The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities

Stephen B. Weisberg, Bruce Thompson[1], J. Ananda Ranasinghe, David E. Montagne[2], Donald B. Cadien[2], Daniel M. Dauer[3], Douglas Diener[4], John Oliver[5], Donald J. Reish[6], Ronald G. Velarde[7] and Jack Q. Word[8]

Abstract

Benthic infaunal communities are frequently used to assess aquatic environmental condition, but interpretation of benthic data is often subjective and based on best professional judgment. Here, we examine the repeatability of such assessments by providing species-abundance data from 36 sites to nine independent benthic experts who ranked the sites from best to worst condition. Their site rankings were highly correlated, with an average correlation coefficient of 0.92. The experts also evaluated the sites in terms of four condition categories: 1) Unaffected, 2) Marginal deviation from reference, 3) Affected, or 4) Severely affected. At least two-thirds of the experts agreed on site categorization for 91% of the samples and they disagreed by more than one category for less than 1% of the assessment pairs. The experts identified seven parameters they used in making their assessments, with four of those parameters (dominance by tolerant taxa, presence of sensitive taxa, species richness and total abundance) used by all of the experts. Most of the disagreements in site categorization were due to philosophical rather than technical differences, such as whether the presence of invasive species indicates a degraded community. Indices are increasingly being used as an alternative to best professional judgment for assessing benthic condition, but there have been inconsistencies in how sites are selected for validating such indices; the level of agreement found among experts in this study suggests that consensus expert opinion can be a viable benchmark for such evaluations.

Introduction

Biocriteria are increasingly being used to assess ecological integrity, with both the US Environmental Protection Agency (Gibson *et al.* 2000) and the European Water Framework Directive (Borja 2005, 2006, Jonge *et al.* 2006) providing guidance that promotes the use of biocriteria for coastal and estuarine assessments. Benthic infauna are prominent indicators in this guidance because their habitat exposes them to many anthropogenic influences: contaminants accumulate in the sediment, eutrophication leads to excess organic matter on the bottom and water column stratification facilitates hypoxia below the pycnocline. Additionally, the wide range of physiological tolerances, feeding modes, trophic interactions, and limited mobility among the diverse benthic taxa makes them responsive as a group to this array of environmental stressors (Bilyard 1987, Diaz *et al.* 2004).

The European and US directives recognize four approaches to developing biocriteria: comparison to historical conditions, comparison to present reference conditions, models and consensus professional judgment. Many numerical indices have been developed that minimize the need for subjective judgment to assess attainment of biocriteria (Weisberg *et al.* 1997, Engle and Summers 1999, Van Dolah *et al.* 1999, Borja *et al.* 2000, Paul *et al.* 2001, Smith *et al.*

[1]*San Francisco Estuary Institute, Oakland, CA*
[2]*County Sanitation Districts of Los Angeles County, Whittier, CA*
[3]*Old Dominion University, Department of Biological Sciences, Norfolk, VA*
[4]*Private Consultant, Oceanside, CA*
[5]*Moss Landing Marine Laboratories, Moss Landing, CA*
[6]*California State University, Long Beach, Department of Biological Sciences, Long Beach, CA*
[7]*City of San Diego Marine Biology Laboratory, San Diego, CA*
[8]*Weston Solutions, Port Gamble Environmental Laboratories, Port Gamble, WA*

2001, Thompson and Lowe 2004). However, even these objective indices involve subjectivity in several steps of their development and application, such as metric selection, site selection for index calibration, and index approach selection (Boyle *et al.* 1990).

Application of best professional judgment (BPJ) often follows from general models of benthic community response to stress (Pearson and Rosenberg 1978, Dauer 1993). However, experts with different backgrounds may emphasize different aspects or elements of these models, leading to uncertainty regarding the extent to which experts agree in their application of BPJ. The objective of this paper is to quantify concordance among experts in their application of BPJ to assess benthic impairment when each is provided with the same data.

Methods

Nine benthic ecologists were provided species composition and abundance data, as well as depth, salinity and sediment texture information, from 36 sites and asked to determine condition of the benthos at each site. These experts were selected to represent a range of affiliations and experience. Three were from academic institutions, three from municipalities that implement benthic monitoring programs to assess the effect of discharge outfalls, two from private consulting firms, and one from a nonprofit research organization. Their experience in benthic monitoring ranged from 20-50 years, with an average of 32 years. All had experience with benthos from the west coast of the United States, though one is presently working on the east coast.

Twenty-four of the sites were from southern California coastal bays and 12 from San Francisco Bay. The sites were selected from a large California sediment database by ordering the data base using Long and MacDonald's (1998) mean Effects Range-Median quotient (mERMq) and systematically selecting samples within each geography so that a range of conditions were represented.

