APPENDIX G

Evaluation Of Five Indicators Of Benthic Community Condition In Two California Bay And Estuary Habitats

J. Ananda Ranasinghe[1], Stephen B. Weisberg[1], Robert W. Smith[2], David E. Montagne[3], Bruce Thompson[4], James M. Oakden[5], David D. Huff[6], Donald B. Cadien[7], and Ronald G. Velarde[8]

[1]Southern California Coastal Water Research Project, 3535 Harbor Blvd, Costa Mesa, CA 92626, USA
[2]Deceased
[3]PO Box 2004, Penn Valley, CA 95946, USA
[4]San Francisco Estuary Institute, 7770 Pardee Lane, Oakland, CA 94621, USA
[5]Moss Landing Marine Laboratory, Moss Landing, CA 95039, USA
[6]Dept. of Fisheries, Wildlife and Conservation Biology, University of Minnesota, St. Paul, MN, USA
[7]County Sanitation Districts of Los Angeles County, P.O. Box 4998, Whittier, CA, 90607, USA
[8]City of San Diego, Marine Biology Laboratory, 2392 Kincaid Rd., San Diego, CA 92101, USA

Abstract

Many types of indices have been developed to assess benthic invertebrate community condition, but there have been few studies evaluating the relative performance of different index approaches. Here we calibrate and compare the performance of five indices: the Benthic Response Index (BRI), Benthic Quality Index (BQI), Relative Benthic Index (RBI), River Invertebrate Prediction and Classification System (RIVPACS), and the Index of Biotic Integrity (IBI). We also examine whether index performance improves when the different indices, which rely on measurement of different properties, are used in combination. The five indices were calibrated for two geographies using 238 samples from Southern California marine bays and 125 samples from polyhaline central San Francisco Bay. Index performance was evaluated by comparing index assessments of 35 sites to the best professional judgment of nine benthic experts. None of the individual indices performed as well as the average expert in ranking sample condition or evaluating whether benthic assemblages exhibited evidence of disturbance. However, several index combinations outperformed the average expert.

Introduction

Index-based approaches to summarizing data have facilitated the use of benthic infauna as indicators of sediment condition in marine and estuarine environments (Hyland *et al.* 1999, Bergen *et al.* 2000, Dauer *et al.* 2000, Summers 2001, Hyland *et al.* 2003, Diaz *et al.* 2004). While reducing complex biological data to a single value has disadvantages, the resulting indices remove much of the subjectivity associated with interpreting data. The indices also provide a simple means for communicating complex information to managers and for correlating benthic responses with stressor data (Dauer *et al.* 2000, Hale *et al.* 2004, Bilkovic *et al.* 2006).

There have been a number of approaches to creating benthic indices (Diaz *et al.* 2004). Some integrate information at the community level and rely on parameters such as abundance, diversity, functional feeding groups, and depth beneath the sediment surface (Weisberg *et al.* 1997, Engle and Summers 1999, Van Dolah *et al.* 1999, Diaz *et al.* 2004). Other indices focus on species composition, comparing sample composition to an expected species mix or quantifying the average pollution tolerance of species found at the site (Borja *et al.* 2000, Hawkins *et al.* 2000, Smith *et al.* 2001, Smith *et al.* 2003, Leung *et al.* 2005, Van Sickle *et al.* 2006). Although community-level approaches often include measures of sensitive and tolerant biota, these measures are usually based on just a few indicator organisms, while species composition indices include many taxa.

Despite the broad range of benthic index approaches, there have been few comparisons of benthic index performance. When comparisons have been conducted, they have been limited to just a few indices and have not included comparison of community-level or species composition indices (Ranasinghe *et al.* 2002, Labrune *et al.* 2006, Quintino *et al.* 2006, Borja *et al.* 2007, Zettler *et al.* 2007). As a result, there are no widely accepted generalizations about the relative efficacy of indices at these two levels of organization.

In this study, we compare the performance of five benthic indices that rely on different sets of community or species composition measures. The five index approaches were (i) the Relative Benthic Index (RBI; Hunt *et al.* 2001), (ii) the Index of Biotic Integrity (IBI; Thompson and Lowe 2004), (iii) the Benthic Response Index (BRI; Smith *et al.* 2001, Smith *et al.* 2003; Ranasinghe *et al.* 2004), (iv) the River Invertebrate Prediction and Classification System (RIVPACS; Wright *et al.* 1993, Van Sickle *et al.* 2006), and (v) the Benthic Quality Index (BQI; Rosenberg *et al.* 2004). The RBI and IBI are based on community measures, the BRI and RIVPACS on species composition, and the BQI on both. The comparisons were conducted in two ecologically and geographically distinct habitats: (a) the marine bays of southern California and (b) polyhaline central San Francisco Bay. The objective was to evaluate the relative performance of these indices alone and in combination in each habitat.

Methods

The performance of the five benthic indices was evaluated in four steps:

- Data for sampling sites in each of the two habitats were identified, acquired, and adjusted to create consistency across sampling programs.
- The five benthic indices were calibrated using a common set of data for all indices.

- Threshold values were selected for each index to assess benthic condition on a four-category scale.
- Performance of the indices, and all possible index combinations, was evaluated by applying them to independent data and comparing the condition assessments to that of nine benthic experts.

Data

Data from projects that collected benthic species abundance and sediment chemistry data synoptically from marine bays in southern California and polyhaline central San Francisco Bay (Table 1) were identified, acquired, evaluated for methodological consistency, normalized for units of measure, and assembled into a database. Data about habitat conditions such as depth, bottom water salinity, sediment grain-size distributions, and acute toxicity to amphipods were included, if available.

