

# A Performance Comparison of Metric Scoring Methods for a Multimetric Index for Mid-Atlantic Highlands Streams

**KAREN A. BLOCKSOM**

Ecological Exposure Research Division  
National Exposure Research Laboratory  
U.S. Environmental Protection Agency  
Cincinnati, Ohio 45268, USA

**ABSTRACT** / When biological metrics are combined into a multimetric index for bioassessment purposes, individual metrics must be scored as unitless numbers to be combined into a single index value. Among different multimetric indices, methods of scoring metrics may vary widely in the type of scaling used and the way in which metric expectations are established. These differences among scoring methods may influence the performance characteristics of the final index that is created by summing individual metric scores. The Macroinvertebrate Biotic Integrity Index (MBII), a multimetric index, was developed previously for first through third order

streams in the Mid-Atlantic highlands of the United States. In this study, six metric scoring methods were evaluated for the MBII using measures related to site condition and index variability, including the degree of overlap between impaired and reference distributions, relationships to a stressor gradient, within-sample index variability, temporal variability, and the minimum detectable difference. Measures of index variability were affected to a greater degree than those of index responsiveness by both the type of scaling (discrete or continuous) and the method of setting expectations. A scoring method using continuous scaling and setting metric expectations using the 95th percentile of the entire distribution of sites performed the best overall for the MBII. These results showed that the method of scoring metrics affects the properties of the final index, particularly variability, and should be examined in developing a multimetric index because these properties can affect the number of condition classes (e.g., unimpaired, impaired) an index can distinguish.

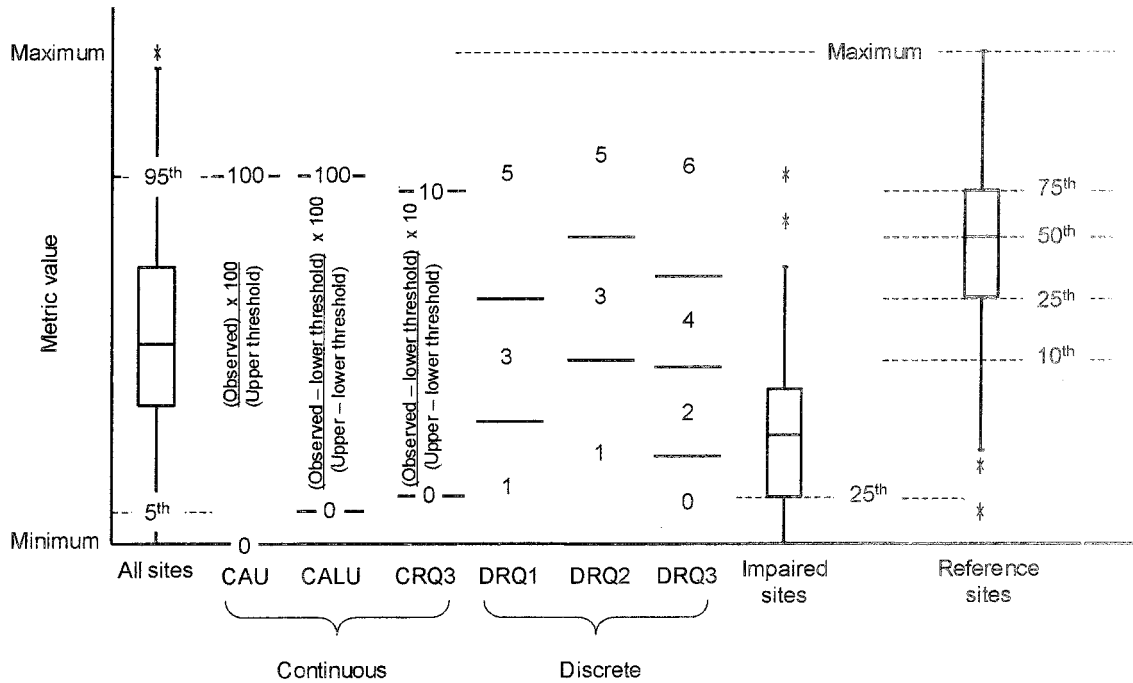
Multimetric indices are commonly used for evaluating the biological condition of water bodies in the United States (Davis and others 1996). Most multimetric indices consist of a number of measures, or metrics, describing a specific assemblage (e.g., fish, macroinvertebrates, or periphyton), which are combined into a single "multimetric" value representing the condition of a water body. Metrics can represent a wide range of ecological characteristics, including taxa richness, pollution tolerance, taxonomic composition, functional feeding groups, and behavioral habits. Thus, the metrics included in an index may have a variety of units, such as the number of taxa, the percentage of total taxa, or the percentage of individuals represented by a particular group. In addition, some indices have incorporated diversity (e.g., Shannon Diversity Index, Margalef 1958) or pollution tolerance (e.g., Hilsenhoff Biotic Index, Hilsenhoff 1987) indices (e.g., Plafkin and others 1989, Barbour and others 1996, Maxted and others 2000, Klemm and others 2003). The component