The experts were asked to rank the relative benthic condition of each site from best to worst within each region. They were also asked to rate each site into one of four categories of absolute 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. While a measure of chemical contamination (mERMq) was used to select the range of sites to be examined, the experts were not provided the chemical data. They were also not asked to differentiate among potential causes for affected condition as it is generally recognized that current models of benthic response to stress do not discriminate between chemical contamination and other sources of disturbance (Borja *et al.* 2003).

The benthic experts were asked to list the attributes of the benthos that they used to determine site rankings and condition categories, and to rate the importance of the attributes as follows: 1) Very important; 2) Important, but secondary; 3) Marginally important; 4) Useful, but only to interpret the other factors. Attributes that were not used by an expert for site classification were assigned a rank of 5 for the purpose of calculating an average importance of that attribute among experts. As all the experts identified indicator species as one of the attributes used in their assessment, they were also asked to list the organisms they used as indicator species and rank the species importance using the same scale.

Results

The relative ranking of sites was highly correlated among all the experts, with an average correlation coefficient of 0.92 (Table 1). There was no difference in the average correlation among experts between sites in San Francisco Bay and southern California. None of the experts deviated notably from their peers, with the correlation coefficient for each reviewer in relation to the average of the other reviewers ranging between 0.90 - 0.94.

All the experts agreed completely on condition category for only four sites. However, eight of the nine experts agreed on condition category for more than 50% of the sites, and seven of the nine experts agreed on the condition category for 75% of the sites (Table 2). Only three sites elicited less than 66% agreement among the experts. Moreover, when disagreement occurred, the difference was almost always limited to a single category; the experts disagreed by more than one category for less than 1% of the assessment pairs.

The experts used seven criteria for assessing benthic assemblage condition and all but one were used as assessment parameters by at least half of the experts (Table 3). The three most important criteria were the presence of sensitive indicator species, species richness, and the proportion of tolerant taxa. Total abundance was also used by all of the experts, but many of them ranked this criterion as of lesser importance because they only used it as an indicator when abundance was low. Other criteria that were used included the abundance of selected higher taxa, presence of nonindigenous species and the diversity of functional/feeding groups.

There was considerable consistency in the indicator taxa identified by the experts (Table 4). The taxa most frequently recognized as tolerant were the polychaetes *Capitella capitata* complex and *Streblospio benedicti*, and oligochaetes. The most frequently recognized sensitive taxa were ophiuroids and amphipods. Although the tolerant taxa were generally identified at the species or genus level, most of the sensitive taxa were higher-level taxonomic groups. Some of the experts also indicated that they placed different emphasis or used different indicator taxa for southern California and San Francisco Bay.

Discussion

The experts generally agreed on the criteria used for assessment, but often disagreed on their relative importance. Nevertheless, conclusions about community condition were robust to these differences. This probably reflects a high degree of correspondence among many of the preferred assessment parameters, suggesting that benthic assessments are robust to differences in metrics commonly used in benthic assessment approaches.

When there was disparity in interpretation among the benthic ecologists, the differences were generally associated with philosophical issues rather than technical ones. For example, the experts disagreed about whether communities altered by the presence of an invasive species, such as the mussel *Musculista senhousia*, should be classified as an affected site; *M. senhousi*a affects community composition by adding habitat structure and heterogeneity, which can facilitate an increase in species abundance and diversity (Ranasinghe *et al.* 2005). Another example of classification uncertainty related to communities where the presence of a mature filter feeder can lower species richness by impeding recruitment through consumption of larvae. In these examples, the differences in condition classification were limited to a single category because the effects manifested in only a subset of parameters, such as number of taxa; other factors, such as the types of species that were present, minimized differences in interpreting overall condition.

Benthic indices are increasingly being used as an alternative to best professional judgment for assessing condition of benthic community condition in many estuarine and marine systems (Weisberg *et al.* 1997, VanDolah *et al.* 1999, Engle and Summers 1999, Borja *et al.* 2000, Paul *et al.* 2001, Smith *et al.* 2001, Llansó *et al.* 2002, Thompson and Lowe 2004). Most of these indices include abundance or proportions of sensitive and tolerant taxa as important assessment metrics. For the sensitive and tolerant taxa parameters at least, benthic indices might provide a means of improving upon the experts' assessments because the list of species relied upon by an individual expert is

Table 1. Spearman correlation coefficients between rankings of samples by benthic ecologists. Each letter represents a different benthic ecologist.