Only benthic data from samples sieved through the most frequently used screen sizes were included: 1-mm sieve data for southern California marine bays and 0.5-mm sieve data for polyhaline central San Francisco Bay. Taxonomic inconsistencies among programs were eliminated by cross-correlating the species lists, identifying differences in nomenclature, and resolving discrepancies by consulting the taxonomists from each program. Species abundances were normalized to the most frequently occurring sample area by combining data from small samples or adjusting abundances to $0.1m^2$ in southern California marine bays and $0.05m^2$ in polyhaline central San Francisco Bay.

A portion of the available data was used to calibrate benthic indices while another was set aside to evaluate them. Approximately 90% of the data from each habitat (Table 1) were used for index calibration. Samples for evaluation were selected by ordering the data in each habitat by the mERMq (Long and MacDonald 1998) and systematically selecting sites from within quartile groups in each habitat. While it is generally accepted that current models of benthic response do not discriminate between chemical contamination and other sources of stress (Borja *et al.* 2003), this approach ensured that a range of benthic conditions were represented in the calibration and evaluation data.

An additional subset of the calibration data was set aside to select index threshold values. Similar to selecting evaluation samples, the subset of 35 samples from southern California and 33 samples from San Francisco Bay was selected by ordering the calibration data in each habitat by the mERMq (Long and MacDonald 1998) and systematically selecting sites within quartile groups in each habitat.

Benthic Index Calibration

All the indices, other than the BQI, have previously been calibrated, validated and used successfully in California, although RIVPACS was used only in freshwater streams. The BQI was previously calibrated and used in Europe. Our index calibration involved applying these previous calibration procedures to data from the southern California marine bays and polyhaline central San Francisco Bay. Each index was calibrated separately for each habitat.

(i) Benthic Response Index (BRI)

We calibrated the Benthic Response Index (BRI) using the methods of Smith *et al.* (2001, 2003) and Ranasinghe *et al.* (2004), with slight variations in the first and third of their four steps. The first step in BRI calibration is identifying a disturbance (or pollution) vector in an ordination space to facilitate calculation of species tolerance scores based on the distribution of species abundances along the vector. The BRI (Smith *et al.* 2001) was originally developed offshore, where a well-understood gradient of point-source disturbance allowed a disturbance vector to be identified from *a priori*-selected disturbed and undisturbed sites. Such simple disturbance gradients do not exist in bays and estuaries because there are many types of disturbance, a number of contaminant sources and circulation patterns that often redistribute contaminants throughout the system. Therefore, the BRI disturbance vector was selected using the vector with the maximum value for $T = R_{MSR} - R_{NSP}$ where R_{MSR} is the Spearman rank correlation between vector position and the observation mean species range (MSR) and R_{NSP} is the Spearman rank correlation between vector position and the observation number of species (Table 2). The MSR quantifies the average species range along the disturbance vector for the species occurring at a site. The range for each species was calculated as the difference between the last and first occurrence on the disturbance gradient; the MSR for a site is the average of the ranges for the species occurring at that site. We identified the disturbance vector by creating test vectors in the ordination space using an optimizing algorithm and selecting the vector with the highest value for T. The R_{NSP} computations excluded observations toward the undisturbed end of the vector to prevent the use of observations that might be to the left of a Pearson-Rosenberg species diversity peak. Species diversity would be negatively correlated with the disturbance gradient to the right of the diversity peak, leading to the negative sign for R_{NSP}.

The second BRI calibration step was application of an optimization procedure to determine data transformations to be used in subsequent computations (see Smith *et al.* 2001, 2003; Ranasinghe *et al.* 2004). Tolerance scores were calculated for abundance transformations with exponents (e in the tolerance score equation) of 0, 0.25, 0.33, 0.5, and 1.0 in combination with BRI calculations using transformations with exponents (f in the BRI equation) of 0.25, 0.33, 0.5, and 1.0. The combination with the highest Spearman correlation between optimized index values and the disturbance vector was used in each habitat (Table 3).

The third BRI calibration step selects the maximum number of occurrences used for species tolerance score calculations. In each habitat, the iteration with the highest Spearman correlation between optimized index values and the disturbance vector was selected using another iterative optimization procedure. Where previous versions of the BRI optimized the same maximum number of occurrences for all species in a habitat, we customized values for each species with the objective of including low abundances in tolerance score calculations only if they contribute signal, rather than noise. We used maximum occurrence values from iterations with Spearman correlations of 0.937 and 0.957 between the disturbance vector and the occurrence adjusted index values in southern California marine bays and polyhaline central San Francisco Bay, respectively.

In the final step, pollution tolerance scores were calculated for species occurring in two or more samples in each habitat as the position of the weighted-average of the abundance distribution on the disturbance vector. Tolerance values were calculated for 460 species in southern California marine bays and 154 species in polyhaline central San Francisco Bay. Higher BRI values are associated with higher pollution levels.

