating answers to two questions: 1) Is there biological degradation at the site, and 2) Is chemical exposure at the site high enough to potentially result in a biological response? The efficacy of the framework was assessed by applying it to data from 25 sites and comparing the site classifications to those of six experts who were provided the same data. The framework produced an answer that better matched the median classification of the experts than did five of the six experts. Moreover, the bias in response was less than that obtained from some of the experts, and the errors were relatively evenly divided between sites classified as more impacted or less impacted than the median expert classification. The framework was also applied and found to distinguish well sites from known degraded and reference areas within California. While the framework suggested here is not the only one possible and should be supplemented with BPJ when additional data beyond that included in the framework are available, the framework provides a validated means for using a triad based approach in large-scale assessments, such as those for Clean Water Act (CWA) 305b programs or regulatory decisions, where transparency in the decision process is critical.

INTRODUCTION

Assessments of sediment quality for the effects of chemical contamination frequently use a triad of chemical concentration, sediment toxicity, and benthic infaunal community condition data (Long and Chapman 1985). These are used in combination because sediments are a complex matrix and chemical concentration data alone fails to differentiate between the fraction that is tightly bound to sediment and that which is biologically available. Toxicity tests improve on chemical measurements because they integrate the effects of multiple contaminants, but toxicity tests are typically conducted under laboratory conditions using species that may not occur naturally at the test site and with a range of sensitivities, making it difficult to interpret the ecological significance of the results when used alone. Benthic community condition is a more direct ecological indicator because these are resources at risk from sediment contamination, but their use alone is problematic because they are potentially affected by other factors, such as eutrophication, physical disturbance or hypoxia.

Several approaches for integrating these multiple lines of evidence (MLOE) data have been developed (Chapman *et al*. 2002). These integration approaches rely mostly on a similar suite of indicators for each LOE, but differ in how the LOEs are combined into a single assessment. Some are based on combinations of binary responses for each LOE, while others use a more complex statistical summarization. For example, some approaches weight the three LOEs equally, while others weight them differently. Further, even within an integration framework, thresholds must be determined for each LOE. Consensus thresholds for these most commonly used LOEs do not yet exist, with the result that these threshold decisions become particularly important when the integration is based on a binary decision for each LOE.

As a result of these current limitations, most triad applications rely on some degree of BPJ (Burton *et al*. 2002, Chapman and Anderson 2005). Despite the many decisions inherent in the integration of LOEs, BPJ has been found to be reasonably repeatable for interpretation of triad data (Bay *et al*. 2007). Thus, BPJ can be an acceptable means of integration for site-specific assessments, but it is not easily applicable to large-scale assessments where many sites are involved. It is also problematic in a regulatory setting, where the assessment protocol must be transparent and consistently reproducible over time and space (Forbes and Calow 2004). The State of California is developing a framework for standardizing such triad-based assessments as part of establishing sediment quality objectives. The present study describes that framework and evaluate the extent to which it captures and reproduces the assessments of experts employing BPJ on the same data.

Methods

Integration Framework

The framework integrates three lines of evidence (LOE) to assess sediment quality at a site (i.e., specific location or station). Integration of the data involves a three-step process (Figure 1). First, the response for each LOE is assigned into one of four response categories: 1) no difference from background conditions, 2) a small response that might not be statistically distinguishable from background conditions, 3) a response that is clearly distinguishable from background, and 4) a large response indicative of extreme conditions.

Second, the individual LOEs are combined to address two key elements of a risk assessment paradigm: 1) Is there biological degradation at the site? and 2) Is chemical exposure at the site high enough to potentially result in an adverse biological response? To answer the first question, the benthos and toxicity

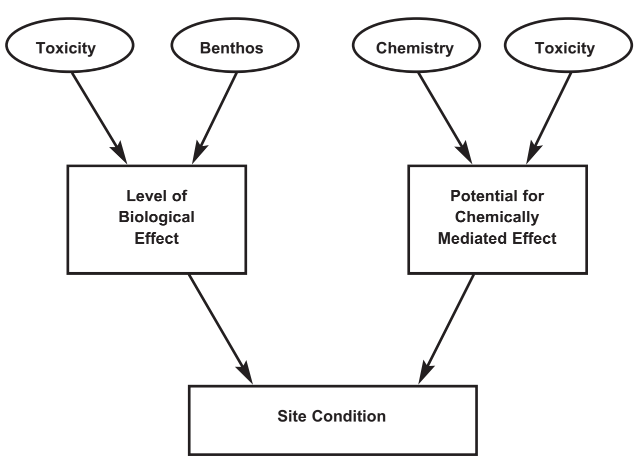

Figure 1. Conceptual model for the integration of MLOEs in the assessment framework.

