Marine Pollution Bulletin 60 (2010) 589–600

Contents lists available at ScienceDirect

Marine Pollution Bulletin

journal homepage: www.elsevier.com/locate/marpolbul

Assessing coastal benthic macrofauna community condition using best professional judgement – Developing consensus across North America and Europe

Heliana Teixeira [a,*], Ángel Borja [b], Stephen B. Weisberg [c], J. Ananda Ranasinghe [c], Donald B. Cadien [d], Daniel M. Dauer [e], Jean-Claude Dauvin [f], Steven Degraer [g], Robert J. Diaz [h], Antoine Grémare [i], Ioannis Karakassis [j], Roberto J. Llansó [k], Lawrence L. Lovell [d], João C. Marques [a], David E. Montagne [l], Anna Occhipinti-Ambrogi [m], Rutger Rosenberg [n], Rafael Sardá [o], Linda C. Schaffner [h], Ronald G. Velarde [p]

[a] IMAR, Institute of Marine Research, Faculty of Sciences and Technology, University of Coimbra, 3004-517 Coimbra, Portugal
[b] AZTI–Tecnalia, Marine Research Division, Herrera Kaia Portualdea s/n, 20110 Pasaia, Spain
[c] Southern California Coastal Water Research Project, 3535 Harbor Blvd., Costa Mesa, CA 92626, USA
[d] Sanitation Districts of Los Angeles County, Ocean Monitoring and Research Group, 24501 S. Figueroa St., Carson, CA 90745, USA
[e] Department of Biological Sciences, Old Dominion University, Norfolk, VA 23529, USA
[f] Université de Lille 1 Laboratoire d'Océanologie et de Géosciences, UMR CNRS 8187 LOG, Station Marine de Wimereux, BP 80, F-62930 Wimereux, France
[g] Royal Belgian Institute of Natural Sciences, Management Unit of the North Sea Mathematical Models, Marine Ecosystem Management Section, Gulledelle 100, 1200 Brussels, Belgium
[h] Department of Biological Sciences, School of Marine Science, Virginia Institute of Marine Science, The College of William and Mary, Gloucester Point, VA 23062, USA
[i] Université Bordeaux 1, UMR 5805, EPOC, Station Marine d'Arcachon, 2 Rue du Pr Jolyet, 33120 Arcachon, France
[j] University of Crete, Department of Biology, Marine Ecology Lab, GR-71409 Iraklion, Crete, Greece
[k] Versar, Inc., 9200 Rumsey Road, Columbia, MD 21045, USA
[l] P.O. Box 2004, Penn Valley, CA 95946, USA
[m] Dept. of "Ecologia del Territorio", Section of Ecology, Via S.Epifanio 14, I-27100 Pavia, Italy
[n] Department of Marine Ecology, University of Gothenburg, Kristineberg 566, 450 34 Fiskebøckskil, Sweden
[o] Centre d'Estudis Avançats de Blanes, CSIC, Cta. Accés a la Cala Sant Francesc, 14, 17300 Blanes, Girona, Spain
[p] City of San Diego, Marine Biology Laboratory, 2392 Kincaid Road, San Diego, CA 92101, USA

ARTICLE INFO

Keywords:
Best professional judgment
Coastal benthic macrofauna
Anthropogenic disturbance
Quality assessment
North America
Europe

ABSTRACT

Benthic indices are typically developed independently by habitat, making their incorporation into large geographic scale assessments potentially problematic because of scaling inequities. A potential solution is to establish common scaling using expert best professional judgment (BPJ). To test if experts from different geographies agree on condition assessment, sixteen experts from four regions in USA and Europe were provided species-abundance data for twelve sites per region. They ranked samples from best to worst condition and classified samples into four condition (quality) categories. Site rankings were highly correlated among experts, regardless of whether they were assessing samples from their home region. There was also good agreement on condition category, though agreement was better for samples at extremes of the disturbance gradient. The absence of regional bias suggests that expert judgment is a viable means for establishing a uniform scale to calibrate indices consistently across geographic regions.

1. Introduction

Benthic invertebrate community condition is used worldwide to assess the effects of many impacts, including physical disturbance, organic loading and chemical contamination (Pearson and Rosenberg, 1978; Dauer et al., 2000; Borja et al., 2000, 2003; Muxika et al., 2005). These assessments often use benthic indices to translate community composition into a quality classification (Weisberg et al., 1997, 2008; Van Dolah et al., 1999; Borja et al., 2000, 2004b; Rosenberg et al., 2004; Dauvin and Ruellet, 2007; Dauvin et al., 2007; Muxika et al., 2007; Ranasinghe et al., 2009). In a recent review of ecological indicators for coastal and estuarine systems, Marques et al. (2009) presented most of the benthic indices available, many of which have proven to be accurate and sensitive indicators of the condition of the sediments in which benthos live (Diaz et al., 2004; Marques et al., 2009; Pinto et al., 2009).

Using benthic indices for assessment over large geographic areas can be problematic though, because they are usually developed within specific habitats and ecoregions (Borja and Dauer, 2008). Benthic species composition varies naturally across habitats

* Corresponding author. Address: IMAR – Institute of Marine Research, a/c Departamento de Zoologia, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, 3004-517 Coimbra, Portugal. Tel.: +351 239 836386; fax: +351 239 823603.
E-mail address: helianateixeira@ci.uc.pt (H. Teixeira).

0025-326X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.marpolbul.2009.11.005

and expectations for reference conditions should vary accordingly (de Paz et al., 2008; Borja et al., 2009a). Consequently, there is no certainty that indices developed in different regions or habitats assess biological condition on the same scale. Interpreting different benthic indices developed for different habitats to yield a common assessment for management purposes is further complicated when the indices are based on different combinations of metrics (Diaz et al., 2004; Borja et al., 2009a,b).

One potential solution is to apply expert best professional judgment (BPJ) to establish a set of samples across regions that provide a uniform scale for calibrating any index, but this assumes there is consensus about benthic community condition classifications among experts across regions. Weisberg et al. (2008) found a high level of agreement in expert BPJ in a benthic quality assessment for two United States West Coast habitats, but that assessment was limited to experts from within the region making an assessment of biota with which they had great familiarity. Agreement in benthic condition assessments of experts with varying familiarity with resident benthic fauna would be necessary for establishment of a credible scale applicable across broader geographic regions.

Here, we evaluate the level of agreement among experts using BPJ to assess the condition of marine coastal benthic communities from four widely separated geographic regions. Our objectives were to evaluate whether (1) BPJ assessments were independent of the home regions of the experts, and (2) whether the level of agreement among expert BPJ was sufficient to establish a universal benthic assessment scale for the four regions that could be used to intercalibrate benthic indices and assessment methodologies across habitat boundaries.

2. Methods

Sixteen benthic experts from four geographic regions were provided species-abundance data for twelve sites from each region and asked to determine the condition of the benthos at each site. The four regions included the West (W) and East (E) coasts of the United States (US), and the Atlantic (A) and Mediterranean (M) coasts of Europe. Of the 16 benthic ecologists, nine were from academic institutions, four from municipalities that implement benthic monitoring programs to assess the effect of discharge outfalls, two from non-profit research organizations, and one from a private consulting firm. Their experience in benthic monitoring ranged from 16 to 38 years. Each benthic ecologist was provided species-abundance data for each sample and limited habitat data (region, salinity, depth, and percent fines as a measure of sediment grain size) sufficient to establish an expectation for what kinds of organisms should occur there under undisturbed conditions.

The experts were asked to rank the relative condition of the sites from "best" to "worst" within each region as well as across all four regions. "Best" means least likely to have been disturbed while "worst" means most likely to have been subjected to disturbance, with ties designated as liberally as each expert desired. The experts were also asked to assign each site to one of four condition categories based on narrative descriptions: (1) "unaffected": a community at a least affected or unaffected site; (2) "marginal deviation from unaffected": a community that shows some indication of stress, but within the measurement error of unaffected condition; (3) "affected": where there is confidence that the community shows evidence of physical, chemical, natural, or anthropogenic stress; and (4) "severely affected": where the magnitude of stress is high. The experts were also asked to identify the criteria they used to evaluate the benthos and rate their importance as follows: (1) very important; (2) important, but secondary; (3) marginally important; (4) useful, but only to interpret other factors. Criteria that were not used by an expert were assigned a rank of five for the purpose of calculating an average importance of that attribute among the experts. Since many of the experts identified tolerant and sensitive indicator species as evaluation criteria, they were also asked to list their indicator species and rank their importance on the same scale.