(ii) Benthic Quality Index (BQI)

We calibrated the Benthic Quality Index (BQI) for each habitat using the method of Rosenberg *et al.* (2004). First, for each sample in the calibration data, the expected number of species for a subset of 50 individuals was calculated as

$$ES50_k = \sum_{i=1}^{s}\left[1-\frac{(N_k - N_{ki})!(N_k - 50)!}{(N_k - N_{ki} - 50)!N_k!}\right],$$

where s is the number of species in sample k, N_k is the total abundance of all species in sample k, and N_{ki} is the abundance of species i in sample k. Next, species tolerance scores were computed for species that were found in at least three samples in each habitat as the 5th percentile of the distribution of expected numbers of species for the samples in which the species occurred. Tolerance scores were calculated for 346 species in southern California marine bays and 132 species in polyhaline central San Francisco Bay. Once species tolerance scores were calculated, the BQI value for each sample k was computed as

$$BQI_k = \left(\sum_{i}^{n}\left(\frac{A_i}{totA}ES50_{0.05i}\right)\right)\left(\log_{10}(S+1)\right),$$

where n is the number of species in the sample with tolerance scores, A_i is the abundance of species i, $totA$ is the total abundance in the sample, and S is the number of species in the sample. Higher BQI values are associated with lower pollution levels.

(iii) Relative Benthic Index (RBI)

We calculated Relative Benthic Index (RBI) values following the method of Hunt *et al.* (2001). The RBI was first calibrated to each habitat by selecting negative and positive indicator taxa. Then, RBI values were calculated as the weighted sum of (a) four community parameters (total number of species, number of crustacean species, number of crustacean individuals, and number of mollusc species), and abundances of (b) three positive and (c) two negative indicator organisms. The negative indicator taxa selected for both habitats were oligochaeta and *Capitella capitata* complex, which have been used for this purpose in previous versions of the RBI. For positive indicator taxa, we followed the practice of selecting an amphipod, a bivalve, and a polychaetes, which is typical of previous applications of the RBI. For southern California marine bays, we selected the amphipod *Monocorophium insidiosum*, the bivalve *Asthenothaerus diegensis*, and the polychaete *Goniada littorea*. For polyhaline central San Francisco Bay positive indicator taxa, we selected the amphipod *Sinocorophium heteroceratum*, the bivalve *Rochefortia* spp., and the polychaete *Prionospio lighti.* The RBI was scaled from 0 to 1.0, with 0 being the "worst" sample and 1 being the "best" sample in the calibration data.

(iv) River Invertebrate Prediction and Classification System (RIVPACS)

We used the methods of Wright *et al.* (1993) and Van Sickle *et al.* (2006) to calibrate the River Invertebrate Prediction and Classification System (RIVPACS) approach and calculate index values. We first used cluster analysis to define site-groups of reference samples in the calibration data, based on the presence or absence of species occurring there. Discriminant function analysis of habitat variables at the site-groups was then used to build discriminant functions that can be used to classify future sampling sites into site-groups based on habitat variable values. Minimally impacted reference sites for this calibration were selected by eliminating samples with high toxicity (control-adjusted survival < 50%) to amphipods, one or more chemicals exceeding ERM concentrations (Long *et al.* 1995), three or more chemicals exceeding their ERL concentrations (Long *et al.* 1995) or from sites influenced by point source discharges.

Several different habitat models explaining site groupings based on species abundances were explored in the southern California marine bays and polyhaline central San Francisco Bay by altering the numbers of site groupings and by varying the habitat variables used to explain the groupings. Based on the proportion of variance explained, 12 and 4 site group models based on latitude, longitude, and depth were selected for the southern California marine bays and San Francisco Bay, respectively. The probability of belonging to each of the site groups was calculated for each test site, based on the habitat variables. The site-group mean abundance for each taxon was then combined with the group probabilities to generate an expected taxon list specific to each test site. All permutations and combinations of numbers of groups and habitat variables were tested, and the combination with the greatest RIVPACS score improvement over an equivalent, non-predictive null model was selected (Van Sickle *et al.* 2005). Predictive improvement was quantified by calculating the reduction in root mean squared error (RMSE) of the predictive model (i.e., the model built using a discriminant function) from the null model. The chosen discriminant function model was then used to establish predictions for the species that would be expected to occur at reference sites in each group. The discriminant functions developed during calibration were used on the evaluation samples, first to identify the habitat site-group to which a sample belonged, and then to evaluate the observed species in relation to expectations for a minimally disturbed reference site. The difference between expected and observed assemblages measures the departure of the site from reference condition. For southern California marine bays, 619 species with > 50% probability of occurring in reference samples were included in the predictive model, while 365 species were included for polyhaline central San Francisco Bay. Summary statistics for the models are presented in Table 4. Based on a one to one ratio of modeled expected to observed (O/E) species present at validation sites they explained 89% and 96% of the variance, respectively.

(v) Index of Biotic Integrity (IBI)

The Index of Biotic Integrity (IBI) approach developed by Thompson and Lowe (2004) was applied in San Francisco Bay without modification. The same approach, which was also used by Ranasinghe *et al.* (2004), was applied to the calibration data for the southern California marine bays. First, twenty-two candidate metrics were evaluated for suitability as indicators, based on

criteria such as conforming to current conceptual models of benthic response to contamination and demonstrating measurable response to sediment contamination. Plots of candidate indicators vs. mERMq were examined, multiple regression analysis was conducted to evaluate the relationships between candidate IBI metrics and percent fines, TOC, and mERMq (independent variables), and four metrics were selected. Next, 59 reference samples were identified and reference ranges calculated for the four selected metrics as the maximum and minimum values for the reference samples. Reference sample selection was based on the same four criteria as Ranasinghe *et al.* (2004), including the absence of toxicity to amphipods. Table 5 presents the benthic assessment measures and reference ranges that were selected for each habitat. The assessment measures selected for southern California marine bays were based on the present study and reference ranges were established using the 59 designated reference samples. The measures and ranges for polyhaline central San Francisco Bay are those of Thompson and Lowe (2004).

