ENVIRONMETRICS
Environmetrics 2002; **13**: 473–498 (DOI: 10.1002/env.548)

Statistical verification of the Chesapeake Bay benthic index of biotic integrity

Raymond W. Alden III[1,*,†] D. M. Dauer[2], J. A. Ranasinghe[3], L. C. Scott[4] and R. J. Llansó[4]

[1]*University of Nevada, Las Vegas, NV 89154-1002, U.S.A.*
[2]*Old Dominion University, Norfolk, VA 23523, U.S.A.*
[3]*SCCWRP, Westminster, CA 92683, U.S.A.*
[4]*Versar, Inc., Columbia, MD 21045,U.S.A.*

SUMMARY

The benthic index of biotic integrity (B-IBI) developed for the Chesapeake Bay was statistically verified using simulations and a suite of multivariate statistical techniques. The B-IBI uses a simple scoring system for benthic community metrics to assess benthic community health and to infer environmental quality of benthic habitats in the Bay. Overall, the B-IBI was verified as being sensitive, stable, robust and statistically sound. Classification effectiveness of the B-IBI increased with salinity, from marginal performance for tidal freshwater ecosystems to excellent results for polyhaline areas. The greater classification uncertainty in low salinity habitats may be due to difficulties in reliably identifying naturally unstressed areas or may be due to regional ecotones created by stress gradients. Pollution-indicative species abundance, pollution-sensitive species abundance, and diversity (Shannon's index) were the most important metrics in discriminating between degraded and non-degraded conditions in the majority of the habitats. Single metrics often performed as well as the multi-metric B-IBI in correctly classifying the relative quality of sites. However, the redundancy in the multi-metric B-IBI provided a stable 'weight of evidence' system which increased confidence in general conclusions. Confidence limits developed for the B-IBI scores were used to distinguish among habitats that were degraded, non-degraded, of intermediate quality, or of indeterminate condition. Copyright © 2002 John Wiley & Sons, Ltd.

1. INTRODUCTION

Environmental managers and scientists are often called upon to determine whether estuaries have been impacted by anthropogenic stresses. Recently, researchers have developed quantitative indices of benthic community health as indicators of environmental quality of estuarine ecosystems (e.g. Engle *et al.*, 1994; Weisberg *et al.*, 1997; Hyland *et al.*, 1998; Engle and Summers, 1999; Hyland *et al.*, 1999; Van Dolah *et al.*, 1999). Such indices provide scientists with tools to summarize and communicate the meaning of complex benthic biological data to managers, politicians and the public at large. Thus, the reliability and sensitivity of benthic indices are becoming of increasing importance due to the significance and scope of the environmental assessments and resource management decisions that

*Correspondence to: Raymond W. Alden III, University of Nevada, Las Vegas, NV 89154-1002, U.S.A.
†E-mail: alden@unlv.edu
Contract/grant sponsor: Maryland Department of Natural Resources; contract/grant number: CA-99-03 07-4-30565-3734.

these indices may support. The purpose of this study is to rigorously assess the statistical validity, sensitivity and robustness of a benthic index developed by Weisberg *et al.* (1997) for application in the Chesapeake Bay watershed.

The development of recent estuarine benthic indices has typically involved a calibration and a validation phase. The calibration phase generally involves one of two approaches. The first approach involves the establishment of a quantitative multivariate relationship (e.g. via canonical discriminant analysis) between benthic data sets taken from areas known to be non-stressed (designated as 'reference' sites) compared to biological data from stressed areas (designated as 'degraded'; see, for example, Engle *et al.*, 1994; Engle and Summers, 1999). The second type of approach involves the development of a multimetric system known as a benthic index of biotic integrity (B-IBI) that is based upon expected biological conditions at sites free of anthropogenic stress ('reference' sites; see, for example, Weisberg *et al.*, 1997; Hyland *et al.*, 1998; Hyland *et al.*, 1999; Van Dolah *et al.*, 1999). The validation phase in all of these studies involved the assessment of independent benthic biological data sets from sites of known environmental quality. Validation success was evaluated by comparing the percent correct classification of the sites by the index to a priori classification based upon the degree of anthropogenic stress and/or by the graphical agreement between the values of the biological index and measures of sediment contamination. Thus, the index is validated inferentially by a reasonably high degree of correct classification of reference and degraded sites from the independent data set(s).

There are few studies that have rigorously verified benthic indices through multivariate statistical analyses. The discriminant-based benthic index developed by Engle and Summers (1999) was independently verified by Rakocinski *et al.* (1997) with canonical correspondence analysis (CCA). The CCA separated sites that had been shown previously to have high discriminant-based benthic index scores from those with low scores. Statistical verification has been conducted on freshwater indices (Reynoldson *et al.*, 1997), but no multivariate statistical verifications have been reported for marine or estuarine B-IBIs.

The present study employed a series of multivariate statistical and simulation techniques to evaluate and verify the B-IBI developed by Weisberg *et al.* (1997) to aid in determining the status and trends of environmental conditions in Chesapeake Bay ecosystems. The study used a large data set from sites known to be uncontaminated/unstressed (reference) or degraded from each of seven benthic habitats in the Bay. The evaluation process focused on verifying various aspects of the B-IBI: (i) whether the threshold values employed to calculate scores for B-IBI metrics are effective; (ii) whether the sets of metrics comprising the B-IBI can detect statistically significant differences between sites known to be degraded and sites known to be unstressed (reference); (iii) which of the metrics provide the greatest discriminating power; (iv) the minimum and optimal sets of metrics that should be used to produce defensible B-IBIs for each habitat; (v) confidence limits of B-IBI scores for reference areas and degraded areas in each habitat; (vi) assessment of the sensitivity of the B-IBIs to changes in the raw values of individual metrics.

2. METHODS

2.1. General

The data sets employed in the present study included benthic samples taken from seven habitats defined by Weisberg *et al.* (1997) for the Chesapeake Bay. The macrobenthic infaunal samples were collected during summer months by a number of sampling programs: EPA's Environmental Monitoring and Assessment Program (EMAP; Paul *et al.*, 1992); the Chesapeake Bay Monitoring Program (Dauer and Alden, 1995 for the monitoring program for Virginia; Ranasinghe *et al.*, 1994 for the

monitoring program for Maryland); a James River Study (Diaz, 1989); as well as an ad hoc study that provided supplementary coverage of tidal freshwater and oligohaline habitats (a parallel study by Scott *et al.*, unpublished data). Except for the latter study, the sampling programs were the same ones that provided samples for Weisberg *et al.* (1997). However, Weisberg *et al.* (1997) averaged all samples collected at each site to produce site means, while the present study employed data from individual samples. Except for the tidal freshwater and oligohaline habitats, the habitat-specific sets of metrics (benthic variables or combinations of variables used to calculate B-IBI scores) and methods developed by Weisberg *et al.* (1997) were used to calculate the B-IBI scores for each sample. The tidal fresh and oligohaline data sets had too few samples for Weisberg *et al.* (1997) to calculate B-IBI scores with any degree of certainty. A study by Scott *et al.* (Versar, Inc., unpublished report) produced putative metric/threshold combinations for these habitats that were evaluated during the present study. Table I presents thresholds for these habitats, as well as several corrections discovered for the Weisberg *et al.* (1997) thresholds for other habitats. The Appendix presents the details of the pollution-indicative and pollution-sensitive metrics from tidal freshwater and oligohaline habitats.

The metrics employed to calculate B-IBI scores fall into four general categories described by Weisberg *et al.* (1997): diversity (Shannon–Wiener index); productivity (biomass or total abundance); species composition (percentage of the benthic community represented by pollution-indicative or pollution-sensitive taxa based on relative abundance or relative biomass contribution); and trophic composition (e.g. percentage of total abundance represented by carnivores and omnivores, or deep deposit feeders). As suggested by Weisberg *et al.* (1997), biomass values were used for the productivity and species composition metrics when available. Otherwise, abundance-based metrics were used. The 'depth distribution' metrics presented by Weisberg *et al.* (1997) were not employed in the present study since only a small proportion of the samples in the data sets that were used had information related to stratification of the benthos within the sediments.

When a metric is indicative of good habitat quality (e.g. the relative abundance or biomass of pollution sensitive species, the diversity described by Shannon's index, etc.), a lower threshold for B-IBI score calculations is established by the 5-percentile value of the metric from a reference data set, and a second threshold is established by the median value for the reference data distribution. When applied to metric data from samples from a site of unknown quality, a B-IBI score of '1' is given to samples with values less than the 5-percentile threshold of the reference data set, a B-IBI score of '3' is given to samples with a value between the 5-percentile threshold and the median threshold, and a value of '5' is given to samples with a value that is greater than or equal to the median. Conversely, when a metric describes a biotic characteristic that is indicative of poor habitat quality (e.g. the relative abundance or biomass of pollution indicative species), an upper threshold for the B-IBI score calculations is established by the 95-percentile and a second threshold is established by the median of the reference data set. B-IBI scoring involves the following coding scheme: giving samples with metric values above the 95-percentile threshold a score of '1'; giving samples with values between the median and the 95-percentile threshold a score of '3'; and giving samples with values less than or equal to the median a score of '5'. A third scoring scenario has been developed for metrics that may indicate good quality when the values are not too small or too great (e.g. abundance or biomass could represent poor conditions if found in extremes). For these metrics, both lower (5-percentile) and upper (95-percentile) thresholds are established, along with intermediate thresholds from the 25-percentile and 75-percentile levels of the reference data sets. The scoring of samples involves the following scheme: giving a score of '1' to samples with values that are less than the lower threshold or greater than or equal to the upper threshold; giving a score of '3' to values that are between the 5-percentile and the 25-percentile threshold, or values that are greater than or equal to the 75-percentile and less

Table I. Metrics and thresholds for tidal freshwater and oligohaline habitats developed by Scott *et al.* (Versar, Inc., unpublished). Also presented are corrected thresholds for metrics from the high salinity mesohaline mud habitat for which errata were discovered in the Weisberg *et al.* (1997) article. Intermediate thresholds are for medians, 25th and 75th centiles. See text for explanation of application of thresholds

				Intermediate threshold(s)			
Habitat	Metrics	Units	Low threshold	25th	Median	75th	High threshold
Tidal freshwater	Abundance	# m^{-2}	800	1050		4000	5500
	Abundance of pollution-indicative taxa	%			39		87
	Tolerance score* (Mean of scores for all taxa)	—			8		9.35
	Abundance of deep-deposit feeders	%			70		95
Oligohaline	Abundance of carnivores and omnivores	%	15		35		
	Abundance of pollution-indicative taxa	%			27		95
	Abundance of pollution-sensitive taxa	%	0.2		26		
	Tanypodinae to chironomidae Abundance ratio	%			17		64
	Tolerance score* (Mean of scores for all taxa)	—			6		9.05
	Abundance	# m^{-2}	180	450		3350	4050
High meso. mud	Abundance of pollution-indicative taxa	%			20		50
	Abundance of pollution-sensitive taxa	%	10		30		
	Biomass	g m^{-2}	0.5	2.0		10.0	50
	Abundance of carnivores and omnivores	%	10		25		

*See Lenat, 1993 for scoring methods.