In each of the four regions, the twelve samples were selected to encompass a range of conditions from unimpacted to highly disturbed, from continental shelf and near shore areas with salinity >30 psu. The US West Coast, European Atlantic Coast, and Mediterranean Coast samples were collected with 0.1 m^2 Van Veen grabs and sieved through 1 mm screens, while the US East Coast samples were collected with 0.04 m^2 Young grabs and sieved through 0.5 mm screens. For consistency, abundances for the US East Coast samples were standardized to 0.1 m^2. The data sets from which the samples were selected, and the assessment measures used to order them, are described below.

2.1. United States West Coast

Twelve samples were selected from 493 in the data set used by Smith et al. (2001) to develop the benthic response index (BRI). These samples were collected between 1973 and 1994, from 25–130 m depths along the southern California mainland shelf. Samples were ordered by their BRI values and selected at even BRI intervals.

2.2. United States East Coast

Samples were selected from a 338 sample data set collected between Cape Cod, Massachusetts and the mouth of Chesapeake Bay, Virginia, by the US Environment Protection Agency (EPA) for the Virginian Province Coastal Environmental Monitoring and Assessment Program (Strobel et al., 1995), the New York–New Jersey Harbor Regional Environmental Monitoring and Assessment Program (Adams et al., 1998), and the Mid Atlantic Integrated Assessment (US Environmental Protection Agency, 1998). Samples were selected by arranging the data set according to their effects-range median (ERM) quotients (Long et al., 2000, 2006) and picking twelve samples at even ERM quotient intervals.

2.3. European Atlantic Coast

Twelve samples from Spain (2), the United Kingdom (5), Ireland (1), Belgium (2), Denmark (1) and Norway (1) were selected from the European dataset of 589 samples used to intercalibrate four different methodologies for assessing benthic quality within the Water Framework Directive (WFD) (Borja et al., 2007, 2009b). Samples were ordered from best to worst using the Ecological Quality Ratio (EQR; EC, 2000) and selected at even intervals. Only samples classified in the same WFD ecological status for all four methodologies and with EQR standard error <0.1 among the four methodologies were included.

2.4. European Mediterranean Coast

Twelve samples were selected from published (Muxika et al., 2005) and unpublished data compiled by AZTI–Tecnalia from three areas in Spain and three areas in Greece. Samples were ordered from best to worst and selected at even intervals using several measures, with generally coincident assessments using biotic indices such as the AZTI's marine biotic index (AMBI) (Borja et al., 2000); trophic indices, such as the infaunal trophic index (ITI) (Word, 1978, 1980a,b, 1990); and multivariate analyses.

Table 1
Condition categories assigned by the benthic experts to each of the 48 samples. EU: Europe, US: United States, A: Atlantic, M: Mediterranean, E: East Coast; W: West Coast. Key to condition categories: 1 – "unaffected"; 2 – "marginal deviation from unaffected"; 3 – "affected"; 4 – "severely affected".

	EU Atlantic experts				Mediterranean experts				US East Coast experts				US West Coast experts			
Samples	A1	A2	A3	A4	M1	M2	M3	M4	E1	E2	E3	E4	W1	W2	W3	W4
EU_A1	3	1	3	3	1	1	3	3	3	3	3	1	1	1	2	2
EU_A2	2	1	3	2	2	1	3	3	1	3	3	2	3	1	2	3
EU_A3	1	1	1	1	1	1	1	2	1	1	2	1	1	2	1	1
EU_A4	4	3	4	3	4	3	4	4	3	4	4	3	4	4	3	4
EU_A5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
EU_A6	3	2	4	2	3	3	3	3	3	3	3	2	3	3	2	3
EU_A7	1	1	1	1	2	1	1	3	1	2	1	1	2	1	1	3
EU_A8	4	3	4	2	4	4	4	4	3	4	4	3	4	4	3	4
EU_A9	2	1	1	1	1	1	1	2	1	1	3	2	2	1	2	2
EU_A10	3	3	4	2	4	3	4	3	3	4	3	2	4	4	3	4
EU_A11	1	1	1	1	2	3	1	2	1	2	1	1	1	1	1	1
EU_A12	2	2	4	4	3	3	3	3	4	4	3	3	3	1	3	3
EU_M1	1	2	4	1	3	1	1	3	1	3	1	1	2	2	2	3
EU_M2	1	1	1	1	1	1	1	2	2	1	2	1	1	1	3	2
EU_M3	2	1	3	2	2	1	2	3	2	3	1	1	2	2	2	3
EU_M4	4	3	4	3	4	4	4	4	3	4	3	3	4	4	3	4
EU_M5	4	2	4	3	3	3	4	4	3	4	3	1	3	4	3	4
EU_M6	3	3	4	3	3	3	4	4	3	4	3	2	3	2	3	4
EU_M7	4	3	4	3	4	3	4	4	3	4	3	2	4	4	3	4
EU_M8	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
EU_M9	1	1	1	1	2	1	1	2	1	2	1	2	1	1	1	3
EU_M10	2	2	1	3	1	1	2	2	2	3	2	1	2	1	3	1
EU_M11	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
EU_M12	1	1	3	1	2	1	1	2	1	3	1	2	2	1	2	3
US_E1	2	1	3	1	3	2	2	3	1	3	2	4	2	2	2	2
US_E2	1	1	1	2	2	3	1	2	1	2	1	3	1	2	1	1
US_E3	2	1	2	1	2	3	1	2	1	2	1	2	2	2	1	1
US_E4	2	1	3	2	3	2	3	3	1	3	3	3	3	3	3	2
US_E5	2	1	2	1	2	1	3	3	2	2	2	2	1	1	2	2
US_E6	4	3	3	3	1	1	3	3	3	3	3	1	1	1	2	2
US_E7	2	2	3	2	3	2	3	3	2	4	3	4	3	3	3	2
US_E8	2	2	3	2	3	2	3	3	3	3	3	4	3	3	3	4
US_E9	2	1	3	1	2	1	2	3	1	2	2	1	2	2	1	2
US_E10	2	1	2	1	1	1	1	2	1	2	2	1	1	2	1	1
US_E11	2	2	4	2	4	4	2	2	2	2	1	2	3	3	3	3
US_E12	2	1	2	2	1	1	2	2	1	2	1	2	2	1	2	1
US_W1	2	2	2	2	3	3	1	2	2	3	3	2	3	3	2	3
US_W2	3	1	2	2	1	2	1	2	1	2	2	2	2	1	1	1
US_W3	3	1	3	3	3	4	3	3	2	3	3	3	3	3	2	3
US_W4	1	1	1	1	2	2	1	2	1	2	1	1	1	2	1	2
US_W5	4	3	4	4	4	3	4	4	3	4	4	4	4	4	3	4
US_W6	2	1	2	2	3	3	1	2	2	3	3	2	2	3	2	3
US_W7	2	1	1	2	1	1	1	2	1	2	3	1	1	1	1	1
US_W8	1	1	1	1	2	2	1	2	1	1	3	1	1	1	1	1
US_W9	4	3	4	3	4	3	4	4	3	4	4	4	4	4	3	4
US_W10	3	2	3	3	3	4	3	3	2	4	3	4	3	3	3	3
US_W11	2	1	1	2	1	1	1	2	1	1	3	1	2	1	1	1
US_W12	4	3	4	4	4	3	4	4	3	4	3	3	4	4	3	4

2.5. Data analysis

Patterns attributable to familiarity of experts with "home region" fauna were evaluated in three ways, using regional assessments. First, the sample categorization of each expert was compared to the median categorization of experts from that region, and was quantified as the sum of the deviations (including the positive or negative sign) from the median category for each set of regional samples. Second, Permutational Multivariate Analysis of Variance (PERMANOVA) was used to determine whether there were significant differences in category assignments among groups of experts. The experimental design for this PERMANOVA (Anderson, 2001; McArdle and Anderson, 2001) included 'Sample Region' (four ecoregions) and 'Expert Region' (four ecoregions) as fixed factors, and a third 'Experts' (four levels) fixed factor nested within the 'Expert Region' factor, with $n = 12$ samples for each 'Sample Region' × 'Expert Region' × 'Experts' block. Bray–Curtis dissimilarities were used as distance measures in the PERMANOVA and

Table 2
Expert condition rankings for all 48 samples. EU: Europe, US: United States, A: Atlantic, M: Mediterranean, E: East Coast, W: West Coast, SD: standard deviation.