Index Threshold Scaling

All five index approaches were calibrated to the same four-category scale of benthic condition: 1) Unaffected − a community that would occur at a reference site for that habitat; 2) Marginal deviation from reference − a community that exhibits some indication of stress, but might be within measurement variability of reference condition; 3) Affected − a community that exhibits clear evidence of physical, chemical, natural, or anthropogenic stress; 4) Severely Affected − a community exhibiting a high magnitude of stress. Affected and severely affected communities are those believed to be showing clear evidence of disturbance, while unaffected and marginal communities do not. Disturbed communities could be due to the effects of one or more types of anthropogenic or natural stress while undisturbed communities likely indicate minimal stress of all types.

Three approaches were used to establish threshold values for each index and the threshold set that performed best with the evaluation samples was selected. The first, or developer set of thresholds, was established by applying the principles used in the original index approach to the calibration data. Two other sets of thresholds were established by applying statistical optimization methods to compare index values and benthic condition categories.

For the BRI, the developer thresholds were based on reductions in the numbers of species along the disturbance gradient. Thresholds were established at index values along the disturbance gradient where the number of species declined to 95%, 75% and 25% of the reference species pool. These thresholds are equivalent to those established for the southern California mainland shelf by Smith *et al.* (2001) because similar reductions in numbers of species accompanied the changes in community structure and function on which those thresholds were based (see Smith *et al.* 2003).

The BQI developer thresholds were selected by dividing the index range into four with three equally spaced thresholds, following the approach used by Rosenberg *et al.* (2004). RBI developer thresholds were based on the distribution of index values, following Hunt *et al.* (2001). Reference thresholds were selected to segregate clusters of stations with high RBI values, high values for community parameters, and the presence of at least two of three positive

indicator taxa. The threshold differentiating between disturbed and undisturbed areas (i.e., between Marginal and Affected) was designated as the minimum RBI value where all three positive indicator taxa were found; 0.26 was selected in polyhaline central San Francisco Bay because *Prionospio lighti* first occurred at this RBI value. The Reference-Marginal threshold was selected at a mode of first occurrence for 18-20 species in the southern California marine bay calibration data; when a number of species have their first station of occurrence around a certain RBI value, that probably indicates a combination of factors that represent a significant change in habitat quality. Because there was no obvious mode in first stations of occurrence for San Francisco Bay, the threshold between Moderate and Severely Affected was chosen at 0.10, the RBI value of the first station of occurrence of the positive indicator species *Sinocorophium heteroceratum*.

For the RIVPACS approach, developer thresholds were set at 0.5, 1.0 and 2.0 standard deviations of the calibration score mean on either side of an observed to expected (O/E) ratio of 1.0. For the IBI, the same threshold evaluation process was used for both habitats, although the San Francisco Bay IBI (Thompson and Lowe 2004) was not recalibrated because it was based on the same data. Sample IBI values were evaluated graphically and statistical comparisons of IBI values and sediment contamination (mERMq) in disturbed and undisturbed samples were used to evaluate whether the assessment results reflected significant differences in sediment contamination. In southern California, sites with no IBI measures outside a reference range were considered Reference, sites with only one measure outside a reference range were considered Marginal, sites with two measures outside the ranges were considered Affected, and sites with three or four measures outside their ranges were considered Severely Affected. A slightly different scheme was used in San Francisco Bay. Sites with no measures or only one measure outside a reference range were considered Reference, sites with two measures outside their reference ranges were considered Marginal, sites with three measures outside their reference ranges were considered Affected, and sites with four measures outside their reference ranges were considered Severely Affected (Thompson and Lowe 2004).

Non-developer thresholds were selected by applying optimization techniques that maximized agreement between the indices and the consensus condition assigned to 68 sites by four benthic experts. One optimization technique was based on the Kappa statistic (Cohen 1960, Cohen 1968), which maximizes categorical agreement by identifying thresholds that minimize the number of sites with severe disagreement. The other optimization maximized classification accuracy with a lesser correction for severity of disagreement.

Evaluation of Index Performance

Index performance was assessed by comparing index results to the consensus assessment of nine benthic experts that were given species abundances, together with habitat, depth, salinity and sediment grain-size information for 35 sites that were not used in index development or calibration (Weisberg *et al.* In press). The experts were asked to (1) rank the sites in each habitat from best to worst condition and (2) classify each site on the four-category scale of benthic condition to which the benthic indices were calibrated. Index condition rank order was evaluated against the average expert rank order using Spearman rank correlation coefficients. Condition

category assessments by the benthic indices, and by all possible index combinations, were compared to the consensus expert condition assessment in three ways:

1. Status classification accuracy, the accuracy with which an index differentiated benthos identified by the nine experts as disturbed (affected or severely affected categories) from benthos identified as undisturbed (reference or marginal categories). This mimics the evaluation approach used in most previously published benthic indicator development efforts.

2. Categorical classification accuracy with respect to the four categories established for index calibration (Reference, Marginal, Affected or Severely Affected).

3. Bias in category designation; the sum of differences between index (or index combination) category and the consensus categorical classification of the experts when categories are expressed numerically (Reference=1, Severely Affected=4). Positive bias indicates a tendency to score samples as more disturbed than the expert consensus, while negative bias indicates a tendency to score samples as less disturbed. Larger absolute values indicate stronger bias.

Index combinations were evaluated as the median of the numeric categories (Reference=1, Severely Affected=4). If the median for the indices in a combination fell between categories, it was rounded to the higher effect category. Comparisons to the experts were performed for each of the three threshold approaches associated with each index, with the best performing thresholds used when combining indices.