LOEs are integrated to assess the severity of effects (Table 1). The effects assessment is equivalent to the benthic condition in most cases, except where there is extreme disagreement between the benthos and toxicity LOEs. Benthos is given greater weight in this assessment as it is the ultimate endpoint of interest (Chapman 2007). The second question arises because the biological response may be attributable to factors other than chemical contaminants. This framework is intended to assess impacts on sediment quality due to anthropogenic chemical contamination, not impacts from physical or biological processes. The potential that effects are chemically mediated is assessed using the sediment chemistry and toxicity LOEs (Table 2). Chemistry is the more direct measure, but toxicity is given equal weight because of the potential that unmeasured chemicals are present and because of uncertainties in thresholds used to interpret chemical data (Ingersoll *et al*. 2005).

The final data integration step combines the severity of effect and potential for chemically-mediated effects to assign a site into one of six impact categories:

- Unimpacted. Confident that contamination is not causing significant adverse impacts to aquatic life living in the sediment at the site.
- Likely Unimpacted. Contamination is not expected to cause adverse impacts to aquatic life in the sediment, but some disagreement among LOEs reduces certainty that the site is unimpacted.
- Possibly Impacted. Contamination at the site may be causing adverse impacts to aquatic life in the sediment, but the level of impact is either small or is uncertain because of disagreement among LOEs.
- Likely Impacted. Evidence of contaminant-related impacts to aquatic life in the sediment is persuasive, in spite of some disagreement among LOEs.
- Clearly Impacted. Sediment contamination at the site is causing clear and severe adverse impacts to aquatic life in the sediment.
- Inconclusive. Disagreement among the LOE suggests that either the data are suspect or that additional information is needed before a classification can be made.

The decision process for determining the site

Table 1. Severity of effect classifications, derived from the benthos and toxicity LOEs.

Benthos LOE Category	Toxicity LOE Category			
	Nontoxic	Low Toxicity	Moderate Toxicity	High Toxicity
Reference	Unaffected	Unaffected	Unaffected	Low Effect*
Low Disturbance	Unaffected	Low Effect	Low Effect	Low Effect
Moderate Disturbance	Moderate Effect	Moderate Effect	Moderate Effect	Moderate Effect
High Disturbance	Moderate Effect*	High Effect	High Effect	High Effect

*Extreme disagreement between LOE is present indicating atypical conditions or suspect data. Review of additional information about the site before making an assessment is recommended.

assessment category is based on a foundation that there must be some evidence of biological effect in order to classify a site as impacted (Table 3). Additionally, there must be some evidence of elevated chemical exposure in order to classify a site as chemically-impacted.

Application of the Framework

Application of the framework requires measuring sediment chemistry, toxicity, and benthic community condition at each site using standardized methods. The response of each measurement is compared to established thresholds to categorize each of the individual LOEs into one of four possible response categories (Table 4).

Chemistry

A combination of two sediment chemistry indices was used to determine the magnitude of chemical exposure at each site: the logistic regression models calibrated to California data (CA LRM; Bay *et al*. 2008) and the Chemical Score Index (CSI; Ritter *et al*. 2008). The CA LRM was developed using the logistic regression modeling approach that estimates the probability of toxicity based on the chemical concentration (Field *et al*. 2002, USEPA 2005). The CSI uses the chemistry data to predict the occurrence and severity of benthic community disturbance. Index-specific thresholds were applied to each index to classify the result into one of four chemical exposure categories: Minimal, Low, Moderate, and High. The resulting exposure categories were assigned a score of 1 - 4 (e.g., Minimal Exposure = 1) and the average of the scores for each chemistry index was used to determine the overall chemistry LOE category. Average scores were rounded up to the next whole number in the case of intermediate results (e.g., average score of 2.5 = Moderate Exposure).

Toxicity

The results of multiple toxicity tests are used to determine the magnitude of sediment toxicity at each site. The tests must include a 10-day amphipod survival test and a sublethal test (e.g., embryo development or juvenile growth). Thresholds based on per-

Table 2. Potential that effects are chemically mediated classifications, derived from chemistry and toxicity LOE.

Chemistry LOE Category	Toxicity LOE Category			
	Nontoxic	Low Toxicity	Moderate Toxicity	High Toxicity
Minimal Exposure	Minimal Potential	Minimal Potential	Low Potential	Moderate Potential*
Low Exposure	Minimal Potential	Low Potential	Moderate Potential	Moderate Potential
Moderate Exposure	Low Potential	Moderate Potential	Moderate Potential	Moderate Potential
High Exposure	Moderate Potential*	Moderate Potential	High Potential	High Potential

*Extreme disagreement between LOE is present indicating atypical conditions or suspect data. Review of additional information about the site before making an assessment is recommended.