Table II. Number of samples in reference and degraded site data sets from various habitats in Chesapeake Bay

Habitat	Number of reference site samples	Number of degraded site samples
Tidal freshwater	75	211
Oligohaline	32	92
Low salinity mesohaline	60	85
High salinity mesohaline mud	117	189
High salinity mesohaline sand	42	15
Polyhaline mud	72	250
Polyhaline sand	117	11
Total, all habitats	515	853

than 95-percentile threshold; and giving a score of '5' to values that are greater than or equal to the 25-percentile threshold and less than the 75-percentile threshold. Once each metric has been converted to the 1–3–5 scale (hereafter designated as the 'metric B-IBI scores'), the overall B-IBI score for the sample is calculated by averaging the values.

The sites used for the present study were classified a priori by abiotic characteristics (degree of contamination, bottom oxygen conditions and sediment organic content) as to whether they represented 'reference' (non-stressed) or 'degraded' (stressed) conditions. In all analyses using a priori classification, the criteria of Weisberg *et al.* (1997) were used. Overall, 515 reference samples and 853 degraded samples were employed in the study (Table II).

Following the convention established by Weisberg *et al.* (1997), B-IBI scores with values at or above 3.0 are used throughout the present study as the breakpoint between 'reference' and 'degraded' sites when correct classification efficiencies are employed as test criteria.

2.2. Evaluation of metric thresholds

The threshold evaluation investigations were conducted by establishing new thresholds for the reference conditions in an iterative fashion and determining the effects on the correct classification of the groups determined a priori as to quality. New lower thresholds for each metric were tested iteratively for values within the range between the percentile represented by the record with the smallest value in the reference data set and the 25-percentile level. Percent correct classifications produced by the overall B-IBI scores as well as by the metric B-IBI scores were used to determine the effectiveness of new potential thresholds within this range of values. Likewise, metrics with upper thresholds were tested for classification performance by decreasing the threshold between the percentile represented by the record with the greatest value and the 75-percentile of the reference data set. For metrics with both upper and lower thresholds, the values were changed simultaneously (e.g. if the lower threshold was set at the 1-percentile level, the upper threshold was the 99-percentile; for 2-percentile, a corresponding 98-percentile level was produced, etc.) so that the definition of extreme conditions (i.e. values for which scores of '1' were assigned) iteratively expanded in the scoring calculations. During the assessment of the extreme thresholds of each metric, the intermediate thresholds for the metric and the thresholds for all other metrics used in calculating the overall B-IBI were held constant at the levels established by Weisberg *et al.* (1997) for mesohaline and polyhaline habitats or presented in Table I for tidal freshwater and oligohaline habitats. The a priori criteria for establishing whether new thresholds for any given metric significantly improved the B-IBI were as follows: to be considered significant, a change in threshold(s) must produce at least a 5 per cent

improvement in overall correct classification by the resulting B-IBI, with no concomitant decreases in classification of reference or degraded sites that exceed 5 per cent.

2.3. Multivariate statistical evaluations of B-IBI scoring systems

Two parallel multivariate statistical approaches were employed to determine whether the metrics could statistically distinguish data from reference and degraded habitats. Canonical variate discriminant analysis (CANDISC procedure; SAS Institute Inc., Version 8, 1999) and Soft Independent Modeling of Class Analogies (SIMCA; see also Wold, 1976; Kvalheim *et al.*, 1983; Droge and van't Kloster, 1987a,b; Kvalheim and Karstang, 1987; Kvalheim, 1993) were employed to determine overall statistical separation of biological data sets that were classified a priori by abiotic characteristics. Separate analyses were conducted on two types of biological data: unit deviate standardized raw data for the metrics; and the B-IBI scores (i.e. the '1–3–5' coded data) for the metrics. Analyses were conducted on data sets from each of the seven habitats.

Discriminant analysis is a standard multivariate technique for determining whether a priori classes are statistically separated. Unfortunately, our experience has shown that this technique tends to be susceptible to Type I errors (i.e. a tendency to discriminate between classes that are not truly different) when environmental data sets are analyzed. SIMCA is a more recently developed technique based upon principal components analysis that does not appear to force differences between groups when there are none. It was assumed that confirmation of statistical separation by these two different multivariate tests would be more definitive than if either were used alone. The discriminant analysis and the SIMCA analysis produced overall multivariate tests of statistical differences between the classes for each habitat, as well as the percent correct classification for the multivariate models.

2.4. Metrics that provide the greatest discriminatory power

The discriminant analysis and SIMCA also allowed the determination of the relative discriminatory power of each metric. The standardized canonical coefficients from the discriminant analysis and the relative discriminatory power (i.e. ratio of residual variance between groups to residual variance within groups for each metric; see Wold, 1976) from the SIMCA were assessed to determine the metrics most responsible for the separation of the groups. These measures of relative discriminatory power were also employed in weighting schemes to attempt to optimize the B-IBI (see below).

2.5. Minimum and optimal sets of metrics

Alternative B-IBI systems based on certain combinations of metrics were assessed by comparing the percent correct classification of the results relative to the a priori classification of the samples from each habitat. The types of models assessed were as follows: *Single-metric Models*—B-IBI scores for each metric alone; *Abundance–Biomass Models*—the averages of B-IBI scores for total abundance and biomass; *Multi-metric Models* (i.e. the existing B-IBI described by Weisberg *et al.*, 1997 or described in Table I)—Mean B-IBI scores for all metrics; *Discriminant-weighted Models*—weighted B-IBI systems based upon standardized discriminant coefficients; and *SIMCA-weighted Models*—weighted B-IBI systems based upon the relative SIMCA discriminatory power for each metric. For the final two types of models, the B-IBI scores for the individual metrics were multiplied by weighting factors before they were summed to produce an overall B-IBI score for each sample. The weighting factors were calculated by taking the relative proportion of the standardized discriminant coefficients

for each metric to the sum of all coefficients (for the *Discriminant-weighted Models*) or the proportion of relative discriminatory power for each metric to the sum of all distances (for the *SIMCA-weighted Models*).

Each of these models was selected due to specific operational interests. The *Single-metric Models* were assessed to determine the classification effectiveness of B-IBIs produced for individual metrics. The *Single-metric Models* were of interest because they provided an independent determination (i.e. in addition to the results of the multivariate statistical assessments) of the 'key' metrics that best separated the reference and degraded sites in each habitat. Among these models, those involving abundance and biomass were of particular interest, since these metrics do not require time and expense of taxonomic identifications. *Abundance–Biomass Models* assessed the classification power of B-IBIs produced by the combination of abundance and biomass data. *Multi-metric Models* were the existing B-IBIs against which other models were compared. The final two types of weighted models involved objective schemes (based upon relative discriminatory power of the metrics) for weighting the relative contribution of the metrics in order to attempt to produce optimal B-IBIs.

The a priori criteria for establishing successful models were as follows: overall classifications that were no lower than 10 per cent less than the classification of the existing *Multi-metric Model*; and classifications of reference and degraded samples that were within 20 per cent of the performance of *Multi-metric Models* for both.

2.6. Confidence limits and probability limits for B-IBIs

Bootstrap simulations similar to those employed by Alden (1992) for indices related to the sediment quality triad were used to determine 95 per cent probability and 95 per cent confidence limits for B-IBIs calculated for single samples and for means of nine samples, respectively. These probability and confidence limits were calculated for the reference and degraded data sets from each habitat. Each simulation consisted of 10 000 runs. Means of nine samples are important to the Chesapeake Bay Benthic Monitoring Program because status assessments at certain collection sites traditionally have been based upon means of overall B-IBI scores that are calculated from sets of three samples taken each year for a three year period. A similar approach could be used for the means of any number of samples in order to create confidence limits for overall B-IBI scores for reference and degraded sites for any given collection regime.

2.7. Sensitivity of the B-IBI to change in metric values

The sensitivities of the multi-metric B-IBI scoring systems to major changes in the raw values of the individual metrics were evaluated. Bootstrap simulations were also employed to conduct the sensitivity analyses on individual metrics. Raw values for samples were selected randomly from the reference or the degraded sites data sets in groups of nine at a time (to parallel studies which focused upon means of nine samples; see above). As each metric was studied, the raw values of the selected samples for that one metric were modified by a weighting factor prior to the B-IBI scoring process. The weighting factors that were sequentially tested represented a series that ranged from 0.1 to 2.0 in increments of 0.1 (i.e. a series that produced values that were 10 per cent to 200 per cent of the original values). All other metrics for the samples were not weighted. B-IBI scores were then calculated for each sample and the mean B-IBI value for the nine samples was determined. Each metric/weighting factor combination was run through 10 000 simulations to produce an overall B-IBI grand median and 95 per cent confidence limits for each level of change.

Table III. Results of threshold assessments. The thresholds for 47 metric–habitat combinations were evaluated to determine effects of iterative changes on correct classification by B-IBI scores

Habitat	Metrics improved	Classification with existing thresholds			Increase in classification			Centile
		Ref. (%)	Deg. (%)	All (%)	Ref.	Deg.	All	
Tidal freshwater	None of 4	67	69	66	—	—	—	
Oligohaline	None of 6	78	71	73	—	—	—	
Low salinity mesohaline	Biomass (of 7)	83	74	78	12	−1	5	3
High salinity mesohaline sand	None of 6	79	93	82	—	—	—	
High salinity mesohaline mud	None of 8	76	96	88	—	—	—	
Polyhaline sand	None of 8	96	100	97	—	—	—	
Polyhaline mud	None of 8	88	97	95	—	—	—	

3. RESULTS

3.1. Evaluation of metric thresholds

The thresholds developed by Weisberg *et al.* (1997) and Scott *et al.* (Versar, Inc., unpublished report) were found to be acceptable for use in the remainder of the study. Table III presents the results of the threshold evaluations for 47 metric–habitat combinations.[‡] Only one metric–habitat combination (biomass in low salinity mesohaline habitats) met the criteria for improvement and it only represented an increase from 78 per cent to 83 per cent in correct classification. Changes in thresholds for other metrics–habitat combinations produced little (<2 per cent change) or no improvement in overall correct classification by the resulting B-IBIs.

3.2. Multivariate statistical evaluations of B-IBI scoring systems

The two multivariate tests indicated highly significant ($p < 0.0001$) differences between the a priori groups for all habitats tested. Strong statistical differences were detected between the samples from reference and degraded sites by both multivariate tests, regardless of whether the raw metric data or the metric B-IBI scores were analyzed. In addition, univariate *F*-tests indicated that all of the metrics from all habitats displayed highly significant differences ($p < 0.001$) in B-IBI scores between the two groups.