Results

Spearman correlation coefficients between index condition ranks and the average expert ranks for the 35 evaluation samples ranged from 0.70 to 0.89 (Table 6). The strongest correlation coefficient for an index (0.89) was slightly stronger than the weakest correlation coefficient for an expert in polyhaline central San Francisco Bay (0.88) and slightly weaker than the weakest expert (0.90) in the southern California marine bays. All the Spearman correlations were highly significant ($p < 0.01$), except for the IBI, which was only applied to five of the San Francisco Bay evaluation samples.

Index condition categories were evaluated for 34 of the 35 samples, as the experts were evenly split as to the condition of one site. In the southern California marine bays, the RIVPACS index performed best, with 87.5% correct status classification, 66.7% correct category classification and low bias (Table 7). The status classification accuracy was higher than one of the nine experts and tied with two others, but was not as high as the average expert (91.2%). The RIVPACS category classification accuracy was higher than the lowest expert. The BRI also had 87.5% correct status classification, but category classification accuracy was not as high as the lowest expert. None of the other indices had a status classification accuracy higher than the lowest expert but, except for the IBI and RBI, all were higher than 75%, which has frequently been used as a standard for indices developed in other estuarine systems (e.g., Engle and

Summers 1999, Van Dolah *et al.* 1999). In polyhaline central San Francisco Bay, at 100%, status classification accuracy for all five indices was the same as the three highest experts. All five indices had higher category classification accuracy than the weakest expert, but only the BQI was higher than the average expert.

When there were differences, indices based on species composition almost always had higher classification accuracy both for status and for four-category assessments than indices based only on community measures. In southern California marine bays, the RIVPACS, BRI and BQI, which are based on species composition, had status classification accuracy of 87.5%, 87.5% and 79.2%, which is lower than the 91.2% classification accuracy for the average expert. The RBI and IBI, which are based on community measures, both had status classification accuracy of 70.8% (Table 7). Four-category classification accuracy was 66.7%, 58.3% and 62.5% for the species composition based RIVPACS, BRI, and BQI, and 50.0% for the community measure-based RBI and IBI. Category bias was also lower for RIVPACS and the BRI than for either of the community measure based indices. In polyhaline central San Francisco Bay, category classification accuracy for the species composition based RIVPACS and BQI was 80.0% and 90.0%, and 70.0% and 75.0% for the community measure based RBI and IBI, respectively. The category classification accuracy for the BQI here was 70.0%, which was the only instance where accuracy for a species composition based index was lower than any community measure based index.

Index combinations generally performed better than individual indices, and combinations of three or more indices generally performed better than combinations of two. In southern California marine bays, seven combinations of three or more indices achieved the highest status classification accuracy of 91.7% (Table 7). One of these combinations, #29, had the highest four-category classification accuracy of 79.2%. The accuracy for this four-index combination of the BRI, BQI, IBI and RIVPACS was only slightly less than the accuracy of 80.1% for the average expert. Another five of these combinations were in second place for category classification accuracy at 75%. In polyhaline central San Francisco Bay, the percentage of index combinations with category classification accuracy of 80% or higher increased from 40% for single indices to 50%, 80%, 100% and 100% for combinations of two, three, four and five indices.

When results for both habitats were combined, the three index combinations that performed best were #24, a three-index combination of the BRI, RBI, and RIVPACS, #26, a four-index combination of the BRI, the RBI, the IBI and RIVPACS, and #29, a four-index combination of the BRI, the BQI, the IBI and RIVPACS. These combinations had the highest status classification accuracy (94.1%), the highest category classification accuracy (79.4%) and low bias (3, 5, and 5, respectively). These combinations outperformed the average expert for status classification, but were outperformed by five of the nine experts for categorical classification. All three of the best-performing combinations include a mixture of community measures and species composition indices.

Discussion

Indices that include measures of species composition generally outperformed indices that include only community measures. This is consistent with Weisberg *et al.* (1997), who found that relative dominance of pollution-tolerant and pollution-sensitive species were the metrics in their index that had the best relationship to pollution gradients. Pearson and Rosenberg (1978) suggest that the initial benthic response to low levels of stress is a shift in species composition, with shifts in community metrics, such as loss of species richness and biomass, manifesting at later stages of stress. Thus, indices based on community metrics should be more effective at differentiating sites subject to high levels of stress, but less effective at differentiating sites with low to intermediate levels of stress that are more typical of the estuarine sites encountered in California.

Combinations of indices consistently outperformed individual indices. Each of the indices relies on a subset of metrics used by experts. Generally, these metrics correlate among themselves and produce the same answer as the experts. However, there are circumstances when these metrics can differ considerably, such as when the presence of a large filter feeder reduces species richness and abundance, or when only a few individuals of a few sensitive species occur. Use of multiple indices incorporates a larger number of metrics and presumably balances the occasional erratic behavior of some metrics. In addition, some of the indices showed biases, with the RBI assessing samples as more disturbed than the experts and the IBI behaving the opposite. Use of multiple indices apparently balances out those biases.

Conclusions about relative performance of indices are reliant upon proper implementation of the index approaches. Our study team included the original developers of the index approach, or investigators who had previously published applications of these indices in other habitats, for four of the five indices evaluated. The team had less experience with the BQI, but this method involves the least amount of developer judgment in its calibration. One indication that our results reflect successful implementation was the high classification accuracy for discriminating among undisturbed and disturbed benthic community status for all of our indices. Our range of 70-100% classification accuracy achieved for the individual indices compares favorably with the average status classification accuracy of 85% that Weisberg *et al.* (1997) achieved for seven Chesapeake Bay habitats, the 85% that Van Dolah *et al.* (1999) achieved in the best of his four southeastern USA estuaries, and the 76% that Engle and Summers (1999) achieved for Gulf of Mexico estuaries.