metric values must be converted into unitless numbers in order to combine them into a single value (Karr and others 1986, Barbour and others 1995). Several methods of standardizing or scoring metric values in this way have been used in various areas of the U.S. (Barbour and others 1999). These different methods, in turn, may influence the properties of the final index because they usually alter the original distributions of individual metrics. Although a limited number of studies have examined certain statistical properties of existing or newly developed indices (Fore and others 1994, Hughes and others 1998, Blocksom and others 2002), little research has directly addressed how different metric scoring methods affect a given multimetric index.

There are two major features of an index that may be influenced strongly by the method of scoring metrics. First, the metric scoring method may affect the relationship between the final index and biological condition relative to that exhibited between individual metrics and biological condition. When individual metrics are scored, the distributions of those metric scores may be very different from the original distributions of the metric values. Thus, relationships of individual metric values to stressor gradients may be strengthened or weakened by the choice of scoring method. This alter-

**KEY WORDS:** Metric scoring; Standardizing; Multimetric index; Mid-Atlantic highlands; Index performance

*Email:* blocksom.karen@epa.gov



**Figure 1.** Example of discrete and continuous scaling and ways of setting metric expectations for the six scoring methods examined in this study. Dotted lines indicate various percentiles of the reference, impaired, and all-site distributions. Method names are coded using the following abbreviations: C = Continuous scaling, D = Discrete scaling, R = Reference sites used to set thresholds, A = All sites used to set thresholds, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used to set thresholds, U = Upper threshold set (all sites only), L = Lower threshold set (all sites only).

ation may be compounded across metrics and affect the ability of the index to discriminate reference and impaired conditions. Temporal, spatial, or measurement variation is another index feature that may be affected by the metric scoring method. If a scoring method results in an index that varies widely in response to minor natural variation in an assemblage (due to either sampling or temporal variation), the index will be less repeatable and reliable as a measure of biological condition. Increased within-site variability also limits the ability of the index to distinguish different levels of biological condition. Among the key characteristics of an ideal indicator are a strong relationship to stressors and low sampling and temporal variability (Cairns and others 1993).

Scoring methods may differ by the type of scaling used (discrete or continuous) and the way in which expectations are set for individual metrics. Discrete scoring was first used by Karr (1981) and has been used extensively for a number of biotic indices (Ohio EPA 1987, Barbour and others 1996, Maxted and others 2000). This type of scoring assigns a series of categorical scores to ranges of metric values and limits each metric to a few possible scores (e.g., 1, 3, or 5) (Figure 1). In contrast, continuous scoring relies on setting upper

and lower thresholds, or expectations, based on the statistical distribution of values, and metric values between the thresholds are scored on a continuous scale as fractions of the expected value (Figure 1).

The manner in which expectations are set for each metric can greatly affect the actual threshold value, but the appropriateness of each method may depend on the nature of the data set. One method of setting expectations uses a set of sites to represent reference conditions and sometimes impaired conditions. Percentiles of the reference and impaired distributions are used as upper and lower scoring thresholds for each metric (Figure 1). Another method of setting expectations is by using a percentile (e.g., 95th percentile) of the entire distribution of sites.