Figure 1 presents the correct classification performance achieved by the discriminant analysis and the SIMCA model. The classification efficiency increased with the salinity of the habitat, from approximately 75 per cent in the tidal freshwater habitat to nearly 100 per cent in the polyhaline habitats. Likewise, multivariate distance between the reference and degraded sites data sets increased dramatically in higher salinity habitats (Figure 2). The canonical correlation between the metrics and the discriminant models tended to increase with the salinity of the habitat (Figure 3). The canonical correlation coefficients ranged from 0.4 (certain metrics from the tidal freshwater and oligohaline habitats) to over 0.9 (polyhaline habitats).

[‡]Note that for habitats with sufficient data for both biomass-based and abundance-based pollution-indicative or pollution-sensitive metrics (low mesohaline, high mesohaline mud, polyhaline sand and polyhaline mud), both types of metrics were assessed separately.

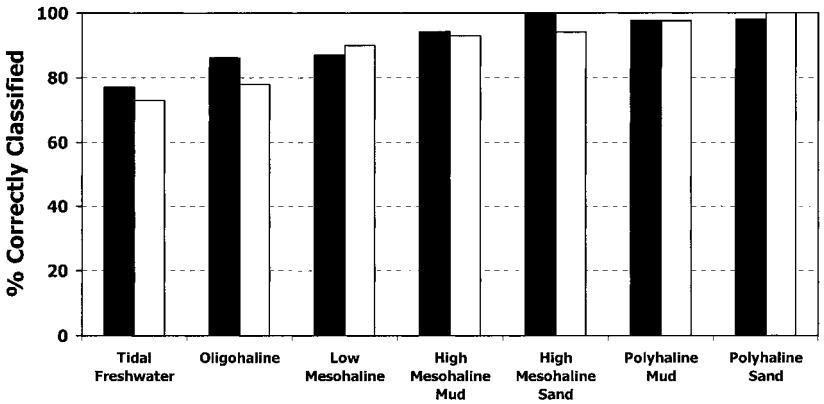

Figure 1. Percent correct classification of a priori groups by discriminant analysis (closed bars) and SIMCA (open bars)

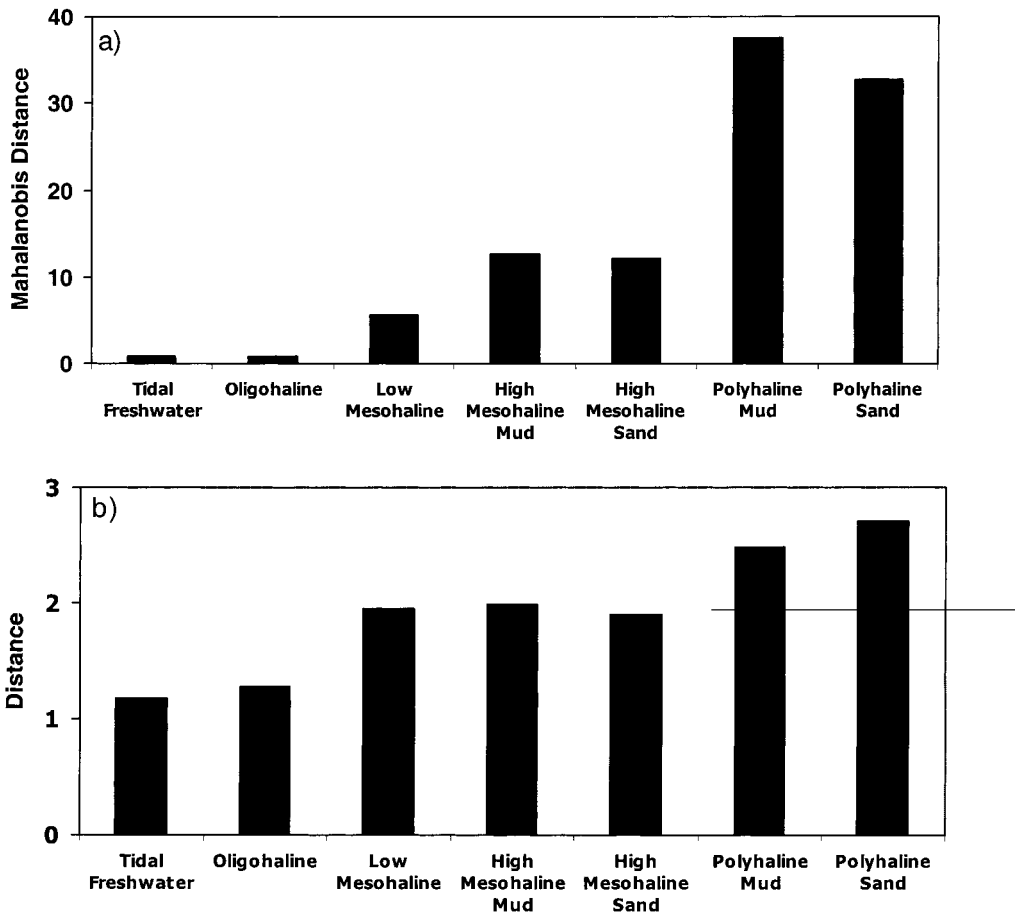

Figure 2. Multivariate distances between reference and degraded data sets: (a) Mahalanobis D^2 distances from discriminant analysis; and (b) multivariate distances (residual variance between to variance within) from SIMCA

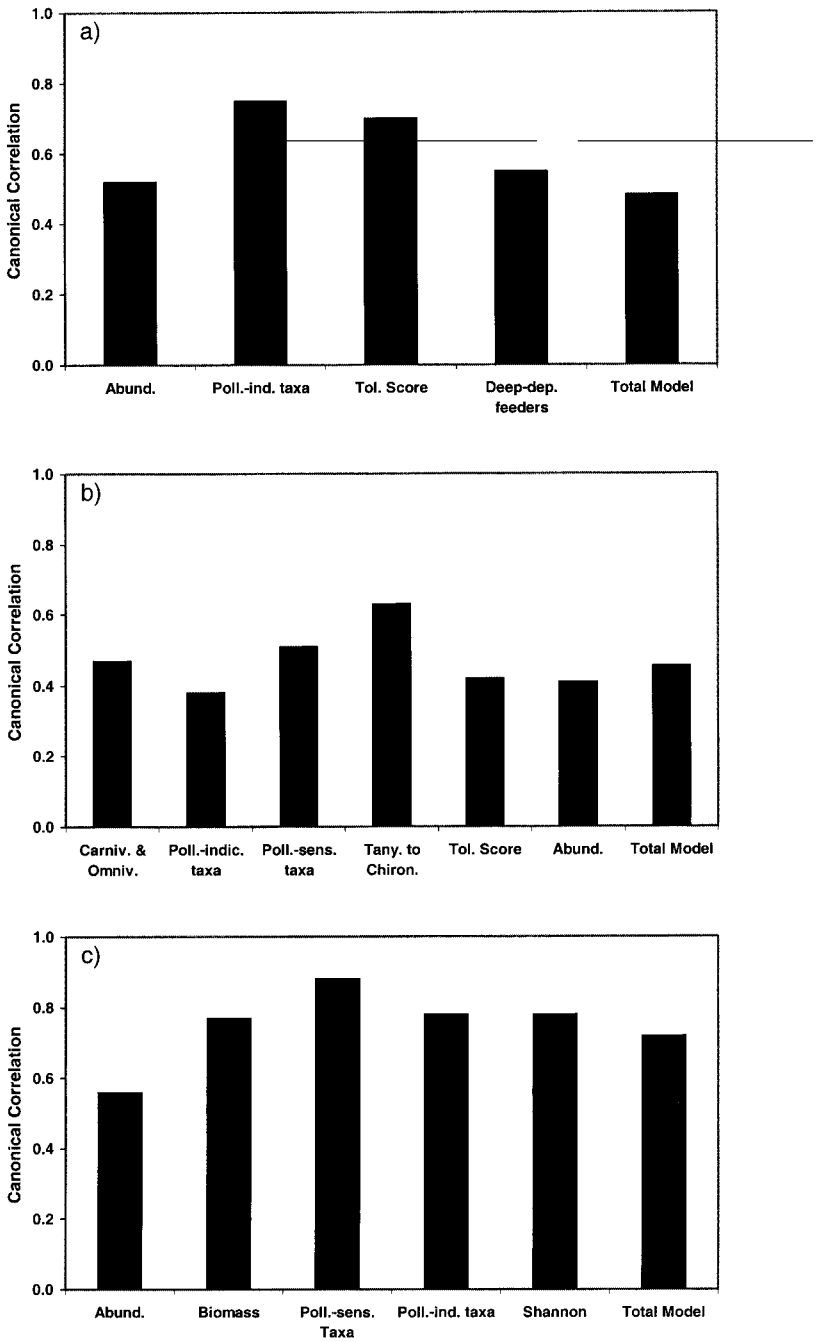

Figure 3. Canonical correlation coefficients for metrics from canonical discriminant analysis of each habitat: (a) tidal freshwater habitat; (b) oligohaline habitat; (c) low salinity mesohaline habitat; (d) high salinity mesohaline sand habitat; (e) high salinity mesohaline mud habitat; (f) polyhaline sand habitat; and (g) polyhaline mud habitat

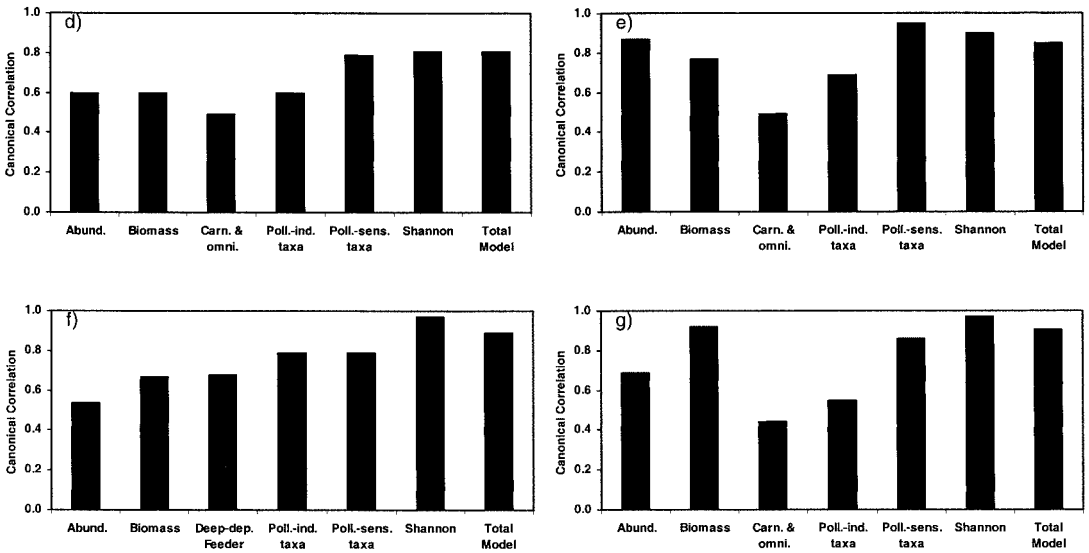

Figure 3. *Continued*

3.3. Metrics that provide the greatest discriminatory power

Certain combinations of metrics were found to be particularly important in the discriminatory power of the multivariate test. Figure 4 presents the relative discriminatory power of the individual metrics. The consensus of the results of the two different approaches were used to identify which metrics are most and least important to discriminating habitat quality. Table IV presents a summary of the consensus agreement between discriminant analysis and the SIMCA in determining the relative discriminatory power of individual metrics. Metrics based upon the relative abundance (or biomass) of pollution-indicative taxa and pollution-sensitive taxa (identified by Weisberg *et al.*, 1997), and diversity (i.e. Shannon's index) were the most important in both multivariate analyses.