One factor that may have led to our slightly higher validation success was our approach to selecting validation sites. Validation site selection has historically been conducted by using chemical and toxicological exposure measures to identify sites of supposedly extreme condition. Here, we used expert professional judgment (Weisberg *et al.* In press) to establish a site's condition. Use of expert judgment to classify sites for validation avoids the concern that unmeasured chemicals or physical disturbance may cause a perturbed site to be incorrectly classified as an undisturbed site. It also avoids the concern of false disturbed site designations due to contaminants that are measured in chemical analysis but are tightly bound to sediments and unavailable *in situ* to benthic organisms.

Using expert judgment to classify sites for index validation has the additional advantage of allowing evaluation of index performance at sites experiencing intermediate levels of disturbance. This cannot be conducted using exposure measures to classify validation sites, as there is no expectation of a linear relationship between biological responses and chemical exposure. Assessment of intermediate conditions is a more difficult, but more relevant, assessment challenge for benthic indices. Interestingly, the indices matched expert opinion for the intermediate sites as well as they did for sites of more extreme condition.

Expert opinion also provides a benchmark for evaluating index performance by comparing to levels of agreement among experts. Historically, index developers have deemed an index successful if it correctly identifies 75-80% of sites with extreme exposure conditions (Van Dolah *et al.* 1999). However, since indices are intended to reproduce the experience of experts in interpreting benthic data using an objective, repeatable, transparent tool, a better evaluation benchmark is whether an index ranks and classifies sites with levels of correlation and accuracy comparable to that among experts. In this study, none of the individual indices achieved this mark, but several index combinations did.

Tables

Table 1. Data sources for calibration and validation samples.

Habitat (Sampling Methods)	Project	Period	Reference	No. of samples	
				Calibration	Validation
Southern California Marine Bays. (0.1-mm sieve; 0.1m^2 sample area)	Bight'98	1998	Ranasinghe *et al.* (2003)	107	5
	Bight'03	2003	Ranasinghe *et al.* (2007)	110	10
	San Diego TMDL	2001-2002	SCCWRP and SPAWAR (2004); Brown and Bay (2005)		4
	EMAP	1999	U.S. EPA (2004)	21	5
	Total			**238**	**24**
Polyhaline central San Francisco Bay. (0.5-mm sieve; 0.05m^2 sample area)	EMAP	2000	U.S. EPA (2004)	22	1
	BADA	1994-1997	Bay Area Dischargers Association (1994)	42	2
	BPTCP	1994, 1997	Hunt *et al.* (2001)	16	4
	RMP	1994-2000	Thompson *et al.* (1999)	45	4
	Total			**125**	**11**

Table 2. Spearman correlation coefficients between the vector in the ordination space selected to represent the disturbance gradient and (a) the mean species range and (b) the number of species in each habitat. The disturbance vector was selected by generating test vectors using an optimization procedure and selecting the vector that maximized the value of T. The mean species range is the average of the species ranges along the disturbance vector for the species occurring at each sampling site (see text).

	Southern California Marine Bays	Polyhaline Central San Francisco Bay
Spearman correlation with mean species range (R_{MSR})	0.9182	0.9007
Spearman correlation coefficient with number of species (R_{NSP})	-0.8457	-0.8632
$T = R_{MSR} - R_{NSP}$	1.7639	1.7639

Table 3. Optimum parameter values for exponents in the Benthic Response Index (BRI) equation for each habitat. The exponent f is used for index calculations, while e is used to develop species tolerance (p_i) values.

	Southern California Marine Bays	Polyhaline Central San Francisco Bay
E	0.25	0.33
F	0	0
Spearman correlation coefficient between the optimized index and the disturbance vector	0.903	0.944

Table 4. Summary statistics for RIVPACS predictive models (See Van Sickle *et al.* 2005). O/E: Observed to expected species ratio. *: Calibration data used for model development validation.

Statistic	Southern California Marine Bays	Polyhaline Central San Francisco Bay
O/E root mean squared error for predictive model based on validation sites	0.270	*
O/E standard deviation for null model (highest variability model)	0.434	0.451
O/E standard deviation for predictive model based on calibration sites	0.301	0.261
Predictive improvement over the null model	0.133	0.190
Standard deviation for calibration pseudoreplicate samples (least variability possible)	0.173	0.259

Table 5. IBI assessment measures and reference ranges for each habitat.

Southern California Marine Bays				Polyhaline Central San Francisco Bay			
Assessment measure	**Reference Range**			**Assessment measure**	**Reference Range**		
	Min.	**Max.**	**Mean**		**Min.**	**Max.**	**Mean**
Number of taxa (per 0.1m^2 sample)	13	99	48.5	Number of taxa (per 0.05m^2 sample)	21	66	40.4
Molluscan taxa (per 0.1m^2 sample)	2	25	10.6	Amphipod taxa (per 0.05m^2 sample)	2	11	5.3
Notomastus sp. abundance (per 0.1m^2)	0	59	2.7	Total abundance (per 0.05m^2 sample)	97	2,931	905.7
Sensitive taxa (%)	9.0	47.1	26.9	*Capitella capitata* abundance (per 0.05m^2)	0	13	2.0

Table 6. Spearman rank correlation coefficients between index condition ranks and average expert condition rankings for evaluation samples. The average, maximum and minimum correlations for the benthic experts are presented to provide context for the index correlations.