A. San Francisco Bay
(n=12; $p < 0.001$ in all cases)

	A	B	C	D	E	F	G	H
B	0.93							
C	0.97	0.96						
D	0.94	0.84	0.93					
E	0.95	0.91	0.92	0.87				
F	0.92	0.89	0.92	0.86	0.97			
G	0.97	0.95	0.99	0.93	0.92	0.92		
H	0.97	0.94	0.98	0.93	0.94	0.94	0.99	
I	0.92	0.86	0.89	0.87	0.97	0.98	0.89	0.90

B. Southern California Bays
(n=24; $p < 0.0001$ in all cases)

	A	B	C	D	E	F	G	H
B	0.88							
C	0.91	0.96						
D	0.92	0.90	0.89					
E	0.92	0.93	0.96	0.90				
F	0.92	0.93	0.92	0.93	0.95			
G	0.93	0.92	0.93	0.94	0.92	0.93		
H	0.93	0.91	0.92	0.93	0.93	0.95	0.96	
I	0.81	0.83	0.84	0.80	0.88	0.90	0.80	0.81

Table 2. Condition categories assigned to samples by the benthic experts. Each column represents a different benthic ecologist. Key to condition categories: R: Reference, M: Marginal deviation from reference A = Affected and S = Severely affected.

A. San Francisco Bay

Sample	A	B	C	D	E	F	G	H	I
1	S	S	S	S	A	A	S	S	S
2	R	R	R	R	R	R	R	R	M
3	R	R	R	M	R	R	R	R	M
4	S	S	S	S	S	S	S	S	S
5	M	R	R	A	R	R	R	R	M
6	A	M	M	M	M	M	M	M	M
7	M	R	R	R	R	R	R	R	R
8	S	S	S	S	S	S	S	S	S
9	S	S	S	S	S	S	S	S	S
10	A	M	A	M	M	M	A	A	M
11	M	R	R	R	R	R	R	R	R
12	A	M	A	A	M	M	A	A	A

B. Southern California Bays

Sample	A	B	C	D	E	F	G	H	I
21	A	M	M	A	M	M	A	A	M
22	M	M	M	M	M	M	M	M	A
23	R	R	R	R	R	R	R	R	M
24	M	M	M	A	M	M	M	M	M
25	R	R	R	R	R	R	R	R	M
26	S	S	S	S	S	S	A	S	S
27	R	R	R	R	R	M	R	R	A
28	S	S	S	A	S	A	S	S	S
29	M	R	R	M	M	M	M	R	M
30	A	M	M	M	A	A	A	A	A
31	A	A	A	M	A	A	A	A	A
32	A	A	M	A	M	A	M	M	A
33	A	M	A	A	A	A	A	A	A
34	S	S	S	S	S	S	A	S	S
35	M	A	M	M	M	M	M	M	A
36	S	S	S	S	S	A	S	S	A
37	R	R	R	R	R	R	M	R	R
38	S	S	S	S	S	S	A	S	A
39	A	S	S	S	S	S	S	S	S
40	R	R	R	R	R	R	M	R	R
41	S	A	S	A	S	A	A	A	A
42	A	A	A	A	A	A	A	A	A
43	M	R	M	M	A	M	R	M	M
44	R	R	R	R	M	R	M	R	R

typically limited or is a broad generalization applied to higher-level taxa (e.g., *Gammaridea*). Every species occurring at a site provides information regarding the site condition and indices that integrate the knowledge of multiple experts to capture information across a larger number of taxa may provide a more accurate or nuanced assessment.

Indices also have the advantage of being objective and transparent, but at a potential cost of information loss associated with a formulaic approach that does not incorporate all aspects of expert judgment. For example, reliance on indicator species alone can lead to misapplication when small numbers of individuals are present (Borja and Muxika

Table 3. Criteria used by benthic experts to rank and categorize samples. Importance is the average importance for all experts where: 1 = very important; 2 = important, but secondary; 3 = marginally important; 4 = useful, but only to interpret the other factors; and 5 = not used. N is the number (out of nine) experts that used the criterion.

Criteria	Importance	N
Dominance by tolerant indicator taxa	1.0	9
Presence of sensitive indicator taxa	1.2	9
Species richness (No. of taxa)	1.4	9
Abundance of, or dominance by, specific higher level taxa	2.7	8
Total abundance	2.8	9
Presence of nonindigenous species	3.6	6
Diverse functional and feeding groups	3.7	4

Table 4. Indicator taxa identified by the experts. Importance is the average importance for all experts, where 1: very important; 2: important, but secondary; 3: marginally important; 4: useful, but only to interpret the other factors; 5: not used. N is the number of experts that identified the taxon as an indicator.