Table 3. MLOE site classifications, derived from intermediate classifications described in Tables 1 and 2.

Potential for Chemically Mediated Effects	Severity of Effect			
	Unaffected	Low Effect	Moderate Effect	High Effect
Minimal Potential	Unimpacted	Likely Unimpacted	Likely Unimpacted	Inconclusive
Low Potential	Unimpacted	Likely Unimpacted	Possibly Impacted	Possibly Impacted
Moderate Potential	Likely Unimpacted	Possibly Impacted or Inconclusive*	Likely Impacted	Likely Impacted
High Potential	Inconclusive	Likely Impacted	Clearly Impacted	Clearly Impacted

*Inconclusive category applies when: chemistry = minimal exposure, benthos = reference, and toxicity = high. Other LOE combinations represented by this cell are classified as Possibly Impacted.

centage survival and statistical significance are applied to classify the test results into one of four toxicity categories (Bay *et al*. 2007a): Nontoxic, Low, Moderate, and High. The toxicity categories are averaged as described previously to determine the overall toxicity LOE category.

Benthos

A combination of four benthic community condition indices was used to determine the magnitude of disturbance to the benthos at each site (Ranasinghe *et al*. 2007). The indices include approaches based on community metrics and abundance of individual species. The benthic indices are:

Benthic Response Index (BRI), which was originally developed for the southern California mainland shelf and extended into California's bays and estuaries (Smith *et al*. 2001, Smith *et al*. 2003). The BRI is the abundance-weighted average pollution tolerance score of organisms occurring in a sample.

Index of Benthic Biotic Integrity (IBI), which was developed for freshwater streams and adapted for California's bays and estuaries (Thompson and Lowe 2004). The IBI identifies community measures that have values outside a reference range.

Table 4. Ranges of values used to define each LOE indicator category. Separate benthic index ranges are used for southern California and San Francisco Bay habitats.

LOE	Indicator	Category			
Chemistry Exposure		**Minimal**	**Low**	**Moderate**	**High**
	CA LRM	<0.33	≥0.33 to <0.49	≥0.49 to ≤0.66	>0.66
	CSI	<1.69	≥1.69 to ≤2.33	>2.34 to ≤2.99	>2.99
Toxicity		**Nontoxic**	**Low**	**Moderate**	**High**
	Survival (%)	≥90	<90 to ≥82	<82 to ≥59	<59
Benthos Disturbance		**Reference**	**Low**	**Moderate**	**High**
Southern California	BRI	<33	≥33 to <51	≥51 to <70	≥70
	IBI	0	1	2	≥3
	RBI	>0.27	≤0.27 to 0.16	≤0.16 to 0.07	≤0.07
	RIVPACS	≥0.9 to <1.1	≥0.74 to <0.89	>0.31 to <0.74	≤0.31
			≥1.1 to <1.27	≥1.27	
San Francisco Bay	BRI	<22.3	≥22.3 to <33.4	≥33.4 to <82.1	≥82.1
	IBI	≤1	2	3	4
	RBI	>0.43	≤0.43 to >0.29	≤0.29 to >0.19	≤0.19
	RIVPACS	≥0.68 to <1.32	≥0.32 to <0.68	>0.15 to <0.32	≤0.15
			≥1.32 to <1.68	≥1.68	

Relative Benthic Index (RBI), which was originally developed for California's Bay Protection and Toxic Cleanup Program (Hunt *et al*. 2001). The RBI combines several parameters: a) several community metrics, b) the abundances of three positive indicator species, and c) the presence of two negative indicator species.

River Invertebrate Prediction and Classification System (RIVPACS), which was originally developed for British freshwater streams (Wright *et al*. 1993, Van Sickle *et al*. 2006) and adapted for California's bays and estuaries. The RIVPACS index calculates the number of reference taxa present in the test sample and compares it to the number expected to be present in a reference sample from the same habitat.

Thresholds specific to regional assemblages were applied to the results in order to classify each benthic index result into one of four disturbance categories: Reference, Low, Moderate, and High. The median of categories for each individual index was used to provide an overall benthos LOE category.

The response category results for each LOE were used to determine the assessment category for each site, based on the relationships shown in Tables 1 - 3. The site category corresponding to each of the 64 possible combinations of the LOE results is shown in Table 5.