3.4. Minimum and optimal sets of metrics

Some of the *Single-metric Models* classified the a priori groups nearly as well as the existing B-IBI *Multi-metric Models* (Table V). The metrics that performed most consistently well were pollution-indicative taxa, pollution-sensitive taxa, and diversity.

Figure 5 presents the classification performance of *Single-metric Models* for abundance and biomass, as well as the existing B-IBIs (*Multi-metric Models*) and the two types of weighted models. *Single-metric Models* for abundance and, especially, biomass performed quite well for the mesohaline and polyhaline habitats. However, the combination of abundance and biomass into a two-metric model did not improve, and sometimes degraded classification performance. The weighted models did not improve classification performance over the existing B-IBIs. For the high salinity mesohaline habitats, the overall classification of the weighted models was slightly less than that of *Multi-metric Models* (Figures 5d and 5e).

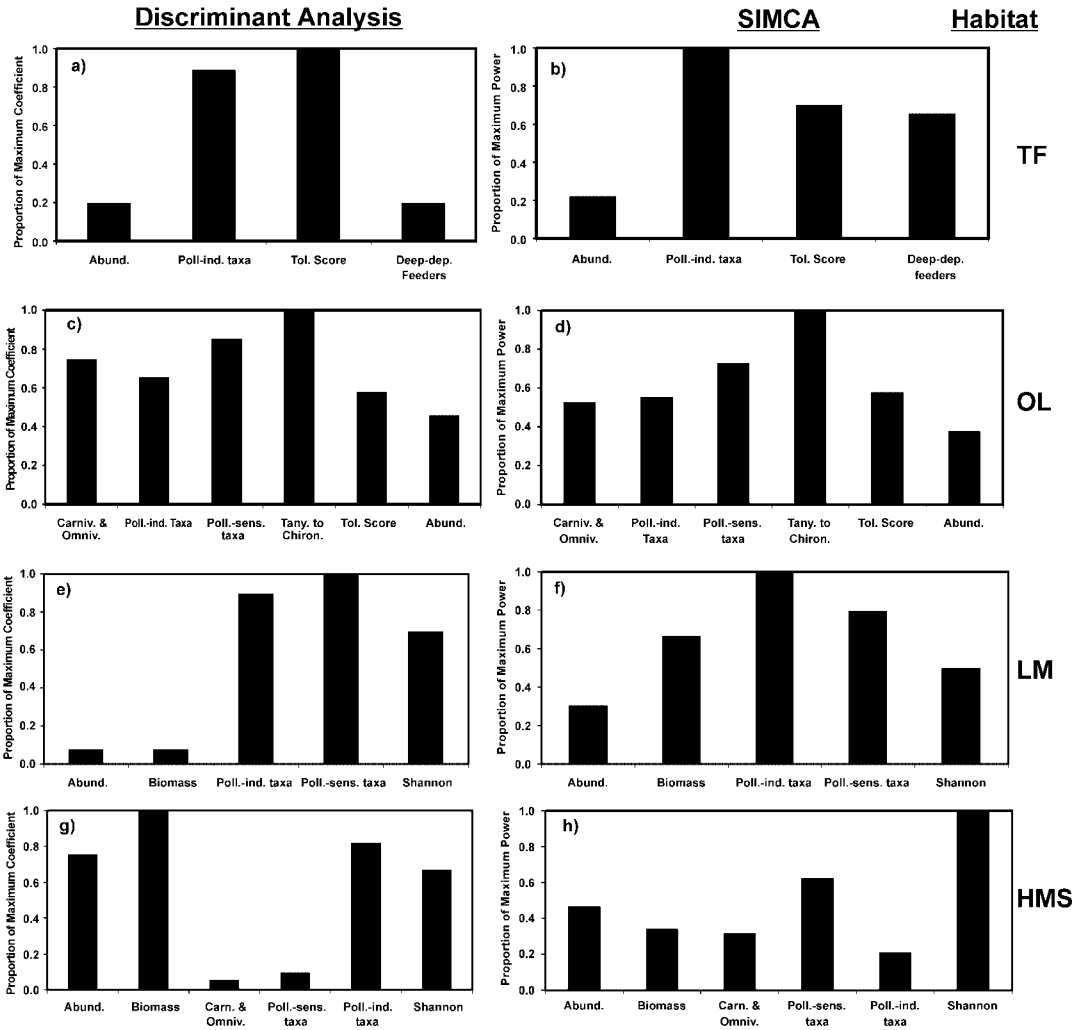

Figure 4. Relative discriminatory power of B-IBI metrics for various habitats: (a) standardized canonical coefficients from discriminant analysis of data from tidal freshwater habitats (TF); (b) standardized power values from SIMCA of data from TF; (c) standardized canonical coefficients for oligohaline habitats (OL); (d) standardized power values for OL; (e) standardized canonical coefficients for low salinity mesohaline habitats (LM); (f) standardized power values for LM; (g) standardized canonical coefficients for high salinity mesohaline sand habitats (HMS); (h) standardized power values for HMS; (i) standardized canonical coefficients for high salinity mesohaline mud habitats (HMM); (j) standardized power values for HMM; (k) standardized canonical coefficients for polyhaline sand habitats (PS); (l) standardized power values for PS; (m) standardized canonical coefficients for polyhaline mud habitats (PM); (n) standardized power values for PM

3.5. Confidence limits and probability limits for B-IBIs

The probability limits for single samples are presented in Table VI. Probability limits for single samples are generally quite large due to the variability of the data set. Thus, while the grand medians for the reference data sets were between 3.3 and 4.3 on the B-IBI scale, the extremes of the 95 per cent probability limits ranged from 2 to 5. The confidence limits for means of nine samples are presented in

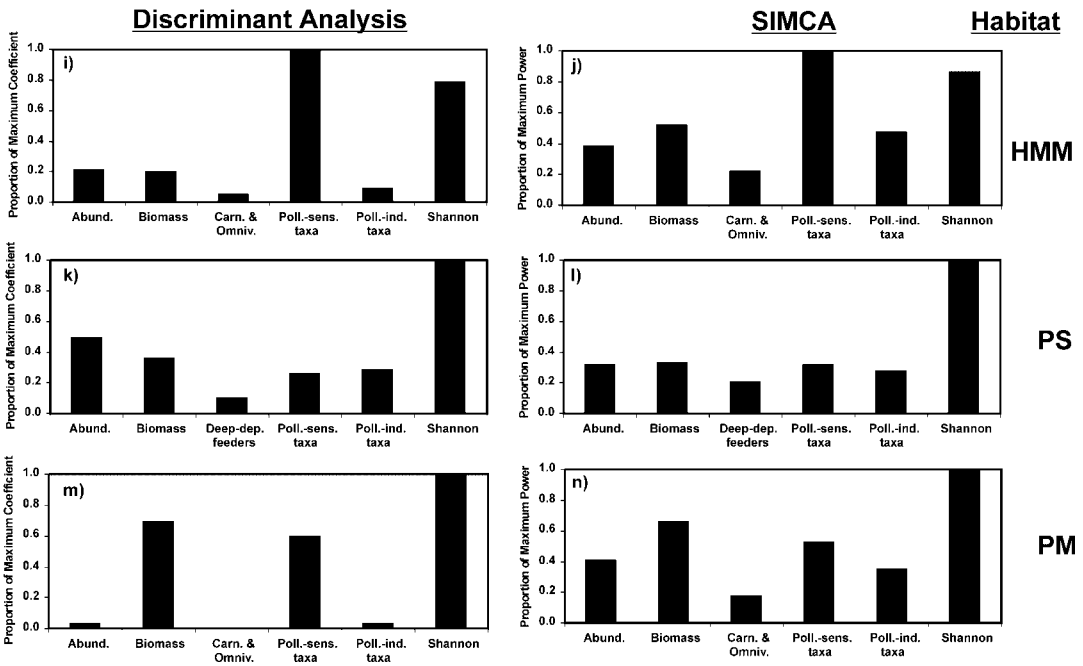

Figure 4. *Continued*

Table VII. As would be expected, the confidence limits for means are much 'tighter' than are probability limits for individual samples. For most of the habitats, the confidence limits of the reference and degraded sites did not overlap. Tidal freshwater and oligohaline habitats were the exceptions.

Table IV. Consensus results of relative discriminatory power of metrics Assessed by discriminant analysis and SIMCA

Habitat	Most important metrics	Least important metrics
Tidal freshwater	Pollution-indicative taxa Deep-dwelling deposit feeders	No consensus
Oligohaline	Tanypodinae to Chrinomidae ratio Pollution-sensitive taxa	Abundance
Low salinity mesohaline	Pollution-sensitive taxa Pollution-indicative taxa diversity	Abundance
High salinity mesohaline sand	diversity	Carnivores and omnivores
High salinity mesohaline mud	Pollution-sensitive taxa diversity	Carnivores and omnivores Pollution-indicative taxa
Polyhaline sand	diversity	Deep-deposit feeders
Polyhaline mud	diversity biomass Pollution-sensitive taxa	Carnivores and omnivores Pollution-indicative taxa

Table V. Correct classification of single metric B-IBI models. Only models with overall classifications within 10% of the performance of the existing multi-metric models, as well as within 20% of the performance for both reference and degraded samples are shown