Samples	EU Atlantic experts				Mediterranean experts				US East Coast experts				US West Coast experts				SD
	A1	A2	A3	A4	M1	M2	M3	M4	E1	E2	E3	E4	W1	W2	W3	W4	
EU_A1	37	21	27	42	1	10	35	31	36.5	20	26.5	11	11	2	27	14.5	12.8
EU_A2	29	20	31	29	23	10	31	36	18	33	28	23	27	8	28	29	7.8
EU_A3	4	2	3	2	11	10	7	2	4.5	2	17.5	2	3	19	8	2	5.6
EU_A4	41	40	44	39	46	41	43	43	32.5	43	46.5	39.5	41	38	33	38.5	3.9
EU_A5	48	48	48	48	48	48	47.5	46	48	48	48	45	48	48	48	48	0.9
EU_A6	35	32	38	27	31	30	35	34	44	22	41.5	23	36	35	26	35	6.1
EU_A7	7	13	13	14	24	10	18	27	18	11	11	11	17	7	14	29	6.6
EU_A8	46	44	47	31	45	46	43	47	32.5	46	46.5	39.5	46	42	44	38.5	5.0
EU_A9	12	5	6	12	7	10	18	10	18	3	26.5	11	15	9	16	14.5	5.9
EU_A10	34	45	43	24	39	38	39	44	32.5	39	41.5	32.5	44	39	34	40	5.5
EU_A11	8	12	10	7	22	32	7	15	18	12	11	11	8	5	9	6	7.0
EU_A12	30	36	36	46	34	31	35	35	46	47	25	32.5	32	6	40	29	9.7
EU_M1	10	27	34	13	26	10	18	29	11.5	29	7	16	24	24	21	24.5	8.2
EU_M2	11	7	7	9	8	10	7	16	28	5	19.5	11	5	15	39	16.5	9.2
EU_M3	14	11	26	30	20	10	20.5	28	32.5	28	7	23	21	23	17	24.5	7.5
EU_M4	42	39	46	38	44	45	43	48	43	45	31.5	39.5	45	45	43	43.5	4.0
EU_M5	44	31	39	36	35	40	37.5	37	40.5	37	29.5	23	38	40	35	43.5	5.3
EU_M6	36	46	33	41	36	35	37.5	38	45	42	31.5	39.5	37	37	42	36.5	4.0
EU_M7	39	41	40	37	40	39	43	40	42	38	29.5	23	39	41	36	43.5	5.2
EU_M8	3	6	1	4	3	10	7	3	4.5	6	7	2	2	16	3	2	3.8
EU_M9	6	18	9	6	18	10	7	17	11.5	13	7	16	7	17	7	24.5	5.7
EU_M10	13	30	8	40	2	10	20.5	11	29.5	27	19.5	32.5	20	11	41	12	11.8
EU_M11	47	47	45	47	47	47	47.5	45	47	44	43	45	47	47	47	47	1.4
EU_M12	9	19	29	10	21	10	7	12	11.5	24	7	16	22	18	18	24.5	6.8
US_E1	28	24	30	11	25	23	25.5	22	11.5	21	15.5	32.5	25	27	25	19	6.1
US_E2	5	4	11	19	15	29	7	4	4.5	14	2.5	23	9	25	10	6	8.3
US_E3	22.5	10	15	8	19	28	7	19	11.5	15	2.5	23	19	26	11	9.5	7.3
US_E4	22.5	14	23	23	30	26	24	23	18	31	23.5	32.5	29	30	32	21.5	5.2
US_E5	18	23	20	16	13	10	29	24	29.5	17	13.5	23	10	12	23	19	6.2
US_E6	45	43	22	43	9	10	33	30	38.5	23	21.5	6	12	10	22	19	13.2
US_E7	24	28	24	28	33	25	28	20	26.5	34	23.5	45	30	31	45	21.5	7.3
US_E8	19.5	35	25	32	37	24	31	25	40.5	32	21.5	45	31	34	46	36.5	7.9
US_E9	17	22	28	15	14	10	22	26	18	16	13.5	6	18	22	12	16.5	5.8
US_E10	21	9	14	5	5	10	15	7	18	10	15.5	6	13	20	13	9.5	5.1
US_E11	27	33	35	25	38	42	23	33	26.5	18	2.5	32.5	28	36	29	33	9.2
US_E12	19.5	15	17	26	6	10	25.5	14	11.5	8	2.5	32.5	23	13	20	11	8.1
US_W1	25	29	18	22	29	27	7	18	23.5	25	36	32.5	33	33	19	33	7.6
US_W2	33	26	19	18	12	22	15	13	4.5	7	17.5	16	14	1	1	2	9.1
US_W3	31.5	25	21	33	27	43	27	21	23.5	30	38.5	32.5	34	29	24	29	6.1
US_W4	2	1	4	3	16	21	7	8	4.5	19	11	2	4	21	4	13	7.1
US_W5	43	42	42	45	43	36	43	42	38.5	41	44.5	32.5	40	43	37	43.5	3.4
US_W6	15	16	16	17	28	34	15	5	23.5	26	38.5	23	26	28	15	29	8.6
US_W7	26	8	12	20	10	10	7	9	4.5	9	34	6	6	14	6	6	8.2
US_W8	1	3	2	1	17	20	7	1	4.5	1	34	16	1	3	2	6	9.5
US_W9	38	37	37	34	41	37	43	39	35	36	44.5	45	43	46	30	43.5	4.6
US_W10	31.5	34	32	35	32	44	31	32	23.5	35	38.5	45	35	32	38	33	5.2
US_W11	16	17	5	21	4	10	7	6	4.5	4	34	6	16	4	5	6	8.4
US_W12	40	38	41	44	42	33	43	41	36.5	40	38.5	45	42	44	31	43.5	4.0

distances were maintained (i.e. not replaced by their ranks) in the analysis. About 4999 permutations were used to achieve an α-level of 0.05 (Anderson, 2005). Third, Spearman rank correlation coefficients (ρ) were used to assess whether levels of agreement in categorizing and ranking sites differed between experts' home regions and other regions. Categories and rankings of experts for each region were compared with the respective regional medians.

The level of agreement on condition categories among all the experts was evaluated using Kappa analysis (Cohen, 1960; Landis and Koch, 1977) by establishing moderate, good, very good, and almost perfect levels of agreement using the equivalence table of Monserud and Leemans (1992). Fleiss–Cohen weights were applied (Fleiss and Cohen, 1973) because misclassifications between distant categories (e.g., between "unaffected" and "affected", or "unaffected" and "severely affected") are more important than misclassifications between closer categories (e.g., between "unaffected" and "marginal deviation from unaffected", or "affected" and "severely affected").

The level of agreement in ranking sites among all the experts was evaluated using Spearman rank correlation analysis to measure associations between sample rankings by each expert and the median of the expert rankings. The variability of the expert rankings for each sample was measured by the median absolute deviation (MAD). Samples were ordered by median rank across all experts and MADs determined as the median of the absolute values of differences between expert ranks and this rank order.

3. Results

There was substantial agreement in condition categories assigned by the experts (Table 1). At least half of the experts agreed on sample condition category for 42 out of the 48 samples. Although there was complete agreement among the experts for only two samples and agreement among 15 of the 16 experts for only one other, at least seven experts agreed on the condition cat-

Table 3
Deviation of expert categories from the median for local experts for each set of regional samples. Expert category deviation sums for regional groups of experts at each regional data set are also presented. Home region results are highlighted. EU: Europe, US: United States.

Samples sets	Experts																			
	A1	A2	A3	A4	EU Atlantic	M1	M2	M3	M4	Mediterranean	E1	E2	E3	E4	US East Coast	W1	W2	W3	W4	US West Coast
EU Atlantic	1.5	−5.5	5.5	−2.5	−1	2.5	−0.5	3.5	7.5	13	−0.5	6.5	5.5	−3.5	8	3.5	−1.5	−1.5	5.5	6
Mediterranean	−1.5	−5.5	4.5	−3.5	−6	0.5	−5.5	−0.5	6.5	1	−3.5	6.5	−4.5	−8.5	−10	−0.5	−2.5	0.5	6.5	4
US East Coast	−1.5	−9.5	4.5	−6.5	−13	0.5	−3.5	−0.5	4.5	1	−7.5	3.5	−2.5	2.5	−4	−2.5	−1.5	−2.5	−3.5	−10
US West Coast	2.0	−9.0	−1.0	0.0	−8	2.0	2.0	−4.0	3.0	3	−7.0	4.0	6.0	−1.0	2	1.0	1.0	−6.0	1.0	−3
Total	0.5	−29.5	13.5	−12.5	−28	5.5	−7.5	−1.5	21.5	18	−18.5	20.5	4.5	−10.5	−4	1.5	−4.5	−9.5	9.5	−3

egory for every sample. In contrast, there were seven samples that were assessed in all four condition categories, but for five of them at least 11 of the 16 experts agreed on their good (“unaffected” or “marginal deviation from unaffected”) or bad (“affected” or “severely affected”) condition. For 32 of the 48 samples, more than 87% of the experts agreed on whether the sample was in good or bad condition.