Evaluation of the Framework

The efficacy of the framework was assessed in two ways. The first was by applying it to data from 25 sites and comparing the site classifications to those of six experts provided the same data. The experts were selected to represent a diverse range of sectors (industry, academia, government), with each having at least 15 years of experience in conducting sediment quality assessments and advising national, state, and local agencies with regards to sediment management and remediation decisions (Bay *et al*. 2007). The experts were asked to classify the sites into one of the six categories of absolute condition described above.

The 25 sites were selected from a California database by rank ordering them according to overall chemical concentrations based on the mean effects range median quotient (ERMq; Long *et al*. 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 included depth, percent sediment fines, percent total organic carbon, chemical concentrations, toxicity, and benthic infaunal condition. The chemical concentration data were for 11 metals, 21 polycyclic aromatic hydrocarbons (PAHs), chlorinated pesticides (DDTs and chlordanes), and total PCBs (sum of 16 congeners). 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, benthic infauna data were provided as a four-category condition assessment developed by consensus of benthic experts (Weisberg *et al*. 2008); the benthic species abundance data were also made available on request.

Agreement between the experts and the framework was quantified in two ways. First, the number of impact categories for which the site assessment derived using the framework differed from the median categorical assessment of the experts was calculated for each site and then summed across sites to indicate the overall rate of disagreement. Second, the bias of the framework was quantified as the net of positive and negative differences from the median expert, calculated by incorporating a sign into the sum of the category differences from the median. For perspective, the framework's agreement with the median of the experts' results was compared to the agreement of each of the individual experts with the median of the other experts.

The second evaluation approach involved applying the framework to geographic regions that have previously been designated as toxic hotspots by the State of California and comparing those data with results when applying the framework to reference areas. The hotspot regions were identified by the State's Bay Protection and Toxic Cleanup Program (BPTCP) as the worst in the state based on a BPJ assessment of sediment chemistry, toxicity and benthic community condition (Anderson *et al*. 2001, Fairey *et al*. 1998, Phillips *et al*. 1998). The reference sites were selected from areas that were distant from known sources of contamination (e.g., outer portion of embayments) and for which previous surveys had consistently shown low toxicity (defined as >80 % amphipod survival) and low chemistry

Table 5. Relationship of individual LOE categories to chemical exposure, biological effects, and final MLOE site assessment categories. Arrows indicate the sequence of classification.