The Macroinvertebrate Biotic Integrity Index (MBII) is a multimetric index developed previously for use in first- through third-order wadeable streams of the Mid-Atlantic Highlands region (MAHR) of the U.S. (Klemm and others 2003). The MBII consists of seven metrics, and the method used in scoring the metrics relied on the distributions of reference and impaired sites. Although the index reflected biological conditions accurately, certain measures indicated that the MBII might have more temporal variability than de-

Table 1. Details of each scoring method for metrics that decrease (increase) with disturbance

Method (reference) <sup>a</sup>	Type of scoring	Range of metric scores	Upper threshold	Lower threshold
CRQ3 (Klemm and others 2003)	Continuous	0–10	75th (25th) percentile of reference	25th (75th) percentile of impaired
CAU (Tetra Tech 2000)	Continuous	0–100	95th (5th) percentile of all sites	0 (100 or maximum possible)
CALU	Continuous	0–100	95th (5th) percentile of all sites	5th (95th) percentile of all sites
DRQ2 (Maryland DNR 1998)	Discrete	1, 3, or 5	>50th percentile of reference	<10th (90th) percentile of reference
DRQ3 (Ohio EPA 1987)	Discrete	0, 2, 4, or 6	75th (25th) percentile of reference	Remaining range quadrisected
DRQ1 (Barbour and others 1996)	Discrete	1, 3, or 5	>=25th (75th) percentile of reference	Remaining range bisected for scores of 3 and 1

<sup>a</sup>Method codes set as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

sired. This concern prompted examination of the scoring method used as a possible source of increased uncertainty in the MBII.

The objective of this empirical study was to evaluate the effect of different metric scoring methods on the performance of the MBII. The original continuous scoring method (CRQ3 method) was compared with five other methods selected to represent different combinations of methods for setting expectations with discrete or continuous scaling (Table 1, Figure 1). I focused on methods currently in use or modifications of those methods. Three scoring methods that use discrete scaling and set expectations using reference distributions were tested. The basis for these scoring methods included the Maryland Benthic Index of Biotic Integrity (B-IBI, Maryland DNR 1998), a modification of the method used for the Ohio Invertebrate Community Index (ICI, Ohio EPA 1987), and the Florida Stream Condition Index (FSCI, Barbour and others 1996). Two continuous scoring methods that relied on the entire distribution of sites were tested. One of these methods is currently used for the West Virginia Stream Condition Index (WVSCI, Tetra Tech 2000), and the other is a modification of this method. For each scoring method, I measured the degree of overlap between impaired and reference distributions, the relationship to a stressor gradient, within-sample index variation, temporal variability, and the minimum detectable difference of overall index scores. These measures or variations on them have been used in various ways to evaluate existing and newly developed multimetric indices (Fore and others 1994, Barbour and others 1996, Hughes and others 1998, Tetra Tech 2000, Maxted and others 2000, McCormick and others 2001, Klemm and others 2003).

## Methods

### Data Sets

The MAHR data used for the development of the MBII (Klemm and others 2003) were used for these analyses. Wadeable stream reaches were sampled in the states of Pennsylvania, Maryland, Virginia, West Virginia, and Delaware, excluding the coastal plains areas. Reaches were selected using a randomized systematic design with a spatial component (Herlihy and others 2000). Data were collected from 506 reaches selected via a randomized systematic design with a spatial component and included first through third order streams (Overton and others 1990, Herlihy and others 2000). Another 68 stream reaches sampled were hand-selected by state and regional biologists (Klemm and others 2002).