There also was a great deal of consensus in ranking of samples (Table 2) among the experts. There were a few samples (EU_A1, EU_A12, US_E11, US_W8, and US_W11) that different experts ranked at opposite extremes of the range, but most of the discrepant ranks were attributable to only a few experts.

3.1. Regional consistency of ecological assessments

No regional bias in expert category assignments was observed (Table 3). The distribution of deviations from regional median categories was similar for experts' home regions and other regions. More importantly, regional deviations were less than individual deviations (Table 3). A slight negative deviation was detected in Atlantic expert assessments, with samples from other regions evaluated in better ecological condition categories than the regional medians (Table 3).

Variability in the category assignments was unrelated to whether the assessments were for home regions. There was no statistical significance for any factor related to ‘Expert Region’ in the PERMANOVA (Table 4), indicating that expert category assignments were independent of the regions in which the experts worked. These results also indicated that patterns of US East Coast category assessments were significantly different from patterns for other sets of regional samples.

High correlations were observed among individual expert category assignments and the regional median category for the European Atlantic, Mediterranean, and US West Coast samples, with few Spearman correlation coefficients less than 0.80 (Table 5). In contrast, for the US East Coast samples, 12 of 16 experts' Spearman correlation coefficients were less than 0.80 and six were not statistically significant. However, PERMANOVA (Table 4) showed that category assignments were similar regardless of whether the experts were assessing their home regions or not, although mean correlations among experts were slightly higher within home region samples, except for US East Coast experts (Table 5).

The patterns observed for regional rank evaluations (Table 6) were similar to those for condition category assignments (Table 5). Correlation coefficients for rankings were higher, on average, than for category assignments indicating that consensus between experts was higher when ranking samples than assigning condition categories. For both categorization and ranking, US West Coast experts had a higher-level of within group concordance than the other regional groups of experts (Tables 5 and 6). They were the only regional group of experts with no significant differences between any expert categorizations (Table 4).

3.2. Level of agreement on the ecological assessments

Kappa analysis indicated a high degree of agreement among experts in their condition category assignments (average κ value of 0.65), with levels of agreement varying from moderate to almost perfect and 78.5% of the comparisons agreeing at “Good” or better (Table 7). Mismatches >30% occurred in less than 10% of the comparisons. At the level of good (‘unaffected’/‘marginal deviation from unaffected’) or bad condition (“affected”/“severely affected”), the experts agreed on approximately 80% of the comparisons.

Sample rankings (Table 2) were highly correlated among experts, with an average Spearman correlation coefficient of 0.85 between expert rank and the median rank (Fig. 1). Seven experts (A2,

Table 4
Results of PERMANOVA on Bray–Curtis distances between category assessments of 48 samples from four regions (Sample Region factor: four levels, $n = 12$ samples each), by groups of experts from those regions (Expert Region factor: four levels, each with four experts).

Source	df	SS	MS	*F*
Sample Region	3	5867.64	1955.88	3.56*
Expert Region	3	2633.24	877.75	1.60
ExpReg (Experts)	12	26083.13	2173.59	3.96**
Sample Region × Expert Region	9	1613.36	179.26	0.33
Sample Region × ExpReg (Experts)	36	14934.57	414.85	0.76
Residual	704	386401.10	548.87	
Total	767	437533.03		
Pair-wise a posteriori comparisons: Sample Region				
Atlantic vs. Mediterranean	0.96			
Atlantic vs. East Coast	3.14**			
Atlantic vs. West Coast	0.84			
Mediterranean vs. East Coast	2.26*			
Mediterranean vs. West Coast	0.52			
East Coast vs. West Coast	2.30*			
Experts	Atlantic	Mediterranean	East Coast	West Coast
Expert Region				
Expert 1 vs. Expert 2	3.00*	1.23	3.71**	0.79
Expert 1 vs. Expert 3	0.96	0.98	2.25*	0.76
Expert 1 vs. Expert 4	1.14	2.44*	0.67	0.61
Expert 2 vs. Expert 3	3.59*	0.35	1.40	0.61
Expert 2 vs. Expert 4	1.87	3.79**	2.99*	1.30
Expert 3 vs. Expert 4	1.93	3.39**	1.57	1.38

Pair-wise *a posteriori* tests using the *t*-statistic between Sample Regions, and between Experts within each region.
* $P \leqslant 0.05$.
** $P \leqslant 0.001$.

Table 5
Spearman correlation coefficients between expert category assignments and the regional median category. A: Atlantic; M: Mediterranean; E: East Coast; W: West Coast.

Experts	Atlantic	Mediterranean	East Coast	West Coast
A1	0.92	0.89	0.43*	0.75
A2	0.87	0.87	0.67	0.88
A3	0.88	0.90	0.63*	0.91
A4	0.86	0.80	0.50*	0.82
Mean	0.88	0.86	0.56	0.84
M1	0.78	0.96	0.52	0.97
M2	0.72	0.88	0.08*	0.76
M3	0.94	0.88	0.85	0.88
M4	0.89	0.94	0.72	0.88
Mean	0.83	0.92	0.54	0.87
E1	0.87	0.80	0.66	0.97
E2	0.92	0.94	0.92	0.93
E3	0.89	0.75	0.82	0.61
E4	0.79	0.62	0.53*	0.86
Mean	0.87	0.78	0.73	0.84
W1	0.78	0.96	0.50*	0.91
W2	0.69	0.94	0.44*	0.99
W3	0.91	0.61	0.82	0.94
W4	0.80	0.92	0.68	0.99
Mean	0.80	0.86	0.61	0.96

$n = 12$; *non-significant correlations: $P \geqslant 0.05$.

Table 6
Spearman correlation coefficients between expert regional sample ranks and the median regional rank. A: Atlantic, M: Mediterranean, E: East Coast; W: West Coast.

Experts	Atlantic	Mediterranean	East Coast	West Coast
A1	0.94	0.83	0.55*	0.62
A2	0.99	0.84	0.74	0.76
A3	0.98	0.93	0.64	0.79
A4	0.80	0.68	0.58*	0.82
Mean	0.92	0.82	0.63	0.75
M1	0.83	0.96	0.55*	0.90
M2	0.84	0.87	0.02*	0.70
M3	0.98	0.87	0.73	0.77
M4	0.94	0.99	0.54*	0.81
Mean	0.90	0.93	0.46	0.79
E1	0.84	0.81	0.73	0.92
E2	0.92	0.89	0.98	0.92
E3	0.85	0.75	0.86	0.85
E4	0.94	0.74	0.59	0.82
Mean	0.89	0.80	0.79	0.88
W1	0.91	0.93	0.69	0.92
W2	0.62	0.96	0.44*	0.99
W3	0.92	0.63	0.92	0.90
W4	0.85	0.92	0.88	0.92
Mean	0.82	0.86	0.73	0.93

$n = 12$; *non-significant correlations: $P \geqslant 0.05$.

A3, M3, M4, E2, W1, and W4) deviated little from the median ranks (Fig. 1). Of the nine that deviated more, five deviated throughout the range (A4, E1, E3, E4, and W3) and four differed primarily for samples in the lower and intermediate ranks (A1, M1, M2, and W2). Overall, the level of agreement between experts was higher at the extremes of the gradient of disturbance than at the centre (Fig. 2). Disagreements with respect to good or bad condition occurred mostly in the intermediate third of samples, where the MAD also was higher, showing that rankings had also higher dispersion near the centre of the gradient (Fig. 2). The three samples with ranking standard deviations > 10 (Table 2) were in the middle third of the gradient (EU_A1, EU_M10, and US_E6). Samples with higher median absolute deviations from the median rank (Fig. 2) were often assigned to three or four categories (Table 1).

The results indicated tendencies in individual experts unrelated to home regions. Assessments by four experts (A2, E1, E2, and M4) deviated from regional medians, with A2 and E1 consistently negative (classifying in better condition than the median), and E2 and M4 consistently positive, classifying in worse condition than the median (Table 3). Within regional groups of experts, *a posteriori* tests showed statistically significant differences between these four experts' category assessments and category assessments of some of the other experts (Table 4).

Table 7
Kappa values with level of agreement in parentheses (lower left) for condition category assignments, and percentage of mismatch between expert classifications (upper right). A: Atlantic, M: Mediterranean, E: East Coast; W: West Coast. Level of agreement: AP – "Almost Perfect"; VG – "Very Good"; G – "Good" and M – "Moderate". The percentage of mismatch is related to the relative number of cases in which one of the experts classified a station as "unaffected" or "marginal deviation from unaffected" and the other as "affected" or "severely affected".