Toxicity	Chemistry	Chemically Mediated Effects	Site Assessment	Biological Effects	Benthos	Toxicity
High	High	High Potential	Clearly Impacted	High Effect	High	High
Moderate	High	High Potential	Clearly Impacted	High Effect	High	Moderate
High	High	High Potential	Clearly Impacted	Moderate Effect	Moderate	High
Moderate	High	High Potential	Clearly Impacted	Moderate Effect	Moderate	Moderate
High	High	High Potential	Likely Impacted	Low Effect	Low	High
Moderate	High	High Potential	Likely Impacted	Low Effect	Low	Moderate
High	High	High Potential	Likely Impacted	Low Effect	Reference	High
High	Moderate	Moderate Potential	Likely Impacted	High Effect	High	High
Moderate	Moderate	Moderate Potential	Likely Impacted	High Effect	High	Moderate
Low	High	Moderate Potential	Likely Impacted	High Effect	High	Low
High	Low	Moderate Potential	Likely Impacted	High Effect	High	High
High	Minimal	Moderate Potential	Likely Impacted	High Effect	High	High
Moderate	Low	Moderate Potential	Likely Impacted	High Effect	High	Moderate
Low	Moderate	Moderate Potential	Likely Impacted	High Effect	High	Low
High	Low	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	High
High	Minimal	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	High
Low	Moderate	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Low
Moderate	Low	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Moderate
Low	Moderate	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Low
High	Moderate	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	High
Moderate	Moderate	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Moderate
Nontoxic	High	Moderate Potential	Likely Impacted	Moderate Effect	High	Nontoxic
Low	High	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Low
Nontoxic	High	Moderate Potential	Likely Impacted	Moderate Effect	Moderate	Nontoxic
Moderate	Minimal	Low Potential	Possibly Impacted	High Effect	High	Moderate
Low	Low	Low Potential	Possibly Impacted	High Effect	High	Low
Nontoxic	Moderate	Low Potential	Possibly Impacted	Moderate Effect	High	Nontoxic
Nontoxic	Moderate	Low Potential	Possibly Impacted	Moderate Effect	Moderate	Nontoxic
Moderate	Minimal	Low Potential	Possibly Impacted	Moderate Effect	Moderate	Moderate
Low	Low	Low Potential	Possibly Impacted	Moderate Effect	Moderate	Low
Moderate	Low	Moderate Potential	Possibly Impacted	Low Effect	Low	Moderate
Moderate	Moderate	Moderate Potential	Possibly Impacted	Low Effect	Low	Moderate
Low	High	Moderate Potential	Possibly Impacted	Low Effect	Low	Low
High	Minimal	Moderate Potential	Possibly Impacted	Low Effect	Low	High
High	Low	Moderate Potential	Possibly Impacted	Low Effect	Low	High
High	Moderate	Moderate Potential	Possibly Impacted	Low Effect	Low	High
High	Low	Moderate Potential	Possibly Impacted	Low Effect	Reference	High
High	Moderate	Moderate Potential	Possibly Impacted	Low Effect	Reference	High
Nontoxic	Minimal	Minimal Potential	Likely Unimpacted	Moderate Effect	Moderate	Nontoxic
Nontoxic	Minimal	Minimal Potential	Likely Unimpacted	Moderate Effect	High	Nontoxic
Nontoxic	Low	Minimal Potential	Likely Unimpacted	Moderate Effect	Moderate	Nontoxic
Nontoxic	Low	Minimal Potential	Likely Unimpacted	Moderate Effect	High	Nontoxic
Low	Minimal	Minimal Potential	Likely Unimpacted	Moderate Effect	Moderate	Low
Low	Minimal	Minimal Potential	Likely Unimpacted	Low Effect	Low	Low
Moderate	Minimal	Low Potential	Likely Unimpacted	Low Effect	Low	Moderate
Low	Low	Low Potential	Likely Unimpacted	Low Effect	Low	Low
Nontoxic	High	Moderate Potential	Likely Unimpacted	Unaffected	Reference	Nontoxic
Nontoxic	High	Moderate Potential	Likely Unimpacted	Unaffected	Low	Nontoxic
Moderate	Low	Moderate Potential	Likely Unimpacted	Unaffected	Reference	Moderate
Moderate	Moderate	Moderate Potential	Likely Unimpacted	Unaffected	Reference	Moderate
Low	Moderate	Moderate Potential	Likely Unimpacted	Unaffected	Reference	Low
Low	High	Moderate Potential	Likely Unimpacted	Unaffected	Reference	Low
Nontoxic	Minimal	Minimal Potential	Unimpacted	Unaffected	Reference	Nontoxic
Nontoxic	Minimal	Minimal Potential	Unimpacted	Unaffected	Low	Nontoxic
Nontoxic	Low	Minimal Potential	Unimpacted	Unaffected	Reference	Nontoxic
Nontoxic	Low	Minimal Potential	Unimpacted	Unaffected	Low	Nontoxic
Low	Minimal	Minimal Potential	Unimpacted	Unaffected	Reference	Low
Moderate	Minimal	Low Potential	Unimpacted	Unaffected	Reference	Moderate
Nontoxic	Moderate	Low Potential	Unimpacted	Unaffected	Reference	Nontoxic
Nontoxic	Moderate	Low Potential	Unimpacted	Unaffected	Low	Nontoxic
Low	Low	Low Potential	Unimpacted	Unaffected	Reference	Low
Moderate	High	High Potential	Inconclusive	Unaffected	Reference	Moderate
Low	Minimal	Minimal Potential	Inconclusive	High Effect	High	Low
High	Minimal	Moderate Potential	Inconclusive	Low Effect	Reference	High

(defined as a mean ERM quotient <0.5). The data sets used for evaluating the framework were independent of the data sets used to identify either the hot spot or reference areas.

Results

The framework evaluation results for the 25 sites produced an answer that differed from that of the median expert for only five of the samples (Table 6). There was only one sample for which the framework and median expert assessments differed by more than a single category, resulting in a total of six category disagreements for all samples. This compares favorably with agreement among the experts (Table 7). Only one of the six experts had a lesser number of disagreements with the median than did the framework, although a second expert had a comparable disagreement rate. The remaining experts disagreed with the median at approximately twice the rate of the framework.

The framework also had little bias, with three of the samples rated as less impacted compared to the expert median and two as more impacted. The overall net bias, which incorporated both the number and sign of the category differences, was -2. Only two of the experts had a lesser bias, while three of the experts had five times greater bias with more than 80% of these errors in a single direction.

The framework also did a good job of differentiating the BPTCP hot spots from reference areas. Almost 90% of the samples from predicted reference areas were classified as Unimpacted or Likely Unimpacted and none of these samples were classified as Clearly Impacted (Figure 2). In contrast, more than 80% of the samples from predicted hot spot areas were classified into one of the impacted categories, with more than 50% of the samples classified as Likely or Clearly Impacted.