Index	**Spearman Rank Correlation Coefficient**	
	Southern California Marine Bays (n=24; $p < 0.0001$)	**Polyhaline Central San Francisco Bay** (n=11; $p < 0.01$ except ‡: n=5; NS)
BQI	0.89	0.89
BRI	0.88	0.77
IBI	0.70	0.71‡
RBI	0.82	0.87
RIVPACs	0.84	0.82
Expert minimum	0.90	0.88
Expert mean	0.95	0.95
Expert maximum	0.98	0.99

Table 7. Classification Accuracy and Bias for Indices and Index Combinations. Classification accuracy is presented for "undisturbed" vs. "disturbed" status and four condition categories. Each of 35 evaluation samples was assessed into one of four numeric categories by the index or index combination and compared with consensus categories from an independent assessment by nine benthic experts. Bias is the sum of differences between index combination and consensus categories; positive values indicate a tendency to score samples as more disturbed than the expert consensus, while negative values indicate a tendency to score samples as less disturbed. The categories were 1: Reference; 2: Marginal; 3: Affected; 4: Severely Affected. Categories 1 and 2 were considered "undisturbed" and 3 and 4 as "disturbed." Index results were combined as the median of the numeric categories; if the median fell between categories, it was rounded to the higher effect category. Results for the benthic experts are presented to provide context.

			Southern California Marine Bays (n=24)			Polyhaline Central San Francisco Bay (n=10)		
No. of indices	**#**	**Measure**	**Category Accuracy (%)**	**Category Bias**	**Status Accuracy (%)**	**Category Accuracy (%)**	**Category Bias**	**Status Accuracy (%)**
One	1	BQI	62.5	8	79.2	90.0	-1	100.0
	2	BRI	58.3	-3	87.5	70.0	-1	100.0
	3	IBI	50.0	-8	70.8	75.0	-1	100.0
	4	RBI	50.0	10	70.8	70.0	3	100.0
	5	RIV	66.7	3	87.5	80.0	0	100.0
Two	6	BQI, BRI	54.2	7	79.2	90.0	1	100.0
	7	BQI, IBI	58.3	6	79.2	90.0	-1	100.0
	8	BQI, RBI	45.8	13	75.0	70.0	3	100.0
	9	BQI, RIV	62.5	11	75.0	80.0	0	100.0
	10	BRI, IBI	66.7	0	83.3	70.0	-1	100.0
	11	BRI, RBI	58.3	9	83.3	70.0	3	100.0
	12	BRI, RIV	62.5	6	83.3	90.0	1	100.0
	13	IBI, RBI	45.8	8	70.8	70.0	3	100.0
	14	IBI, RIV	66.7	3	87.5	80.0	0	100.0
	15	RBI, RIV	45.8	13	75.0	70.0	3	100.0
Three	16	BRI IBI RBI	70.8	-1	87.5	80.0	2	100.0
	17	BQI BRI IBI	66.7	0	87.5	80.0	0	100.0
	18	BQI BRI RBI	70.8	5	83.3	90.0	1	100.0
	19	BQI BRI RIV	70.8	3	91.7	80.0	0	100.0
	20	BQI IBI RBI	66.7	6	83.3	70.0	1	100.0
	21	BQI IBI RIV	75.0	2	91.7	80.0	0	100.0
	22	BQI RBI RIV	66.7	6	83.3	80.0	0	100.0
	23	BRI IBI RIV	62.5	-3	87.5	80.0	0	100.0
	24	BRI RBI RIV	75.0	2	91.7	90.0	1	100.0
	25	IBI RBI RIV	75.0	2	91.7	70.0	1	100.0
Four	26	BRI IBI RBI RIV	75.0	4	91.7	90.0	1	100.0
	27	BQI IBI RBI RIV	66.7	6	83.3	80.0	0	100.0
	28	BQI BRI RBI RIV	70.8	7	83.3	90.0	1	100.0
	29	BQI BRI IBI RIV	79.2	5	91.7	80.0	0	100.0
	30	BQI BRI IBI RBI	70.8	7	83.3	90.0	1	100.0
Five	31	All	75.0	4	91.7	80.0	0	100.0
Expert Consensus		Minimum	62.5	+1, -1	83.3	60.0	0	90.0
		Average	80.1	-0.2	91.2	84.4	0.56	94.4
		Maximum	87.5	+4, -3	100.0	100.0	+4, -2	100.0

Literature Cited

Bay Area Dischargers Association. 1994. Local effects monitoring program, quality assurance project plan. Bay Area Clean Water Association. Oakland, CA.

Bergen, M., D.B. Cadien, A. Dalkey, D.E. Montagne, R.W. Smith, J.K. Stull, R.G. Velarde and S.B. Weisberg. 2000. Assessment of benthic infaunal condition on the mainland shelf of Southern California. *Environmental Monitoring and Assessment* **64**:421-434.

Bilkovic, D.M., M. Roggero, C.H. Hershner and K.H. Havens. 2006. Influence of land use on macrobenthic communities in nearshore estuarine habitats. *Estuaries and Coasts* **29**:1185-1195.

Borja, A., J. Franco and V. Perez. 2000. A marine Biotic Index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. *Marine Pollution Bulletin* **40**:1100-1114.