	Percentage of mismatch														
	A1	A2	A3	A4	M1	M2	M3	M4	E1	E2	E3	E4	W1	W2	W3	W4
A1		10.6	25.0	10.6	26.5	22.4	14.6	25.0	10.4	27.1	25.0	30.6	22.4	22.4	23.4	27.7
A2	0.77 (VG)		33.3	16.7	29.2	25.0	22.9	33.3	10.4	35.4	33.3	25.0	25.0	25.0	20.8	33.3
A3	0.65 (G)	0.47 (M)		24.4	16.7	34.7	14.6	6.4	22.9	8.5	28.0	24.4	14.9	23.4	21.7	20.8
A4	0.80 (VG)	0.68 (G)	0.61 (G)		29.2	25.0	18.8	29.2	14.6	27.1	29.2	25.0	25.0	30.6	20.8	33.3
M1	0.58 (G)	0.56 (G)	0.79 (VG)	0.51 (M)		16.7	17.0	19.1	21.3	12.8	22.4	15.6	8.3	8.3	16.7	16.7
M2	0.58 (G)	0.52 (M)	0.48 (M)	0.55 (M)	0.77 (VG)		25.5	36.2	17.8	29.8	30.6	27.7	16.7	16.7	23.4	20.8
M3	0.83 (VG)	0.64 (G)	0.79 (VG)	0.70 (VG)	0.76 (VG)	0.59 (G)		10.4	12.5	16.7	14.6	19.6	10.4	17.0	18.8	27.1
M4	0.69 (G)	0.48 (M)	0.88 (AP)	0.54 (M)	0.73 (VG)	0.44 (M)	0.84 (VG)		22.9	14.6	25.0	27.7	20.8	27.7	29.2	25.0
E1	0.77 (VG)	0.79 (VG)	0.62 (G)	0.78 (VG)	0.64 (G)	0.68 (G)	0.79 (VG)	0.61 (G)		25.0	22.9	27.1	18.8	24.5	18.8	27.1
E2	0.64 (G)	0.46 (M)	0.88 (AP)	0.59 (G)	0.80 (VG)	0.55 (M)	0.78 (VG)	0.80 (VG)	0.60 (G)		18.8	28.3	18.8	22.9	27.1	17.0
E3	0.68 (G)	0.49 (M)	0.47 (M)	0.58 (G)	0.57 (G)	0.44 (M)	0.74 (VG)	0.60 (G)	0.63 (G)	0.65 (G)		33.3	16.7	20.8	29.2	29.2
E4	0.40 (M)	0.45 (M)	0.54 (M)	0.50 (M)	0.75 (VG)	0.55 (G)	0.63 (G)	0.51 (M)	0.46 (M)	0.56 (G)	0.49 (M)		19.1	19.1	19.1	34.8
W1	0.67 (G)	0.62 (G)	0.81 (VG)	0.61 (G)	0.90 (AP)	0.73 (VG)	0.84 (VG)	0.70 (VG)	0.68 (G)	0.76 (VG)	0.72 (VG)	0.67 (G)		8.3	12.5	16.7
W2	0.67 (G)	0.60 (G)	0.67 (G)	0.47 (M)	0.89 (AP)	0.75 (VG)	0.76 (VG)	0.62 (G)	0.57 (G)	0.65 (G)	0.64 (G)	0.64 (G)	0.88 (AP)		16.7	20.8
W3	0.63 (G)	0.66 (G)	0.67 (G)	0.67 (G)	0.72 (VG)	0.60 (G)	0.72 (VG)	0.55 (G)	0.72 (VG)	0.62 (G)	0.55 (M)	0.66 (G)	0.78 (VG)	0.70 (VG)		29.2
W4	0.63 (G)	0.51 (M)	0.75 (VG)	0.48 (M)	0.81 (VG)	0.60 (G)	0.68 (G)	0.70 (G)	0.61 (G)	0.79 (VG)	0.53 (M)	0.48 (M)	0.80 (VG)	0.71 (VG)	0.58 (G)	

3.3. Criteria used by experts

The experts used eight criteria for assessing benthic assemblage condition. Six were used by more than half of the experts, with the other two used by only two experts (Table 8). The three most widely used criteria were "Dominance by tolerant taxa", "Presence of sensitive taxa", and "Biodiversity number of taxa measures". However, they were not equally important to experts from different regions. Mediterranean and US East Coast experts, respectively, considered "Biodiversity number of taxa measures" and "Presence of sensitive taxa" only marginally important. In turn, two other attributes, "Biodiversity community measures" and "Abundance dominance patterns" also were considered important by Mediterranean and US West Coast experts, respectively.

On average, experts deviating from their peers used less than the average of 5.3 criteria used by the others. Experts who consistently assessed samples in a worse condition than the median used an average of 5.0 criteria compared to an average of 2.8 criteria used by experts assessing samples in better condition than the median. The experts indicated that it was not more difficult to evaluate data from non-home regions because genus and family associations across regions permitted extrapolation from knowledge of local fauna. Most experts identified tolerant taxa at the species or genus level, but mostly relied on the presence of higher-level taxonomic groups for sensitive taxa (Table 9). Most frequently recognized as tolerant taxa were Polychaetes from the *Capitella capitata* complex, *Streblospio benedicti*, *Ophryotrocha* sp., and *Malacoceros fuliginosus*, oligochaetes, and the bivalve *Nucula annulata*. Most commonly identified as sensitive taxa were the Echinoidea, Ophiuroidea (other than Ophiuridae) and Gammaridea higher taxonomic groups, *Amphiodia* spp., and *Tellina agilis*. Different indicator taxa were considered for samples from different regions and, therefore, this list of indicator taxa is not universally applicable throughout all four regions.

4. Discussion

No systematic difference in assessments based on experts' regions of origin was observed, though the level of agreement here was slightly lower than that achieved by Weisberg et al. (2008) in a single region. The slightly higher correlations within the US West Coast group of experts were possibly driven by the particularly close professional ties, since three of them are from the same agency.

There was greater agreement on sample ranks than on sample condition categories. While the experts largely agreed on the relative positions of samples along the disturbance gradient, they had more difficulty establishing assessment thresholds to assign categories. For example, experts A2 and E2 did not differ from the median expert in sample rankings, but there was consistent directional deviation in their condition categories. Other examples of threshold setting being less consensual than sample ranking included experts giving the same rank to a sample, but disagreeing on sample condition (e.g., experts E3 and E4 on samples EU_M5 or EU_M7; Tables 1 and 2). For both types of evaluation, the consensus was less clear near the middle of the disturbance gradient. From a management perspective, having good agreement at the ends of the gradient is of much less utility than having good agreement near its centre. This agreement is of particular importance in categorizations, since the classification of a site has practical implications whose consequences are most apparent at the good/bad threshold (Borja et al., 2009b).

The experts differed in the number of criteria they used for their assessments and those using more criteria generally showed less directional deviation in their category assignments. This is