Discussion

There are potential shortcomings of a formulaic approach to data integration, as there will occasionally be additional information about a site that would be factored in by experts but which are not included in a more structured objective assessment. However, the formulaic approach also offers some advantages. It is transparent, so that all parties will reach the same conclusion using the same data. Moreover, it is not prone to the individual biases associated with use of BPJ. Such biases, or the need for employing a large team of experts to average out individual biases, would be problematic for large-scale assessments.

Table 6. Individual site results for expert and MLOE framework assessments. Shading indicates sites for which the assessments differ.

Site	Expert Median	MLOE Framework
1	Unimpacted	Unimpacted
2	Possibly Impacted	Possibly Impacted
3	Likely Unimpacted	Possibly Impacted
4	Likely Unimpacted	Unimpacted
5	Likely Impacted	Likely Impacted
6	Unimpacted	Unimpacted
7	Likely Unimpacted	Likely Unimpacted
8	Likely Impacted	Likely Impacted
9	Possibly Impacted	Possibly Impacted
10	Likely Impacted	Likely Impacted
11	Clearly Impacted	Clearly Impacted
12	Possibly Impacted	Likely Impacted
13	Possibly Impacted	Possibly Impacted
14	Likely Impacted	Clearly Impacted
15	Likely Impacted	Likely Impacted
16	Possibly Impacted	Unimpacted
17	Possibly Impacted	Possibly Impacted
18	Unimpacted	Unimpacted
19	Clearly Impacted	Clearly Impacted
20	Clearly Impacted	Clearly Impacted
21	Clearly Impacted	Clearly Impacted
22	Clearly Impacted	Clearly Impacted
23	Unimpacted	Unimpacted
24	Unimpacted	Unimpacted
25	Unimpacted	Unimpacted

The selected framework employs unequal weighting among LOEs, which differs from that of the earliest triad integration frameworks (Chapman 1990). The present study also attempted to develop a framework based on equal weighting and found it did not perform as well in reproducing results from the experts. Subsequent discussions with the experts revealed that few of them placed equal weighting in their assessments. Most of them placed greatest emphasis on the benthos because it is the ultimate endpoint of interest and weighted chemistry the least because of potential exposure from unmeasured chemicals. The two-phased assessment approach and its inherent weightings of the different LOEs effectively mimicked the expert's thought process.

Table 7. Summary of categorical assessments for experts and MLOE framework. Differences in the number of sites are due to the exclusion of sites classified as inconclusive. Disagreement values for experts represent the total number of category differences between the expert's assessment and the median of all other experts' assessments. Framework disagreement is the number of category differences between the framework and median of all experts (maximum of 78 for all sites). Bias values reflect the net of positive or negative assessment differences, with positive numbers indicating a bias toward rating the site as more impacted.

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

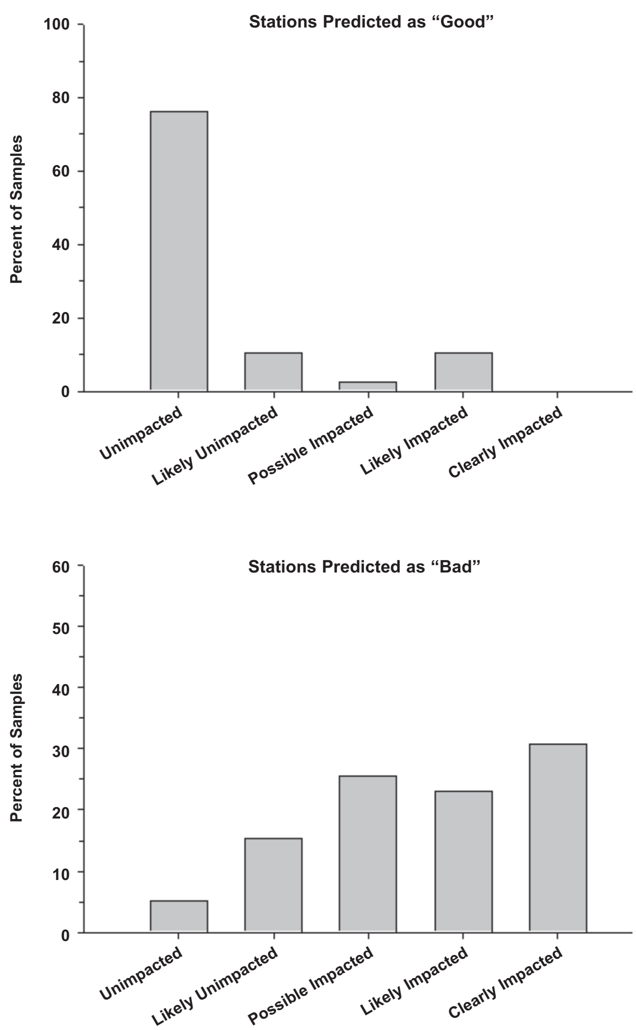

Figure 2. Distribution of MLOE assessment categories for California sites located in locations predicted from previous studies to be either unimpacted or impacted. n = 38 for the unimpacted samples; n = 39 for the impacted samples.