		Correct classification (%)		
Habitat	Metric	Reference	Degraded	Overall
Tidal freshwater	Abundance of pollution-indicative taxa (%)	65	63	64
Oligohaline	Abundance of pollution-sensitive taxa (%)	64	67	66
	Abundance of pollution-indicative taxa (%)	66	68	68
	Tanypodinae to chironomidae abundance ratio	68	66	66
Low salinity mesohaline	Abundance of pollution-indicative taxa (%)	95	69	80
	Abundance of pollution-sensitive taxa (%)	89	68	75
High salinity mesohaline sand	diversity	76	80	77
High salinity mesohaline mud	diversity	75	88	83
	Biomass of pollution-indicative taxa (%)	87	90	88
	Biomass of pollution-sensitive taxa (%)	78	97	85
	Abundance of pollution-indicative taxa (%)	87	82	83
	Abundance of pollution-sensitive taxa (%)	87	93	91
Polyhaline sand	diversity	97	100	98
	abundance	85	82	85
	biomass	84	100	85
	Abundance of pollution-indicative taxa (%)	100	91	98
	Abundance of pollution-sensitive taxa (%)	97	100	97
	Abundance of deep-deposit feeders (%)	91	91	91
Polyhaline mud	diversity	89	94	93
	abundance	90	88	88
	biomass	88	93	91
	Abundance of pollution-sensitive taxa (%)	92	91	91

3.6. Sensitivity of the B-IBI to change in metric values

The overall B-IBI scores for all habitats were quite stable even when individual metrics were subjected to major changes. Figure 6 presents the results from the low salinity mesohaline habitat reference data set as a visual example of the findings of the sensitivity analyses. While the grand medians of some metrics moved towards the 95 per cent confidence limits as the degree of changes became more extreme, only a very few of the overall B-IBI scores became significantly different (i.e. the grand medians cross the 95 per cent confidence limits for the unaltered B-IBIs for the reference sites or degraded sites data sets for the habitat). In fact, of the 47 metric–habitat combinations tested for sensitivity, only 3.3 per cent (3 out of 92 tests; 3 from metric–habitat–reference site combinations and none from metric–habitat–degraded site combinations) displayed a significant change in the overall B-IBI. Since the small percentage of significant differences was not greater than what one could expect from chance alone, no further presentation or interpretation of the findings was deemed necessary.

4. DISCUSSION

4.1. Verification of the B-IBI

The threshold assessments demonstrated that the metric thresholds developed by Weisberg *et al.* (1997) worked as well as a range of potential new thresholds established from the single sample

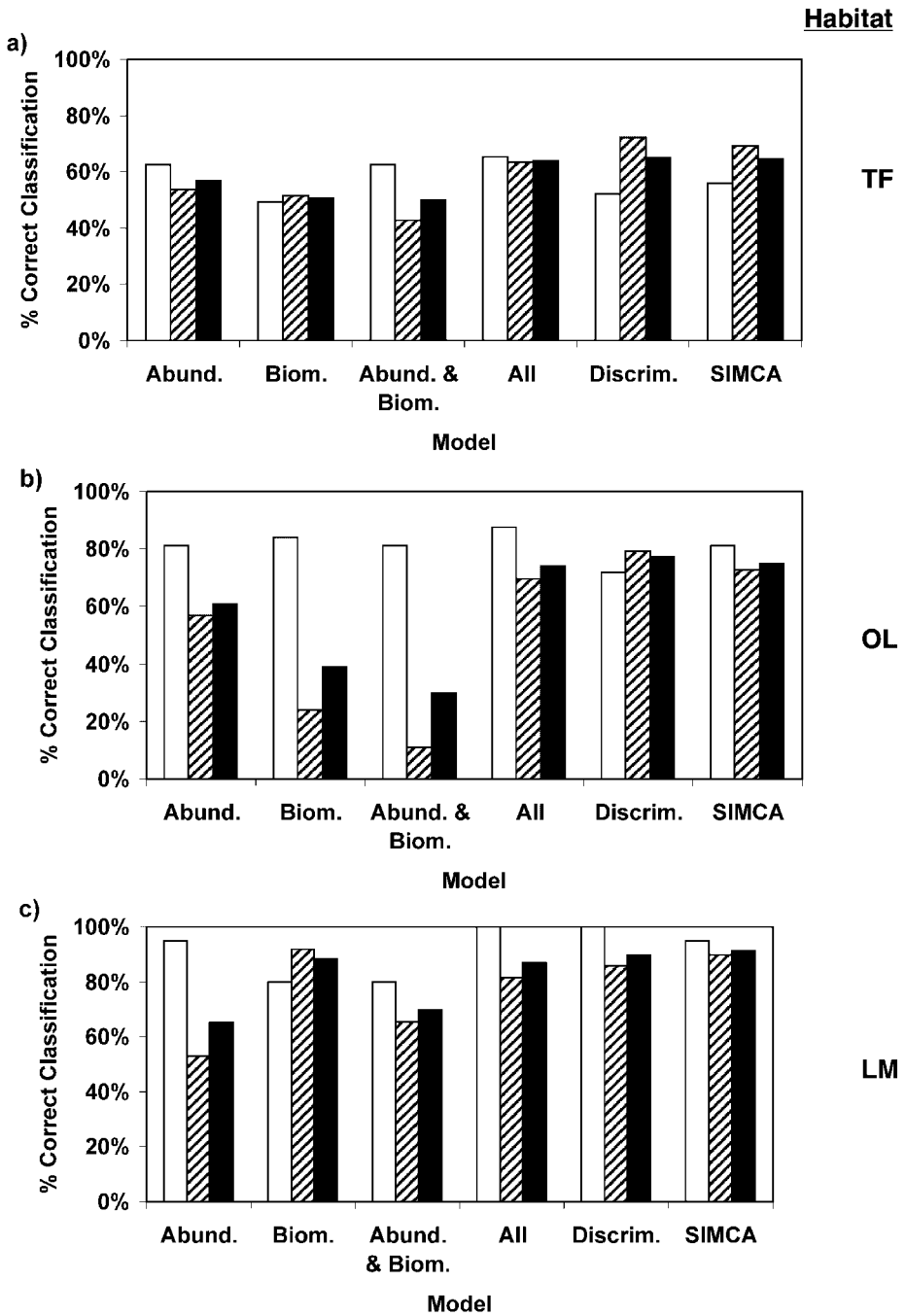

Figure 5. Percent correct classifications for models in various habitats (same abbreviations as for Figure 4): (a) TF; (b) OL; (c) LM; (d) HMS; (e) HMM; (f) PS and (g) PM. The open bars indicate the correct classification of the reference samples, the cross-hatched bars indicate the classification of degraded samples, and the closed bars indicate the overall classification

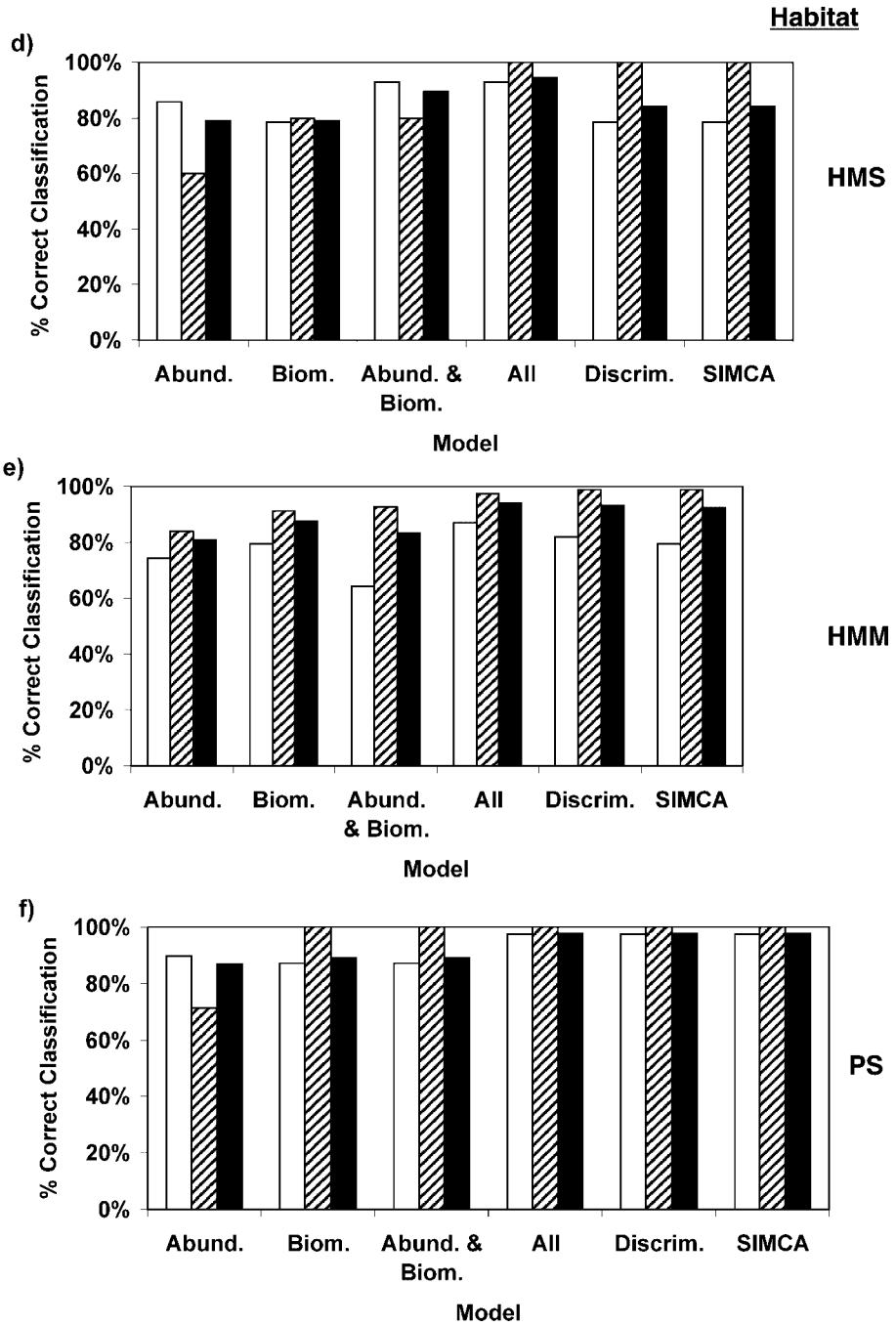

Figure 5. *Continued*

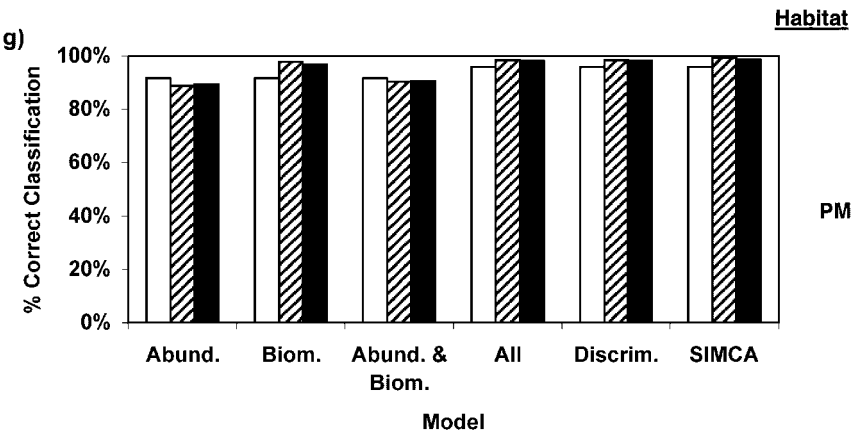

Figure 5. *Continued*

reference data sets (i.e. from the lowest value to the first quartile for the low thresholds; and from the highest value down to the third quartile for the high thresholds). Thus, the published methods, metrics and thresholds work well and were employed throughout our study.