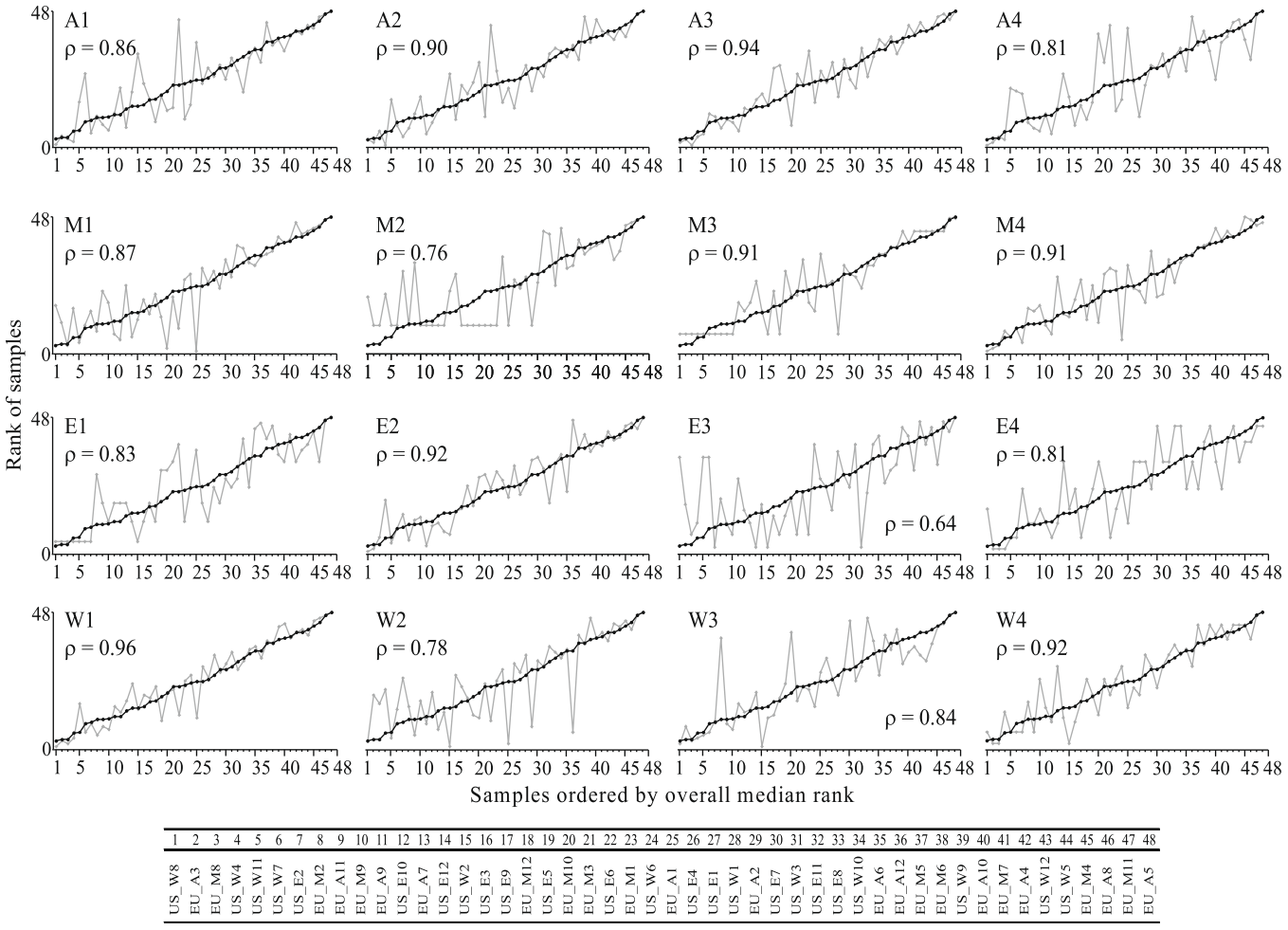

Fig. 1. Deviation of each expert's sample ranks from median ranks along the disturbance gradient (samples ordered by median ranks). Key: A, EU Atlantic; M, Mediterranean; E, US East Coast; W, US West Coast; ρ, Spearman rank correlation; grey dots, expert ranks; black dots, median ranks.

consistent with recommendations to use multiple metrics when assessing ecological status (Weisberg et al., 1997; Borja et al., 2004a, 2007, 2009a; Dauvin et al., 2007; Muxika et al., 2007; Borja and Dauer, 2008; Lavesque et al., 2009). The criteria most widely used by the experts are similar to metrics listed by Alden et al. (2002) as having the greatest discriminatory power within the Chesapeake Bay Benthic Index of Biotic Integrity (B-IBI). However, the number of criteria used was not the only factor affecting individual expert tendencies. Experts who placed higher importance on dominance of tolerant, or presence of sensitive taxa often rated sites more negatively than the average expert. In contrast, those who tended to classify samples in better condition than the median, besides using considerably fewer attributes, often disregarded tolerant species presence or sensitive species presence or both, or did not give any of these criteria the top importance. This was evident in samples with low species richness but high quality species present, and those with high species richness as well as a high percentage of *C. capitata* or other indicators of poor condition (e.g., samples US_E1; US_W3; EU_M5). The use of complementary criteria that measure different benthic community attributes is therefore recommended. Including the presence or dominance of indicator species minimizes the risk of misclassifying disturbed communities as undisturbed.

Some of the differences in how much emphasis experts placed on use of sensitive and tolerant taxa may have been due to their inability to identify relevant taxa outside of their home region. The experts suggested that this was less of a problem for sensitive taxa, which they tended to identify at higher taxonomic levels. In contrast, tolerant taxa tended to be identified at the species level and required local knowledge. For instance, in sample EU_A1 most European Atlantic and Mediterranean and US East Coast experts associated the dominance of *Amphiura filiformis* with organic enrichment, while US West Coast experts considered it a sensitive species. This raises the possibility that species occurring over wide geographical areas may indicate different ecological conditions in different regions. Benthic indices based on indicator species (e.g., AMBI; Borja et al., 2000) may need to adjust accordingly when expanding geographic application (Borja et al., 2008).

Individual expert approaches to assigning condition categories and dealing with uncertainty explain many condition category differences. Some experts, assuming a balanced gradient of disturbance from good to bad, simply split the ranked samples into four classes; others divided the range of values for different metrics by four; and others assigned categories depending on how well benthic community characteristics fit their conceptual view. In doubt, several experts assigned ties to sample rankings. For samples in between categories or on category boundaries, some experts always chose the lower condition category.

However, the full gradient of disturbance might not have been truly achieved for all regional datasets, weakening the assumption of balanced samples across the four categories and contributing to the lower overall level of agreement on categorizations. The

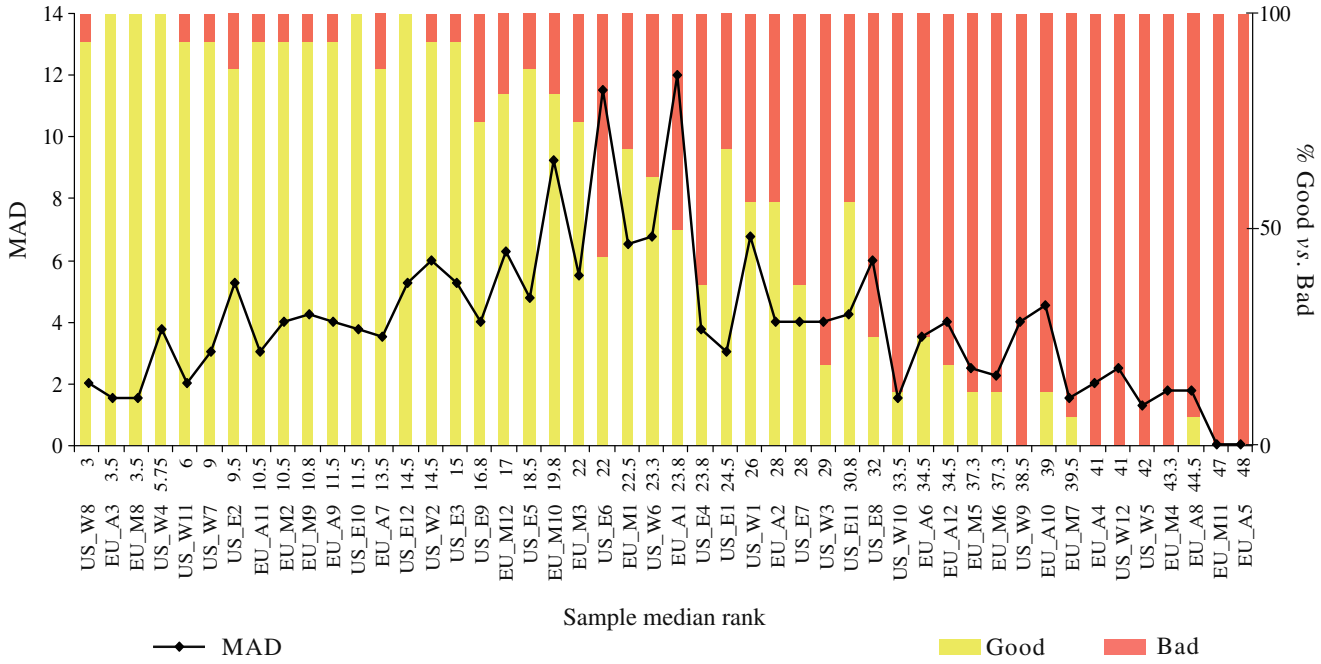

Fig. 2. Median absolute deviations (MAD) from the median rank along the disturbance gradient. Samples are ordered by median rank with histogram bars showing the percentage of experts classifying samples in "Good" ("unaffected" or "marginal deviation from unaffected") and "Bad" ("affected" or "severely affected") condition. Key: EU, Europe; US, United States; A, Atlantic; M, Mediterranean; E, East Coast; W, West Coast.

Table 8
Criteria used by benthic experts to rank and categorize samples. EU: Europe, US: United States, SD: standard deviation. "Importance" is the average importance for all 16 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 used the criterion.