The framework ranks each LOE on a four-category scale, in contrast to the binary framework that was prevalent in the initial triad integration approaches (Long and Chapman 1985). A multi-category scale improves upon the binary approach because it lessens the all-or-none nature of thresholds that are established, and often measured, with great uncertainty (Batley *et al*. 2002). The use of five categories for such applications is prevalent in Europe, but ultimately the choice of number of categories becomes a tradeoff between placing great importance on a small number of thresholds and having more thresholds than there are philosophical bases on which to establish them. For the present study, we chose four categories because we could identify a unifying concept for threshold selection across LOEs. The first threshold, separating the reference and low effect categories, is one at which differences from background initially become apparent. The second threshold is where the differences become substantial enough that they can be detected with statistical certainty. The last threshold, separating the moderate and high effect categories, is one where the difference from background is severe. The last threshold is the most subjective because there is little precedent for distinguishing between moderate and high effects and establishing additional thresholds beyond that seemed increasingly artificial.

In the present study, application of the framework involved using multiple indices to summarize the complex benthic infauna and chemical data. The framework is not dependent on use of multiple indicators within an LOE, but multiple indicators proved helpful in reducing variability associated with individual indices and eliminated some inconsistencies with the experts that would have resulted from extreme values associated with a single index. We did not use a multiple indicator approach for the toxicity LOE because the data sets available for evaluation contained only a single toxicity test for which LOE thresholds were available. However, we believe that integrating multiple toxicity tests is also advisable in order to reduce uncertainty in the evaluation of this line of evidence (Burton *et al*. 1996).

The assessment framework yields six categories of interpretation. This differs from Chapman's original integration framework, which provided a separate interpretation for each combination of LOEs that described the extent to which the outcome was likely caused by chemical contamination. There is an advantage to having a large array of answers that incorporate a causality explanation, but many possible outcomes also complicates the linkage of the result to management outcomes. The six categories used here were selected in consultation with managers from the regulatory, regulated and public advocacy sectors. Their input was that information should be reduced to a linear scale that ranks sites, at least categorically, from best to worst. Linearization is scientifically challenging because it confounds several factors: confidence that there is an effect, magnitude of the effect, and likelihood that the effect is chemically-mediated. The two-phased assessment approach provides the management community with the linear response they need for large scale assessments while retaining a relationship to the individual lines of evidence needed to interpret data from an individual site.

The framework suggested here is not the only one possible. There have been numerous other suggested MLOE integration approaches, including those based on multivariate analysis, statistical summarization, logic models, and scoring systems (Burton *et al*. 2002, Chapman *et al*. 2002). It is also clear that when other data for a site are available, such as toxicity identification evaluations or bioaccumulation testing, they should be incorporated into the assessment process (Chapman and Hollert 2006). However, California's proposed framework was shown to reproduce the assessments of experts provided with the same data and provides a means for using a triad-based approach in large-scale assessments, such as Clean Water Act (CWA) 305b programs, or in a regulatory context where transparency in the decision process is critical.

Literature Cited

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.

Batley, G.E., G.A. Burton and P.M. Chapman. 2002. Uncertainties in sediment quality weight of evidence (WOE) assessments. *Human and Ecological Risk Assessment* 8:1517-1548.

Bay, S., W. Berry, P. Chapman, R. Fairey, T. Gries, E.R. Long, D. McDonald and S.B. Weisberg. 2007. Evaluating consistency of best professional judgment in the application of a multiple lines of evidence sediment quality triad. *Integrated Environmental Assessment and Management* 3:491-497.

Bay, S., D. Greenstein and D. Young. 2007a. 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., K.J. Ritter, D.E. Vidal-Dorsch and L.J. Field. 2008. Comparison of national and regional sediment quality guidelines for predicting sediment toxicity in California. pp. 79-90 *in*: S.B. Weisberg and K. Miller (eds.), Southern California Coastal Water Research Project Annual Report 2008. Costa Mesa, CA.

Burton, Jr., G.A., C.G. Ingersoll, L.C. Burnett, M. Henry, M.L. Hinman, S.J. Klaine, P.F. Landrum, P. Ross and M. Tuchman. 1996. A comparison of sediment toxicity test methods at three Great Lake areas of concern. *Journal of Great Lake Research* 22:495-511.