The two multivariate tests indicated very highly significant differences between the a priori groups for all habitats tested. Thus, the B-IBI scores calculated for the various sets of metrics provide excellent overall discrimination capacity. The degree of discrimination between reference and degraded groups appeared to increase with increasing salinity: the tests for oligohaline and tidal freshwater habitats displayed the lowest level of discrimination, those of mesohaline habitats were intermediate, and analyses of polyhaline habitats displayed the strongest discrimination and the strongest, most effective models (Figures 1–3). Nonetheless, both of the multivariate tests indicated that the B-IBI scores for all habitats significantly separated the reference and degraded conditions. Thus, while the B-IBIs for low salinity systems were somewhat weaker, all of the B-IBI systems have been statistically verified by two complementary multivariate tests.

Table VI. Result of bootstrap simulations: B-IBI grand medians and probability limits for single samples

Habitat	*A priori* condition	Lower probability limit	Grand median	Upper probability limit
Tidal freshwater	Reference	2.40	3.75	5.00
	Degraded	1.00	2.50	4.00
Oligohaline	Reference	2.33	3.60	4.67
	Degraded	1.67	2.33	4.00
Low salinity mesohaline	Reference	3.00	3.80	4.60
	Degraded	1.00	1.80	4.00
High salinity mesohaline sand	Reference	2.33	3.33	4.00
	Degraded	1.33	1.67	2.67
High salinity mesohaline mud	Reference	2.00	3.67	4.67
	Degraded	1.00	1.33	2.33
Polyhaline sand	Reference	3.33	4.33	4.67
	Degraded	1.67	1.67	2.00
Polyhaline mud	Reference	3.00	4.00	5.00
	Degraded	1.00	1.00	2.60

Table VII. Result of bootstrap simulations: B-IBI grand medians and confidence limits for means of 9 samples

Habitat	*A priori* condition	Lower confidence limit	Grand median	Upper confidence limit
Tidal freshwater	Reference	3.50	4.50	5.00
	Degraded	2.00	3.00	4.00
Oligohaline	Reference	3.00	3.80	4.60
	Degraded	2.20	2.60	3.40
Low salinity mesohaline	Reference	3.00	3.80	4.60
	Degraded	1.00	1.40	2.20
High salinity mesohaline sand	Reference	3.33	4.00	4.67
	Degraded	1.33	2.33	3.00
High salinity mesohaline mud	Reference	3.00	3.67	4.33
	Degraded	1.33	1.67	2.00
Polyhaline sand	Reference	4.00	4.33	5.00
	Degraded	1.67	1.67	2.00
Polyhaline mud	Reference	3.33	3.67	4.33
	Degraded	1.67	1.67	2.33

The B-IBIs produced classification efficiencies for samples in the mesohaline and polyhaline data sets that were 4–8 per cent lower than those reported by Weisberg *et al.* (1997). These results are not surprising because the single sample data sets have more spatiotemporal variability than do the data sets of site means. Nonetheless, the classification performance observed for the B-IBI scoring systems applied to the single sample data sets remained quite high (78–96 per cent correct overall classification) for the mesohaline and polyhaline habitats. The B-IBIs for lower salinity habitats produced lower efficiencies, probably due to the greater variability in these systems (see below).

4.2. Key metrics

The two multivariate approaches were also used to evaluate the discriminatory power of the individual metrics (Figure 4). Since all of the metrics are statistically different between the groups and reflect various degrees of positive correlation with the multivariate functions that separated them, these assessments are designed to explore the relative discriminatory power of certain metrics that may be key to the separations. While there were some differences in the metrics shown to be most important by each of the multivariate tests, there is consensus between the two analyses (Table IV). Pollution-indicative taxa, pollution-sensitive taxa, and diversity (Shannon's index) appear to be the most important of the metrics.

It is not too surprising that these same metrics also performed best in the *Single-metric Models* for the mesohaline and polyhaline habitats. As reported by Weisberg *et al.* (1997), the correct classifications of single metric B-IBIs were not quite as high as observed for the *Multi-metric Models*. However, the classification performance of many of these models was still quite high, with many of the models involving these 'key' metrics producing correct classifications of sites exceeding 90 per cent.

4.3. Minimum and optimal sets of metrics

The relatively high classification performances of *Single-metric Models* serve to confirm which key metrics appear to be most important in each habitat. However, since all metrics other than total

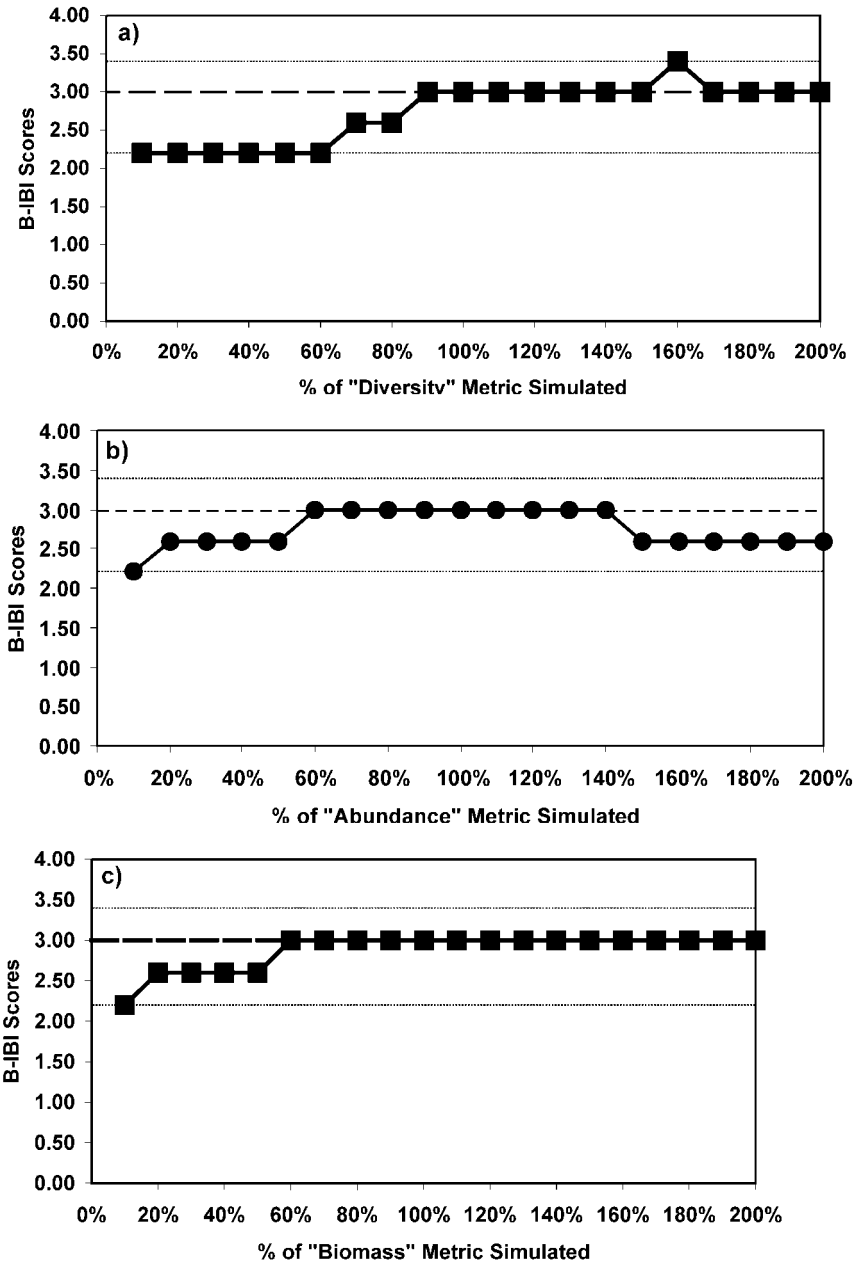

Figure 6. Results of sensitivity analyses of low mesohaline habitat metrics from the reference data set. The data points are the medians of overall B-IBI values from bootstrap simulations (10 000 runs per data point) in which the metrics were varied individually from 10 to 200 per cent of their original values (with all other metrics drawn from their original distributions) prior to calculating the B-IBI scores. The dashed line in each graph is the grand median for the overall B-IBI score produced by 10 000 simulations of the multi-metric models without any introduced change, while the dotted lines represent the 95 per cent confidence limits for these unchanged models. The metrics are: (a) diversity (Shannon's index); (b) abundance; (c) biomass; (d) pollution-indicative taxa (percentage of total abundance); and (e) pollution-sensitive taxa (percentage of total abundance)

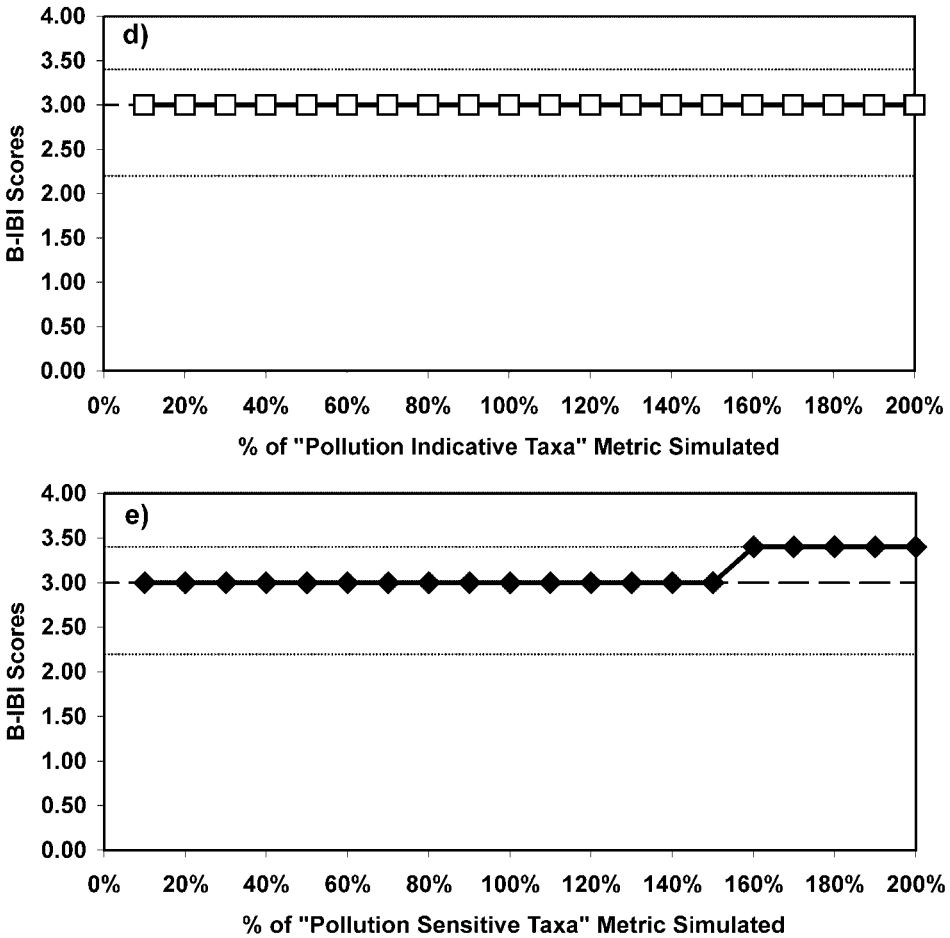

Figure 6. *Continued*

abundance and biomass require a quantitative taxonomic analysis of each sample, this sort of single metric model assessment is more of an academic than of an operational interest. That is, to calculate any one of these metrics would involve nearly the same resources that would be required to produce all, so it would make little sense not to use all metrics, especially since *Multi-metric Models* perform the best for all habitats. On the other hand, data for the metrics abundance and biomass can be generated with lower resource demands. Abundance data can be generated without taxonomic identifications or biomass determinations, while biomass data can be generated without taxonomic identifications or counts. Thus, there was an interest in whether these metrics perform well in single metric B-IBI models.