Criteria	Importance	SD	*N*	Regional average importance			
				EU Atlantic	Mediterranean	US East Coast	US West Coast
Dominance by tolerant taxa	1.8	0.4	14	1.4	1.5	2.1	2.0
Presence of sensitive taxa	2.6	0.5	11	2.8	2.0	3.3	2.3
Biodiversity number of taxa measures	2.7	0.5	13	2.8	3.3	2.0	2.7
Abundance dominance patterns	3.0	0.9	10	3.3	3.8	3.3	1.6
Biodiversity community measures	3.4	1.1	9	3.0	2.0	4.0	4.5
Abundance	3.7	0.9	10	4.0	4.8	3.0	3.0
Complex analyses	4.6	0.9	2	5.0	3.3	5.0	5.0
Invasive and introduced species	4.8	0.2	2	5.0	5.0	4.5	4.8

number of "unaffected" and "marginal deviation from unaffected" categories was higher for US East Coast samples (Table 1) than the other regions, which had samples more evenly distributed across categories (Table 4). While the other regional samples were selected based on characteristics of the biological communities, US East Coast samples were selected based on abiotic factors, and the ERM values used as proxy for disturbance may not have accurately reflected the condition of the local benthic communities.

Another factor that contributed to discrepancies among experts was the challenge of distinguishing anthropogenic disturbance from natural stress, which has most often been identified in estuaries (Dauvin, 2007; Elliott and Quintino, 2007). This was particularly notable for the US East Coast data, which were largely samples from coarse sediments subject to high wave energy or strong currents about which the experts had more disagreements than for the other regions. The high energy led to lower species richness (Hall, 1994) than might otherwise be expected for euhaline areas in that region. Some of the experts ranked the samples as stressed because of low species richness, independent of the cause of the stress. Others recognized the communities as dominated by high energy species, such as bivalves, and modified their species richness expectations accordingly. Thus, the differences in evaluations of these samples can largely be attributed to differences in the interpretation of guidelines on how to deal with natural vs. anthropogenic stress.

This challenge associated with natural stress illustrates that our estimate of the level of agreement among experts was probably a minimum estimate, because we withheld information that they would usually have used when making an assessment. In typical assessments, the experts would know the specific sample locations, which we did not share to avoid interference due to local expert knowledge about particular sites. For example, the experts may have used location specific information to lower their species richness expectations based on susceptibility to wave energy stress. We probably also underestimated the true level of consensus because we asked the experts to conduct their assessments in isolation, where normally they would probably confer with their peers. Following submittal of their site assessments, we held a conference call among the experts to investigate factors that led to differences among them. In many cases, experts deviating from the median indicated that hearing the perspectives (such as the potential for wave energy influence) of the other experts would have

Table 9
Indicator taxa identified by the experts. SD: standard deviation. "Importance" is the average importance for all 16 experts, where: 1, very important; 2, important, but secondary; 3, marginally important; 4, useful but only to interpret the other factors; 5, taxa mentioned but not its importance. *N* is the number of experts that referred to the taxa.

Tolerant taxa	Importance	SD	*N*	Sensitive taxa	Importance	SD	*N*
Solemya reidi	1.0	0.0	3	*Lanice conchilega*	1.0		1
Solemya togata	1.0	0.0	4	Sabellidae	1.0		1
Schistomeringos longicornis	1.0	0.0	3	Terebellidae	1.0		1
Ophryotrocha spp.	1.2	0.4	5	Trichobranchidae	1.0		1
Armandia brevis	1.3	0.6	3	*Amphiura* spp.	1.3	0.5	4
Mulinia lateralis	1.5	0.7	2	*Amphiodia* spp. complex	1.4	0.5	5
Raricirrus beryli	1.5	0.7	2	*Ampelisca* spp.	1.5	0.6	4
Capitella capitata complex	1.6	1.2	14	*Proclea* spp.	1.5	0.7	2
Macoma carlottensis	1.8	0.5	4	Gammaridea (Haustoriidae, Phoxocephalidae)	1.8	1.8	5
Parvilucina tenuisculpta	1.8	0.5	4	*Tellina agilis*	1.8	0.8	5
Mediomastus spp.	1.8	0.8	5	*Cyathura burbancki*	2.0		1
Streblospio benedicti	1.8	0.8	6	Echinoidea	2.0	1.5	6
Mollusca	2.0		1	*Anadara transversa*	2.0		1
Corbula gibba	2.0	0.8	4	*Mercenaria mercenaria*	2.0		1
Thyasiridae	2.0	0.0	2	*Mya arenaria*	2.0	0.0	3
Cossura longocirrata	2.0		1	*Nemocardium centifilosum*	2.0		1
Armandia spp.	2.0		1	*Plagiocardium papillosum*	2.0	1.4	2
Eteone heteropoda	2.0		1	*Tellina* spp.	2.0	1.0	3
Euchone incolor	2.0		1	*Timoclea ovata*	2.0	1.4	2
Levinsenia gracilis	2.0		1	Ophiuroidea (other than Ophiuridae)	2.0	1.5	6
Nephtys hombergii	2.0		1	Ampharetidae	2.0		1
Nucula annulata	2.2	0.4	5	Maldanidae	2.0		1
Malacoceros fuliginosus	2.2	1.1	5	*Pectinaria* spp.	2.0	1.4	2
Polydora spp.	2.3	0.5	4	*Ensis directus*	2.3	0.6	3
Prionospio steenstrupi	2.3	0.5	4	*Macoma balthica*	2.3	0.6	3
Axinopsida serricata	2.3	0.6	3	*Listriella goleta*	2.5	2.1	2
Oligochaeta	2.4	1.4	7	*Spisula* spp.	2.5	2.1	2
Dipolydora spp.	2.5	0.7	2	Arthropoda	2.7	2.1	3
Thyasira flexuosa	2.7	1.2	3	*Spisula solidissima*	2.7	1.2	3
Ampelisca spp.	3.0	1.7	3	Mollusca	3.5	1.7	4
Euphilomedes spp.	3.0	1.4	2	*Chaetopterus variopedatus*	3.7	0.6	3
Mysella spp.	3.0	0.0	2	Brachiopoda	3.7	1.2	3
Mytilus edulis	3.0	1.0	3	*Edwardsia* spp.	4.0	1.4	2
Nassarius mendicus	3.0	1.4	2	Crustacea	5.0		1
Amphiuridae	3.0	2.8	2	Bivalvia	5.0		1
Chaetopterus variopedatus	3.0		1	Polychaeta	5.0		1
Aphelochaeta spp.	3.0	1.2	4	*Lumbrineris* spp.	5.0		1
Chaetozone spp.	3.0	1.0	3	*Pista* spp.	5.0		1
Cirratulus spp.	3.0	1.0	3				
Tharyx spp.	3.0	1.0	5				
Cossuridae	3.0		1				
Streblospio spp.	3.0		1				
Lucinidae	3.3	1.5	3				
Paraonidae	3.3	0.6	3				
Pseudopolydora spp.	3.5	1.3	4				
Prionospio spp.	3.7	1.2	3				
Spionidae	3.8	1.5	4				
Ericthonius brasiliensis	4.0		1				
Polychaeta	4.0	1.4	2				
Cirratulidae	4.0	1.0	3				
Polygordius spp.	4.0	1.4	2				
Malacoceros spp.	4.0		1				
Nucula spp.	4.5	0.7	2				

caused them to change their assessment toward the median, if they had been provided that opportunity.

While there were some sites where the experts disagreed, the generally high level of agreement in our study seems to confirm the European WFD suggestion that BPJ is a viable means for calibrating indices of ecosystem condition (Borja, 2005). More importantly, the agreement we observed across large geographies suggests a means for creating a common calibration scale that facilitates national and international comparisons of benthic condition. Once BPJ consensus is achieved for a small subset of samples along a clear disturbance gradient that are representative of a particular habitat, the BPJ scale can be used to intercalibrate distinct benthic indices results. In the context of large-scale regional or national surveys, this approach allows intercalibrating assessments conducted on an unlimited number of samples with any type of methods or indices, which is fundamental to accurately apply large-scale regulatory quality thresholds.

While the data set from this study has value in that context, it also needs to be expanded. Our data were limited to temperate coastal ocean waters and there are many other geographies and habitats that were not considered here. It is important that any BPJ scale reflect variability associated with anthropogenic disturbance within a habitat. The natural variability across habitats would condemn the use of such a scale since the expectations for community health varies accordingly. Therefore, BPJ scales should be adapted to targeted environments. For example, estuarine habitats, in particular, are a challenge because the distinction between natural and human induced changes is often difficult to infer from community data (Dauvin, 2007; Elliott and Quintino, 2007). Ideally, in the words of Elliott and Quintino (2007), to tackle this "Estuarine Quality Paradox" it is necessary to find methods able to detect anthropogenic stress against a background of natural stress. BPJ can provide an alternative by using criteria more difficult to include in benthic index

formulation, such as those associated with the functioning of the ecosystem.