Indicator Taxon	Importance	N
Tolerant taxa		
Capitella capitata complex	1.0	9
Oligochaeta	1.3	9
Streblospio benedicti	2.0	9
Dorvillea (Schistomeringos) spp.	2.2	8
Mediomastus spp.	2.3	8
Armandia brevis	2.6	7
Pseudopolydora spp.	3.0	7
Exogone spp.	3.0	7
Grandiderella japonica	3.0	8
Euphilomedes spp.	3.1	7
Monocorophium spp.	3.1	8
Neanthes acuminata complex	3.2	7
Musculista senhousia	3.2	7
Notomastus spp	3.4	5
Ophiura spp.	4.7	1
Sensitive taxa		
Ophiuroidea	1.4	8
Amphipod taxa	1.8	8
Gammaridea (most species)	1.9	7
Molluscan taxa	2.2	8
Ampelisca abdita	2.7	6
NCOS*	3.0	6
Corophiidae	3.2	5
Spiophanes duplex and *S. berkeleyorum*	3.2	5
Polychaete dominance	3.6	4
Crustacea	3.7	2
Amphiuridae	4.1	2

* Nemertea, Cnidaria, Opistobranchia and Sipuncula

2005) or soon after a recruitment event that yields large numbers of small juveniles that don't survive to maturity (Dauer *et al.* 1993). These situations would be recognized by experts and assessments adjusted accordingly.

The agreement we found among experts in use of BPJ presents new opportunities for validation of indices. The primary means that has been used previously for validating benthic indices has been to assess whether sites of extreme condition, identified through use of chemical or toxicological measures, can be distinguished by the index. This is a substantial impediment to the development of benthic indices in geographic regions where extreme conditions are rare, because few data are available to evaluate the performance of indices. The agreement among experts found in this study indicates that consensus expert opinion is a viable alternative as an evaluation benchmark, though further study to evaluate the number of experts to reach an appropriate consensus, particularly in other habitats, is warranted.

Consensus expert opinion as an evaluation benchmark may facilitate evaluation of how the indices are performing in assessing sites experiencing intermediate levels of disturbance. This is a more difficult, but more relevant, assessment challenge for indices. The use of expert opinion also provides a benchmark to assess index performance. Index developers have generally identified an index as successful if it correctly differentiated 80% of the extreme sites. A better evaluation benchmark would be that an index predicts sites with a level of correlation comparable to that among experts.

Literature Cited

Bilyard, G.R. 1987. The value of benthic infauna in marine pollution monitoring studies. *Marine Pollution Bulletin* 11:581-585.

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., 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.

Borja, A. and I. Muxika. 2005. Guidelines for the use of AMBI (AZTI's marine biotic index) in the assessment of the benthic ecological quality. *Marine Pollution Bulletin* 50:787-789.

Borja, A. 2005. The European Water Framework Directive: a challenge for nearshore, coastal and continental shelf research. *Continental Shelf Research* 25:1768-1783.

Borja, A. 2006. The new European Marine Strategy Directive: difficulties, opportunities, and challenges. *Marine Pollution Bulletin* 52:239-242.

Boyle, T.P., G.M. Smillie, J.C. Anderson and D.R. Bieson. 1990. A sensitivity analysis of nine diversity and seven similarity indices. *Journal of the Water Pollution Control Federation* 62:749.

Dauer, D.M. 1993. Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin* 26:249-257.

Dauer, D.M., M.W. Luckenbach and A.J. Rodi, Jr. 1993. Abundance biomass comparison (ABC method): effects of an estuarine gradient, anoxic/hypoxic events and contaminated sediments. *Marine Biology* 116:507-518.

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.

Gibson, G.R., M.L. Bowman, J. Gerritsen and B.D. Snyder. 2000. Estuarine and Coastal Marine

Waters: Bioassessment and Biocriteria Technical Guidance. EPA 822-B-00-024. US Environmental Protection Agency, Office of Water. Washington, DC.

Jonge, V.N., M. Elliot and V.S. Brauer. 2006. Marine monitoring: its shortcomings and mismatch with the EU Water Framework Directive's objectives. *Marine Pollution Bulletin* 53:5-19.

Llansó, R.J., L.S. Scott, J.L. Hyland, D.M. Dauer, D.E. Russell and F.W. Kutz. 2002. An estuarine benthic index of biological integrity for the Mid-Atlantic region of the United States. II. Index Development. *Estuaries* 25:1231-1242.

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.

Paul, J.F., K.J. Scott, D.E. Campbell, J.H. Gentile, C.S. Strobel, R.M. Valente, S.B. Weisberg, A.F. Holland and J.A. Ranasinghe. 2001. Developing and applying a benthic index of estuarine condition for the Virginian Biogeographic Province. *Ecological Indicators* 1:83-99.

Pearson, T.H. and R. Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology Annual Review* 16:229 311.

Ranasinghe, J.A., T.K. Mikel, R.G. Velarde, S.B. Weisberg, D.E. Montagne, D.B. Cadien and A. Dalkey. 2005. The prevalence of non-indigenous species in southern California embayments and their effects on benthic macroinvertebrate communities. *Biological Invasions* 7:679-686.

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.

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.

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

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