Borja, A., A.B. Josefson, A. Miles, I. Muxika, F. Olsgard, G. Phillips and J.G. Rodriguez. 2007. An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. *Marine Pollution Bulletin* **55**:42-52.

Borja, A., I. Muxika and J. Franco. 2003. The application of a Marine Biotic Index to different impact sources affecting soft-bottom benthic communities along European coasts. *Marine Pollution Bulletin* **46**:835-845.

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

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**:37-46.

Cohen, J. 1968. Weighted Kappa nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin* **70**:213-220.

Dauer, D.M., J.A. Ranasinghe and S.B. Weisberg. 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in Chesapeake Bay. *Estuaries* **23**:80-96.

Diaz, R.J., M. Solan and R.M. Valente. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. *Journal of Environmental Management* **73**:165-181.

Engle, V.D. and J.K. Summers. 1999. Refinement, validation, and application of a benthic condition index for Northern Gulf of Mexico estuaries. *Estuaries* **22**:624-635.

Hale, S.S., J.F. Paul and J.F. Heltshe. 2004. Watershed landscape indicators of estuarine benthic condition. *Estuaries* **27**:283-295.

Hawkins, C.P., R.H. Norris, J.N. Hogue and J.W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications* **10**:1456-1477.

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.

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

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

Labrune, C., J.M. Amouroux, R. Sarda, E. Dutrieux, S. Thorin, R. Rosenberg and A. Gremare. 2006. Characterization of the ecological quality of the coastal Gulf of Lions (NW Mediterranean). A comparative approach based on three biotic indices. *Marine Pollution Bulletin* **52**:34-47.

Leung, K.M.Y., A. Bjorgesaeter, J.S. Gray, W.K. Li, G.C.S. Lui, Y. Wang and P.K.S. Lam. 2005. Deriving sediment quality guidelines from field-based species sensitivity distributions. *Environmental Science & Technology* **39**:5148-5156.

Long, E.R. and D.D. MacDonald. 1998. Recommended uses of empirically derived, sediment quality guidelines for marine and estuarine ecosystems. *Human and Ecological Risk Assessment* **4**:1019-1039.

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.

Pearson, T.H. and R. Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanogr. Mar. Biol. Ann. Rev.* **16**:229-311.

Quintino, V., M. Elliott and A.M. Rodrigues. 2006. The derivation, performance and role of univariate and multivariate indicators of benthic change: Case studies at differing spatial scales. *Journal of Experimental Marine Biology and Ecology* **330**:368-382.

Ranasinghe, J.A., A.M. Barnett, K.C. Schiff, D.E. Montagne, C. Brantley, C. Beegan, D.B. Cadien, C. Cash, G.B. Deets, D.R. Diener, T.K. Mikel, R.W. Smith, R.G. Velarde, S.D. Watts and S.B. Weisberg. 2007. Southern California Bight 2003 Regional Monitoring Program: III Benthic Macrofauna. Southern California Coastal Water Research Project Authority. Costa Mesa, CA.

Ranasinghe, J.A., J.B. Frithsen, F.W. Kutz, J.F. Paul, D.E. Russell, R.A. Batiuk, J.L. Hyland, K.J. Scott and D.M. Dauer. 2002. Application of two indices of benthic community condition in Chesapeake Bay. *Environmetrics* **13**:499-511.

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

Ranasinghe, J.A., B. Thompson, R.W. Smith, S. Lowe and K.C. Schiff. 2004. Evaluation of benthic assessment methodology in southern California bays and San Francisco Bay. Southern California Coastal Water Research Project. Westminster, CA. Technical Report 432.

Rosenberg, R., M. Blomqvist, H.C. Nilsson, H. Cederwall and A. Dimming. 2004. Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. *Marine Pollution Bulletin* **49**:728-739.

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 Southern California Coastal Water Research Project. Westminster, CA. Technical Report 410.

Southern California Coastal Water Research Project and Space and Naval Warfare Systems Center San Diego. 2004. Sediment assessment study for the mouths of Chollas and Paleta Creek, San Diego. Phase I Draft Report. San Diego Regional Water Quality Control Board, Commander Navy Region Southwest, City of San Diego. San Diego, CA.

Summers, J.K. 2001. Ecological condition of the estuaries of the atlantic and gulf coasts of the United States. *Environmental Toxicology and Chemistry* **20**:99-106.

Thompson, B., B.S. Anderson, J.W. Hunt, K.M. Taberski and B.M. Phillips. 1999. Relationships between sediment contamination and toxicity in San Francisco Bay. *Mar. Environ. Res.* **48**:285-395.

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.

U.S. Environmental Protection Agency. 2004. National Coastal Condition Report II. U.S. Environmental Protection Agency, Office of Research and Development. Washington, DC. EPA-620/R-03/002.

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

Van Sickle, J., C.P. Hawkins, D.P. Larsen and A.T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. *J. N. Am. Benthol. Soc.* **24**:178-191.

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

Weisberg, S.B., J.A. Ranasinghe, L.C. Schaffner, R.J. Diaz, D.M. Dauer and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. *Estuaries* **20**:149-158.

Weisberg, S.B., B. Thompson, J.A. Ranasinghe, D.E. Montagne, D.B. Cadien, D.M. Dauer, D.R. Diener, J.S. Oliver, D.J. Reish, R.G. Velarde and J.Q. Word. In press. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. *Ecological Indicators*

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.

Zettler, M.L., D. Schiedek and B. Bobertz. 2007. Benthic biodiversity indices versus salinity gradient in the southern Baltic Sea. *Marine Pollution Bulletin* **55**:258-270.