More importantly, we used the four assessment categories used by Weisberg et al. (2008) and there is a need to map those to the five ecological quality classes on which the WFD is based or to any other new assessment scheme. Category classifications are important because they usually are the basis for different environmental regulatory and management requirements, which may be associated with substantially different cost. Based on the consistency in sample ranking among the experts in the present study, we expect this mapping will easily be accomplished.

Acknowledgements

This study was supported by a Ph.D. Grant (SFRH/BD/24430/2005) from FCT, the Portuguese National Board of Scientific Research and a portion of the Mediterranean Coast data were provided by the ECASA project supported by Commission of the European Communities Contract No. 006540. Some Atlantic data were provided by the European benthic intercalibration group. We also acknowledge contributions and collaboration by J. Germán Rodríguez (AZTI-Tecnalia), Thierry Ruellet (University of Lille1), and Giulia Forni (University of Pavia).

References

Adams, D.A., O'Connor, J.S., Weisberg, S.B., 1998. Sediment Quality of the NY/NJ Harbor System. Final Report. An Investigation Under the Regional Environmental Monitoring and Assessment Program (REMAP). US Environmental Protection Agency, Edison, NJ, EPA/902/R-98/001.

Alden, R.W., Dauer, D.M., Ranasinghe, J.A., Scott, L.C., Llansó, R., 2002. Statistical verification of the Chesapeake Bay benthic index of biotic integrity. Environmetrics 13, 473–498.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 6, 32–46.

Anderson, M.J., 2005. PERMANOVA: A FORTRAN Computer Program for Permutational Multivariate Analysis of Variance. Department of Statistics, University of Auckland, New Zealand.

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., Dauer, D.M., 2008. Assessing the environmental quality status in estuarine and coastal systems: comparing methodologies and indices. Ecological Indicators 8, 331–337.

Borja, A., Franco, J., Perez, V., 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., Muxika, I., Franco, J., 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., Franco, J., Valencia, V., Bald, J., Muxika, I., Belzunce, M.J., Solaun, O., 2004a. Implementation of the European Water Framework Directive from the Basque Country (northern Spain): a methodological approach. Marine Pollution Bulletin 48, 209–218.

Borja, A., Franco, J., Muxika, I., 2004b. The biotic indices and the Water Framework Directive: the required consensus in the new benthic monitoring tools. Marine Pollution Bulletin 48, 405–408.

Borja, A., Josefson, A.B., Miles, A., Muxika, I., Olsgard, F., Phillips, G., Rodriguez, J.G., 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., Dauer, D., Diaz, R., Llansó, R.J., Muxika, I., Rodriguez, J.G., Schaffner, L., 2008. Assessing estuarine benthic quality conditions in Chesapeake Bay: a comparison of three indices. Ecological Indicators 8, 395–403.

Borja, A., Ranasinghe, J.A., Weisberg, S.B., 2009a. Assessing ecological integrity in marine waters using multiple indices and ecosystem components: challenges for the future. Marine Pollution Bulletin 59, 1–4.

Borja, A., Miles, A., Occhipinti-Ambrogi, A., Berg, T., 2009b. Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the Water Framework Directive. Hydrobiologia. doi: 10.1007/s10750-009-9881-y.

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

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

Dauvin, J.C., 2007. Paradox of estuarine quality: benthic indicators and indices, consensus or debate for the future. Marine Pollution Bulletin 55, 271–281.

Dauvin, J.C., Ruellet, T., 2007. Polychaete/amphipod ratio revisited. Marine Pollution Bulletin 55, 215–224.

Dauvin, J.C., Ruellet, T., Desroy, N., Janson, A.L., 2007. The ecological quality status of the Bay of Seine and the Seine estuary: use of biotic indices. Marine Pollution Bulletin 55, 241–257.

de Paz, L., Patrício, J., Marques, J.C., Borja, A., Laborda, A.J., 2008. Ecological status assessment in the lower Eo estuary (Spain). The challenge of habitat heterogeneity integration: a benthic perspective. Marine Pollution Bulletin 56, 1275–1283.

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

EC, 2000. Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy. Official Journal of the European Communities (OJ L327) 43, 1–73.

Elliott, M., Quintino, V.M., 2007. The estuarine quality paradox, environmental homeostasis and the difficulty of detecting anthropogenic stress in naturally stressed areas. Marine Pollution Bulletin 54, 640–645.

Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619.

Hall, S.J., 1994. Physical disturbance and marine benthic communities: life in unconsolidated sediments. Oceanography and Marine Biology: An Annual Review 32, 179–239.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lavesque, N., Blanchet, H., de Montaudouin, X., 2009. Development of a multimetric approach to assess perturbation of benthic macrofauna in *Zostera noltii* beds. Journal of Experimental Marine Biology and Ecology 368, 101–112.

Long, E.R., MacDonald, D.D., Severn, C.G., Hong, C.B., 2000. Classifying probabilities of acute toxicity in marine sediments with empirically derived sediment quality guidelines. Environmental Toxicology and Chemistry 19, 2598–2601.

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

Marques, J.C., Salas, F., Patrício, J., Teixeira, H., Neto, J.M., 2009. Ecological Indicators for Coastal and Estuarine Environmental Assessment – A User Guide. WIT Press, Southampton.

McArdle, B.H., Anderson, M.J., 2001. Fitting multivariate models to community data: a comment on distance based redundancy analysis. Ecology 82, 290–297.

Monserud, R., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. Ecological Modelling 62, 275–293.

Muxika, I., Borja, A., Bonne, W., 2005. The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. Ecological Indicators 5, 19–31.

Muxika, I., Borja, A., Bald, J., 2007. Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Marine Pollution Bulletin 55, 16–29.

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

Pinto, R., Patricio, J., Baeta, A., Fath, B.D., Neto, J.M., Marques, J.C., 2009. Review and evaluation of estuarine biotic indices to assess benthic condition. Ecological Indicators 9, 1–25.

Ranasinghe, J.A., Weisberg, S.B., Smith, R.W., Montagne, D.E., Thompson, B., Oakden, J.M., Huff, D.D., Cadien, D.B., Velarde, R.G., Ritter, K.J., 2009. Calibration and evaluation of five indicators of benthic community condition in two California bay and estuary habitats. Marine Pollution Bulletin 59, 5–13.

Rosenberg, R., Blomqvist, M., Nilsson, H.C., Cederwall, H., Dimming, A., 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., Bergen, M., Weisberg, S.B., Cadien, D.B., Dalkey, A., Montagne, D.E., Stull, J.K., Velarde, R.G., 2001. Benthic response index for assessing infaunal communities on the southern California mainland shelf. Ecological Applications 11, 1073–1087.

Strobel, C.J., Buffum, H.W., Benyi, S.J., Petrocelli, E.A., Reifsteck, D.R., Keith, D.J., 1995. Virginian Province Statistical Summary 1990–1993. US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI, EPA/620/R-94/026.

USEPA (US Environmental Protection Agency), 1998. Condition of the Mid-Atlantic Estuaries. US Environmental Protection Agency, Office of Research and Development, Washington, DC, EPA/600/R-98/147.

Van Dolah, R.F., Hyland, J.L., Holland, A.F., Rosen, J.S., Snoots, T.R., 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., Ranasinghe, J.A., Schaffner, L.C., Diaz, R.J., Dauer, D.M., Frithsen, J.B., 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. Estuaries 20, 149–158.

Weisberg, S.B., Thompson, B., Ranasinghe, J.A., Montagne, D.E., Cadien, D.B., Dauer, D.M., Diener, D.R., Oliver, J.S., Reish, D.J., Velarde, R.G., Word, J.Q., 2008. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. Ecological Indicators 8, 389–394.

Word, J.Q., 1978. The infaunal trophic index. In: Southern California Coastal Water Research Project Annual Report. El Segundo, CA, pp. 19–39.
Word, J.Q., 1980a. Extension of the infaunal trophic index to a depth of 800 meters. In: Southern California Coastal Water Research Project Biennial Report 1979–1980. Long Beach, CA, pp. 95–101.
Word, J.Q., 1980b. Classification of benthic invertebrates into infaunal trophic index feeding groups. In: Southern California Coastal Water Research Project Biennial Report 1979–1980. Long Beach, CA, pp. 103–121.
Word, J.Q., 1990. The Infaunal Trophic Index, A Functional Approach to Benthic Community Analyses. PhD Dissertation. University of Washington, Seattle, WA.