Macroinvertebrate data were collected using Environmental Monitoring and Assessment Program—Surface Waters (EMAP-SW) field methods (Lazorchak and others 1998) and laboratory (Klemm and Lazorchak 1994) methods. Samples were collected from April through June of 1993–1995 from a reach equal in length to 40 times the wetted width of the stream. Benthic macroinvertebrates were collected at the inner nine of eleven evenly-spaced transects in the reach. At each transect, a single kick net sample of 20 seconds was collected. Samples collected from riffle habitats and those from pool habitats were composited separately. In the laboratory, a random subsample of 300 organisms ( $\pm 10\%$ ) was removed from debris, and all organisms were identified to the lowest practicable taxon. For development of the MBII, pool and riffle samples were treated separately. For simplicity, I included only data from riffle samples in this study be-

cause the vast majority of sites had some riffle data but many sites had no pool data. Water chemistry samples (Lazorchak and others 1998) and Rapid Bioassessment Protocols (RBP) habitat (Barbour and others 1999) data were also collected at each site. Detailed quantitative physical habitat data (Kaufmann and others 1999) were collected at a subset of sites.

Data were divided into calibration and validation data sets for development of the MBII (Klemm and others 2003) and for some analyses in this study. The calibration data set consisted of samples from 448 sites and the validation data set consisted of samples from 101 sites (Klemm and others 2003). Thirty-five sets of within-year revisits to sites, sampled within the same index sampling period, were used for certain analyses in this study and in developing the MBII. Within the data set, a subset of reaches was identified as reference and impaired sites using chemical and RBP habitat criteria (Waite and others 2000, Klemm and others 2003). To be defined as reference, a reach met all of the following criteria: sulfate  $<400 \mu\text{eq/L}$ , Acid Neutralizing Capacity (ANC)  $>50 \mu\text{eq/L}$ , chloride  $<100 \mu\text{eq/L}$ , total phosphorus  $<20 \mu\text{g/L}$ , total nitrogen  $<750 \mu\text{g/L}$ , RBP mean habitat score  $>15$  (of a possible 20), and at least 150 organisms. A reach was defined as having a recognized impairment if any of the following criteria were met: pH  $<5$ , chloride  $>1000 \mu\text{eq/L}$ , sulfate  $>1000 \mu\text{eq/L}$ , total phosphorus  $>100 \mu\text{g/L}$ , total nitrogen  $>5000 \mu\text{g/L}$ , or an RBP mean habitat score  $<10$ . This subset of reaches was used in some analyses for both the MBII development (Klemm and others 2003) and this study.

In developing the MBII, over 100 macroinvertebrate metrics were evaluated for range, precision, responsiveness to abiotic variables, redundancy, and the relationship to catchment area (Klemm and others 2003). Based on the results of these evaluations, seven metrics were selected for the final index: number of Ephemeroptera taxa, number of Plecoptera taxa, number of Trichoptera taxa, number of collector-filterer taxa, percent non-insect individuals, percent of individuals in the dominant five taxa, and a macroinvertebrate tolerance index (modified from the Hilsenhoff Biotic Index, Hilsenhoff 1987). This study focuses on variation in methods used to score these seven metrics and combine them into an index value.

Number of Ephemeroptera taxa, number of Plecoptera taxa, and number of collector-filterer taxa were related to catchment area when reference sites were regressed on the natural log of watershed area. These metrics were adjusted by calculating the residual for each sample based on the estimated regression equation (Klemm and others 2003). For ease of use, the

residuals were then constrained to be positive by adding a constant to each residual value (Urquhart 1982). The constant was derived simply as the predicted value associated with the most negative residual. These adjusted metric values were used to set scoring expectations for all methods. Obviously, there are other potential methods of adjusting for catchment area, but evaluation of those methods was beyond the scope of this study.