Neither abundance nor biomass performs satisfactorily as single metric B-IBIs for the lower salinity habitats (tidal freshwater or oligohaline; see Figure 5). On the other hand, abundance and, particularly, biomass models performed quite well in the mesohaline and polyhaline habitats. Classification by biomass models generally exceeded 80 per cent for these habitats.

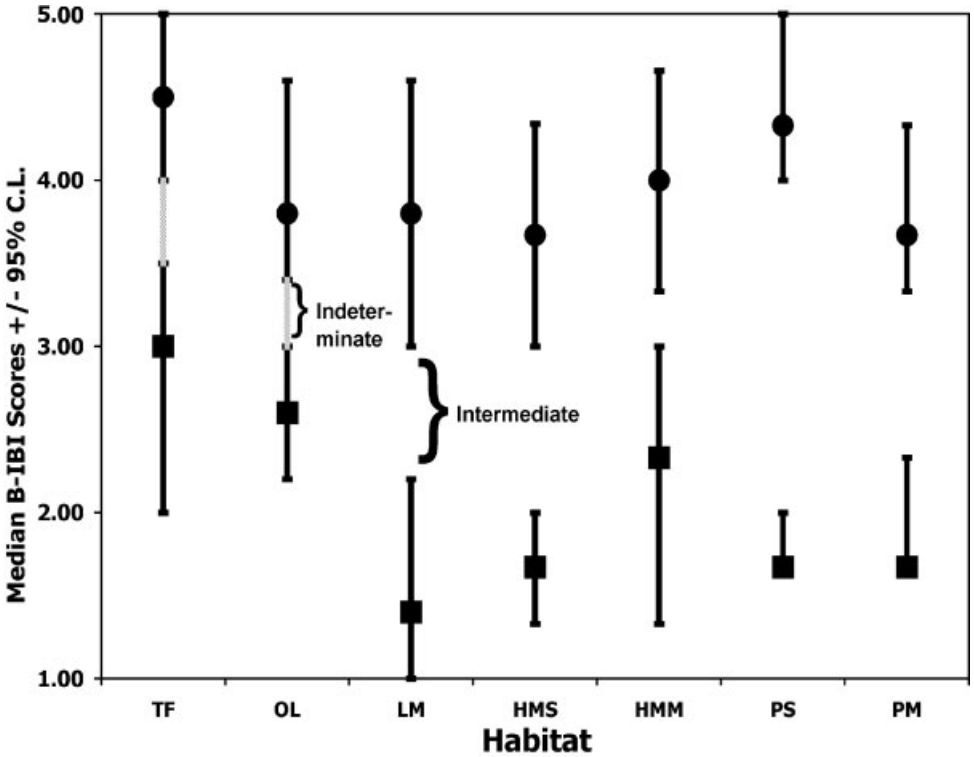

Figure 7. Grand medians and 95 per cent confidence limits for overall B-IBI scores for means of nine samples for a priori reference samples (circles) and for a priori degraded site samples (squares). The areas of overlap of confidence limits for reference and degraded data sets for tidal freshwater and oligohaline habits are shown as dotted regions and are called 'indeterminate'. Areas of separation between reference and degraded confidence limits are called 'intermediate'

Models involving both abundance and biomass (*Abundance–Biomass Models*) generally performed no better than, and often worse than, the *Single-metric Models* for either metric alone (Figure 5). The only exception was for the high salinity mesohaline sand habitat, where the percentage of correct classifications of *Abundance–Biomass Models* improved by 10 per cent compared to the performance of *Single-metric Models* for abundance or biomass. On the other hand, the combined *Abundance–Biomass Models* did quite poorly (more so than either of the metrics alone) in identifying reference sites from the high mesohaline mud habitat. Overall, *Abundance–Biomass Models* performed 6–44 per cent lower in correct classifications than the *Multi-metric Models*. These findings suggest that including both abundance and biomass in the same index is counterproductive.

The stability of a multi-metric 'weight of evidence' approach was one of the desired design features identified by Weisberg *et al.* (1997) when the Chesapeake Bay B-IBI was developed. The most striking finding from the metric sensitivity assessments is that large changes in the values of a single metric do not significantly affect the overall B-IBI scores for either the reference or the degraded sites. The median B-IBI scores produced for the range of changes very seldom deviated beyond the 95 per cent confidence limits for the existing, unchanged *Multi-metric Models*. This finding, plus the observation that significant changes in the thresholds of individual metrics do little to affect the classification

performance of the overall B-IBI, both indicate that the multi-metric B-IBIs are quite robust and stable. That is, when a single metric deviates from the 'weight of evidence' provided by the remaining metrics, it does not greatly influence the overall B-IBI scores or the classification of the samples. Thus, while some single metric B-IBI systems (e.g. biomass alone) may perform quite well for the higher salinity habitats, they do not have the stability and robustness inherent in the *Multi-metric Models*. Furthermore, a multi-metric B-IBI may have greater potential sensitivity and resolution to distinguish differences in habitat quality when conditions are between the extreme reference and degraded data types employed in the present verification study.

The models that were designed to assess the effects of weighting schemes on performance of the B-IBI system did not increase the percentage correct classification by more than a few percentage points and, at least in the high mesohaline sand habitat, did somewhat more poorly than did the existing, unweighted *Multi-metric Models*. Of course, the classification performances of the *Multi-metric Models* were already quite high for many of the data sets being assessed, so great improvements would not be expected or necessary.

It should be noted, however, that weighting schemes may be useful for enhancing performance for data sets from other estuarine systems or for aiding in the discrimination of sites that are not so much at the extremes of the 'non-stressed' to 'stressed' spectrum as those used in the present verification study. In other words, weighting schemes could prove helpful in distinguishing the relative quality of sites that are in the intermediate categories (i.e. in distinguishing between the various 'shades of gray' in the gradient of anthropogenic stress). Until this conjecture can be substantiated with other data sets containing a priori classifications of intermediate habitat quality, there is no compelling reason to weight the B-IBI scores for various metrics.

4.4. Confidence limits and probability limits for B-IBIs

The B-IBI data for the lower salinity habitats were the most variable, producing probability limits for reference and degraded samples that had degrees of overlap ranging from quite high in the tidal freshwater to moderately low in the high salinity mesohaline habitats (i.e. high mesohaline sand and high mesohaline mud). Thus, if one was to use these probability limits to classify the scores of individual samples from these habitats, there could be three possible outcomes: 'reference' for samples with overall B-IBI scores falling above the lower probability limits of the reference condition but above the upper probability limit of the degraded sites data set; 'degraded' for samples with scores that fall below the upper probability limits of the degraded sites data set and the lower probability limit of the reference data set; and 'indeterminate' for samples with scores that fall within the range of overlap of the probability limits of the two data sets. Obviously, for the tidal freshwater and oligohaline habitats, classifications of individual samples will tend to produce a high percentage of samples that would fall into the indeterminate category.

The probability limits for the reference and degraded sites data sets from the polyhaline habitats did not overlap. The classification of samples from polyhaline habitats employing these probability limits would also produce three possible outcomes: 'reference' for samples with overall B-IBI scores falling above the lower probability limits of the reference condition; 'degraded' for samples with scores that fall below the upper probability limits of the degraded sites data set; and 'intermediate' for samples with scores that fall between below the lower probability limit of the reference data set but above the upper probability limit of the degraded sites data set. In this case, the third class is not defined by uncertainty, but by biological characteristics that classify the sample as neither 'non-stressed' nor 'stressed', but statistically intermediate in quality.

The confidence limits for means of samples are much 'tighter' than are probability limits for individual samples. Therefore, an analysis strategy that employs the means of multiple samples reduces the uncertainty of the classification of regions as to sediment quality or benthic health. There were fewer overlaps in the confidence limits for reference and degraded data sets when means of nine samples were analyzed. Tidal freshwater and oligohaline were the only habitats for which there were overlaps in confidence limits between the reference and degraded conditions (Figure 7). Thus, if B-IBI confidence limits were used for classification purposes, all of the other habitats would have the same three potential outcomes as discussed previously for the polyhaline single sample classification using probability limits: 'reference', 'degraded' or 'intermediate'. Tidal freshwater and oligohaline classifications would yield 'reference', 'degraded' or 'indeterminate' results.

While there may be certain advantages to binomial systems for which a sample (or mean of samples) is classified as either 'non-stressed' (similar to reference) or 'stressed' (degraded), the capacity to identify samples that are intermediate in quality improves the resolution of the interpretation of the relative condition of a site and probably better reflects reality. Confidence limits and, to a lesser degree, probability limits could provide important tools to the understanding of environmental quality.

A cautionary note should be made concerning the use of these limits for classification of relative quality of samples, sites or strata (i.e. regions of similar habitat type). The limits were created for data sets from the Chesapeake Bay that were classified a priori as to sediment quality and that contained the full sets of metrics used in the B-IBI system described in the Weisberg *et al.* (1997) investigation. These limits may not be valid for other estuarine ecosystems or for data sets with reduced numbers of metrics. Therefore, they should be used with caution if they are being applied to data sets from other estuaries and/or from investigations designed to employ fewer metrics. Certainly, the best approach is to apply the simulation process of establishing confidence limits for reference and degraded conditions from each new ecosystem.