Burton, Jr., G.A., P.M. Chapman and E.P. Smith. 2002. Weight of evidence approaches for assessing ecosystem impairment. *Human and Ecological Risk Assessment* 8:1657-73.

Chapman, P.M. 1990. The sediment quality triad approach to determining pollution induced degradation. *Science of the Total Environment* 97:815-825.

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* 8:1489-515.

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

Chapman, P.M. and H. Hollert. 2006. Should the sediment quality triad become a tetrad, a pentad or possibly even a hexad? *Journal of Soil and Sediments* 8:4-8.

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

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.

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 sediments using Logistic Regression Models. *Environmental Toxicology and Chemistry* 9:1993-2005.

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.

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

Ingersoll, C.G., S.M. Bay, J.L. Crane, L.J. Field, T.H. Gries, J.L. Hyland, E.R. Long, D.D. MacDonald and T.P. O'Connor. 2005. Ability of SQGs to estimate effects of sediment-associated contaminants in laboratory toxicity tests or in benthic community assessments. pp. 497-556 *in*: R.J. Wenning, G.E. Batley, C.G. Ingersoll and D.W. Moore (eds.), Use of Sediment Quality Guidelines (SQGs) and Related Tools for the Assessment of Contaminated Sediments. Society of Environmental Toxicology and Chemistry. Pensacola, FL.

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

Phillips, B.M., B.S. Anderson, J.W. Hunt, J. Newman, R.S. Tjeerdema, C.J. Wilson, E.R. Long, M. Stephenson, H.M. Puckett, R. Fairey, J.M. Oakden, S. Dawson and H. Smythe. 1998. Sediment chemistry, toxicity and benthic community conditions in selected water bodies of the Santa Ana region. California State Water Resources Control Board. Sacramento, CA.

Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Cadien and R.G. Velarde. 2007. Evaluation of five indicators of benthic community condition in two California bay and estuary habitats. Technical Report 524. Southern California Coastal Water Research Project. Costa Mesa, CA.

Ritter, K.J., S.M. Bay, R.W. Smith, D.E. Vidal-Dorsch and L.J. Field. 2008. Development and evaluation of sediment quality guidelines based on benthic macrofauna responses. pp. 91-105 *in*: S.B. Weisberg and K. Miller (eds.), Southern California Coastal Water Research Project Annual Report 2008. Costa Mesa, CA.

Smith, R.W., M. Bergen, S.B. Weisberg, D.B. Cadien, A. Dalkey, D.E. Montagne, J.K. Stull and R.G. Velarde. 2001. Benthic response index for assessing infaunal communities on the Southern California Mainland Shelf. *Ecological Applications* 11:1073-1087.

Smith, R.W., J.A. Ranasinghe, S.B. Weisberg, D.E. Montagne, D.B. Cadien, T.K. Mikel, R.G. Velarde and A. Dalkey. 2003. Extending the Southern California Benthic Response Index to Assess Benthic Condition in Bays. Technical Report 410. Southern California Coastal Water Research Project. Westminster, CA.

Thompson, B. and S. Lowe. 2004. Assessment of macrobenthos response to sediment contamination in the San Francisco Estuary, California, USA.

Environmental Toxicology and Chemistry 23:2178-2187.

United States Environmental Protection Agency (USEPA). 1994. Methods for assessing the toxicity of sediment-associated contaminants with estuarine and marine amphipods. EPA 600-R94-025. USEPA, Office of Research and Development. Washington, DC.

USEPA. 2005. Predicting Toxicity to Amphipods from Sediment Chemistry (Final Report). EPA/600/R-04/030. USEPA, ORD National Center for Environmental Assessment. Washington, DC.

Van Sickle, J., D.D. Huff and C.P. Hawkins. 2006. Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. *Freshwater Biology* 51:359-372.

Weisberg, S.B., B.E. Thompson, J.A. Ranasinghe, D.E. Montagne and D.B. Cadien. 2008. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. *Ecological Indicators* 8:389-394.

Wright, J.F., M.T. Furse and P.D. Armitage. 1993. RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. *European Water Pollution Control* 3:15-25.

ACKNOWLEDGEMENTS

The authors thank Chris Beegan from the California Water Resources Control Board and Mike Connor, Bruce Thompson and Ben Greenfield of the San Francisco Estuary Institute for their thoughts during many discussions about possible frameworks. Jeff Brown assisted with data compilation and statistical analysis. The authors also thank Peter Landrum, Ed Long, Todd Bridges, Tom Gries, Rob Burgess and Bob Van Dolah for their thoughtful review of the ideas contained within the document. Work on this project was funded by the California State Water Resources Control Board under agreement 01-274-250-0.