4.5. B-IBI effectiveness in low salinity habitats

A trend observed throughout all of the statistical and simulation assessments was that the performance of the B-IBIs improved with the salinity of the habitat. The multivariate separation of the reference and degraded site data sets increased from tidal freshwater to polyhaline. The uncertainty associated with B-IBIs from tidal freshwater and oligohaline, as measured by the overlap of the confidence limits of reference and degraded sites, were greater than that of the mesohaline and, particularly, polyhaline habitats. It is possible that the natural stresses and variability of many tidal freshwater habitats (e.g. the hydrodynamic stresses in rapidly flowing, narrow reaches of rivers; high sedimentation rates of the 'turbidity maximum zone'; unstable, fluidized muds, etc.) may make the B-IBI scoring system more uncertain and less reliable than those established for more saline habitats. If some of the data used for the reference distributions were from stations that were naturally stressed, the resulting scoring system could be variable and the classification performance for B-IBIs calculated from what were believed to be reference data sets could be poor. Thus, while the B-IBIs for the tidal freshwater and oligohaline systems produced marginal to fair classification results, better results may be achieved in the future if the confounding effects of natural stress could be reduced (i.e. if best professional judgement could be used to eliminate reference sites that are believed to be naturally stressed from the data set prior to the development of metric thresholds).

A slightly different hypothesis for the relationship between the relative effectiveness of the B-IBIs and salinity could be due to natural ecotones related to stresses found in the low salinity habitats of tributaries. In other words, rather than the conjecture that the B-IBI development may have been

confounded by some subset of the reference sites having been naturally stressed, it could be that most or all of the benthic communities in these habitats are naturally stressed to a degree not observed for other estuarine habitats. Between the natural stresses listed above, osmotic stresses, and the tendency for organisms (and their reproductive products) to be flushed downstream, it may be difficult for healthy, non-stressed benthic communities to become established in these habitats. Thus, even communities from habitats not impacted by anthropogenic stresses may not be that greatly different in biological characteristics from anthropogenically stressed communities. In going downstream toward more stable higher salinity communities or upstream above the fall-line to true freshwater habitats, one would hypothesize that natural stresses would tend to decrease. This gradient of decreasing stress going upstream and downstream from the low salinity habitats parallels the relative effectiveness of the B-IBIs: many previous studies have established effective B-IBIs for freshwater systems (e.g. Clements *et al.*, 1992; Lenat, 1993; Kerans and Karr, 1994; Lang and Reymond, 1995); and the present study has demonstrated that the more saline habitats have stronger B-IBIs with tighter confidence limits. Multivariate distances between benthic biological data sets from reference and degraded sites are also much less for the low salinity habitats than for the more saline systems. Thus, the low salinity reaches of tributaries may be stressed, highly variable habitats that form ecotones of benthic communities that are more difficult to classify by B-IBIs as to the degree of anthropogenic stress than are the more stable habitats upstream or downstream.

5. CONCLUSIONS

The B-IBI scoring systems employed in the Chesapeake Bay Benthic Monitoring Program have been verified as being sensitive, stable, robust and statistically sound. The B-IBI metrics and thresholds developed by Weisberg *et al.* (1997) and Scott *et al.* (Versar, Inc., unpublished report; see also Table I) are valid for application to both site means and single samples. The published thresholds performed as well in classifying stations as any of a range of values that were studied. Performance of B-IBIs, as measured by correct station classification and statistical discriminatory power, increased with the salinity of the habitats, ranging from marginal in the tidal freshwater habitat to excellent in the polyhaline habitats. While single metric B-IBI systems (e.g. biomass alone) often displayed good classification performance, particularly for the higher salinity habitats, the multi-metric systems displayed the desirable characteristics of being quite stable and robust. The B-IBIs including all metrics generally gave a higher per cent correct classification than those using fewer metrics, and weighting of metrics resulted, at best, in only small improvements. Thus, the existing B-IBI systems are recommended for use in benthic assessments in the Chesapeake Bay. Assessments of the environmental quality of various habitats in the Bay can also employ bootstrapped probability and confidence limits of B-IBIs to distinguish between classes of benthic health. The statistical and simulation approaches employed in this study could be used to verify benthic indices in other ecosystems.

ACKNOWLEDGEMENTS

This study was partially funded by Maryland Department of Natural Resources under Cooperative Agreement No. CA-99-03 07-4-30565-3734.

REFERENCES

Alden RW, III. 1992. Uncertainty and sediment quality assessments: confidence limits for the triad. *Environ. Toxicol. Chem.* **11**: 645–651.

Clements WH, Cherry DS, Van Hassel JH. 1992. Assessment of the impact of heavy metals on benthic communities at the Clinch River (Virginia): evaluation of an index of community sensitivity. *Canadian J. of Fisheries and Aquatic Sci.* **49**: 1686–1694.
Dauer DM, Alden RW, III. 1995. Long-term trends in macrobenthos of the lower Chesapeake Bay. *Marine Pollution Bulletin* **30**(12): 840–850.
Diaz RJ. 1989. Pollution and tidal benthic communities of the James River Estuary, Virginia. *Hydrobiologia* **180**: 195–211.
Droge JBM, van't Kloster HA. 1987a. An evaluation of SIMCA. Part 1—The reliability of the SIMCA pattern recognition method for a varying number of objects and features. *J. of Chemometrics* **1**: 221–230.
Droge JBM, van't Kloster HA. 1987b. An evaluation of SIMCA. Part 2—Classification of pyrolysis mass spectra of Pseudomonas and Serratia bacteria by pattern recognition using the SIMCA classifier. *J. of Chemometrics* **1**: 231–241.
Engle VD, Summers JK. 1999. Refinement, validation, and application of a benthic condition index for Northern Gulf of Mexico estuaries. *Estuaries* **22**(3A): 624–635.
Engle VD, Summers JK, Gaston GR. 1994. A benthic index of environmental condition of Gulf of Mexico. *Estuaries* **17**(2): 372–384.
Hyland JL, Snoots TR, Balthis WL. 1998. Sediment quality of estuaries in the southeastern U.S. *Environ. Monit. and Assess.* **51**: 331–343.
Hyland JL, Van Dolah RF, Snoots TR. 1999. Predicting stress in benthic communities of Southeastern U.S. estuaries in relation to chemical contamination of sediments. *Environ. Toxicol. and Chem.* **18**: 2557–2564.
Kerans BL, Karr JR. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* **4**: 768–785.
Kvalheim OM. 1993. *Sirius* for Windows™ Data into Information. Technical Manual for *Sirius* software. Pattern Recognition Systems A/S.
Kvalheim OM, Karstang TV. 1987. A general-purpose program for multivariate data analysis. *Chemometrics and Int. Lab. Systems* **2**: 235–237.
Kvalheim OM, Oygard K, Grahl-Nielsen O. 1983. SIMCA multivariate data analysis of blue mussel components in environmental pollution studies. *Anal. Chim. Acta* **150**: 145–157.
Lang CG, Reymond O. 1995. An improved index of environmental quality for Swiss rivers based upon benthic invertebrates. *Aquat. Sci.* **57**: 172–180.
Lenat DR. 1993. A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assigning water-quality ratings. *J. N. American Benthological Society* **12**: 279–290.
Paul JF, Scott KJ, Holland AF, Weisberg SB, Summers JK, Robertson A. 1992. The estuarine component of the U.S. EPA's Environmental Monitoring and Assessment Program. *Chem. and Ecol*. **7**: 93–226.
Rakocinski CF, Brown SS, Gaston GR, Heard RW, Walker WW, Summers JK. 1997. Macrobenthic responses to natural and contaminant-related gradients in northern Gulf of Mexico estuaries. *Ecol. Appl.* **7**: 1278–1298.
Ranasinghe JA, Weisberg SB, Dauer DM, Schaffner LC, Diaz RJ, Frithsen JB. 1994. Chesapeake Bay Benthic Community Restoration Goals. Prepared for the U.S. EPA Chesapeake Bay Program Office, the Governor's Council on Chesapeake Bay Research Fund, and the Maryland Department of Natural Resources by Versar, Inc., Columbia, MD.
Reynoldson RH, Resh VH, Day KE, Rosenberg DM. 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *J. N. Am. Benthol. Soc.* **16**(4): 833–852.
Van Dolah RF, Hyland JL, Holland AF, Rosen JS, Snoots TR. 1999. A benthic index of biological integrity for assessing habitat in estuaries of the southeastern USA. *Mar. Environ. Res.* **48**: 1–15.
Weisberg SB, Ranasinghe JA, Dauer DM, Schaffner LC, Diaz RJ, Frithsen JB. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. *Estuaries* **20**: 149–158.
Wold S. 1976. Pattern recognition by means of disjoint principal components models. *Pattern Recognition* **8**: 127–139.

APPENDIX

Descriptions of most of the metrics for tidal freshwater and oligohaline B-IBIs are presented in Ranasinghe *et al.* (1994). The following tables present taxa comprising the 'pollution-indicative taxa' and 'pollution-sensitive taxa' metrics for the tidal freshwater and oligohaline habitats.

Table A1. Tidal freshwater pollution-indicative taxa

Oligochaeta:	*Limnodrilus hoffmeisteri*
	Tubificidae immature without capilliform chaetae

Table A2. Oligohaline pollution-indicative taxa

Polychaeta:	*Streblospio benedicti*	Oligochaeta:	*Aulodrilus limnobius*
	Heteromastus filiformis		*Aulodrilus paucichaeta*
	Leitoscoloplos spp.[a]		*Aulodrilus pigueti*
	Mediomastus ambiseta		*Aulodrilus pluriseta*
	Neanthes succinea		*Branchiura sowerbyi*
	Polydora cornuta		*Haber* cf. *speciosus*
			Ilyodrilus templetoni
Bivalvia:	*Corbicula fluminea*		*Isochaetides freyi*
			Limnodrilus cervix
Amphipoda:	*Leptocheirus plumulosus*		*Limnodrilus claparedianus*
			Limnodrilus hoffmeisteri
Chironomidae:	*Chironomus* spp.[a]		*Limnodrilus udekemianus*
	Cladotanytarsus spp.[a]		*Oligochaetes (unidentifiable)*[b]
	Coelotanypus spp.[a]		*Quistadrilus multisetosus*
	Glyptotendipes spp.[a]		*Telmatodrilus vejdoskyi*
	Polypedilum spp.[a]		Tubificidae immature without capilliform chaetae
	Procladius spp.[a]		Tubificidae with capilliform chaetae
	Tanypus spp.[a]		*Tubificoides* spp.[a]

[a]All species belonging to the genera are classified as pollution-indicative.
[b]Oligochaetes should be identified to the lowest possible taxonomic level. Whenever a specimen is unidentifiable it should be classified as pollution-indicative.

Table A3. Oligohaline pollution-sensitive taxa

Polychaeta:	*Marenzelleria viridis*
Isopoda:	*Chiridotea almyra*