### Scoring Methods

For each of the seven metrics of the MBII, scores were calculated on the calibration data set using six methods. Details on setting expectations and dividing metric values into scores for each scoring method are provided in Table 1 and Figure 1. All of the methods except the CALU method are currently used by at least one bioassessment program. The CALU method is a modification of the CAU method that includes principles of the CRQ3 method. Methods with *C* as the first letter of the abbreviation (CRQ3, CAU, CALU) used continuous scaling, and those with *D* as the first letter (DRQ1, DRQ2, DRQ3) used discrete scaling. Thresholds were set using reference site distributions if the second letter of the method abbreviation is *R* and using all sites where the second letter is *A*. Methods with *Q1*, *Q2*, and *Q3* in the abbreviation used the first, second, and third quartiles of the reference distribution, respectively, as thresholds. Of the two methods using all sites to set thresholds, the CAU method only set upper (U) thresholds and used the minimum value possible for lower (L) thresholds, and the CALU used the distribution to set upper and lower thresholds for scoring. For methods with continuous scaling (CRQ3, CAU, CALU), the metric score was the linear interpolation between the U and L thresholds in the manner used by Minns and others (1994) and Hughes and others (1998). Continuous metric scores were calculated as the difference between the observed value and the lower threshold (0 if no lower threshold was calculated) divided by the difference between the U and L threshold values. For methods with discrete scaling (DRQ1, DRQ2, DRQ3), the thresholds were used in various ways (Table 1, Figure 1) to divide the metric value range into discrete categories. Each category was then assigned one of a discrete number of scores (e.g., 1, 3, or 5). After summing the individual metric scores, the sum was rescaled to a 100-point range by multiplying the total by 100 and dividing by the maximum possible for that index (e.g., 35 for the DRQ1 method).

Table 2. Water chemistry and physical habitat variable weightings from PCA axis 1

Variable	Weighting
<sup>a</sup> RBP channel alteration score	-0.22
RBP embeddedness score	-0.32
RBP epifaunal substrate score	-0.23
RBP riparian vegetation score	-0.29
Percent sand and fines substrate	0.28
Canopy density on bank	-0.21
Riparian Disturbance Index	-0.27
Turbidity	0.25
Chloride ion	0.31
Sulfate	0.17
Total phosphorus	0.27
Total nitrogen	0.31
pH	0.24

<sup>a</sup>RBP = Rapid bioassessment protocols (Plafkin and others 1989).

### Evaluation Methods

The ability of each scoring method to reflect site condition was measured in two ways. The discrimination efficiency of each scoring method was calculated as the percentage of impaired sites scoring below the 25th percentile of reference sites (Tetra Tech 2000). The 25th percentile cutoff was calculated using only the calibration data, and impaired sites from the validation data were used to calculate the discrimination efficiency. This index characteristic provides a measure of the overlap of reference and impaired distributions for a given scoring method, and higher efficiencies indicate smaller overlap of distributions.

Although the ability to distinguish impaired and reference sites is a critical feature of an index, these sites do not represent the full stressor gradient. Thus, I used the results of a Principal Components Analysis (PCA) using physical habitat and log-transformed water chemistry variables (Table 2, Klemm and others 2003) as a second measure of site condition. To more directly assess the ability of an index to reflect this condition, I conducted a Pearson correlation analysis of index scores for each method with the first axis scores from the PCA, which explained approximately 37% of the variation in the chemistry and habitat data. The calibration and validation data sets were combined for this analysis because only a limited number of sites contained data on quantitative physical habitat variables. A higher correlation among methods indicates an index with a stronger relationship to and a better representation of the stressor gradient.

Sources of index variability were measured across scoring methods in three ways. First, each sample was bootstrapped and the six scoring methods were applied to the bootstrapped data to determine the characteris-

tics of the distribution of index scores within sites (Dixon 1992, Fore and others 1994). This analysis was intended to mimic the small differences in metric values that may result from variation due to laboratory subsampling and to evaluate the effect of scoring method on the magnitude of variation at the index level. Fifty samples were selected from the full data set to represent a wide range of index scores. A new bootstrapped sample was generated for each original sample by randomly selecting organisms, with replacement, from the count data for the original sample (typically approximately 300 organisms). This technique results in a new sample with the same number of individuals as the original count of that sample but with varying composition. For selection of each organism for the new sample, the probability of a given species being selected was the same as its proportion in the original count. For laboratory subsampling to be considered valid, it is assumed that the proportion of various species in a random subsample of organisms approximately represents the actual proportions of those species in the entire sample (Fore and others 1994). If this is so, then the new bootstrapped sample simply represents differences in species count data that might result from the variability inherent in laboratory subsampling.

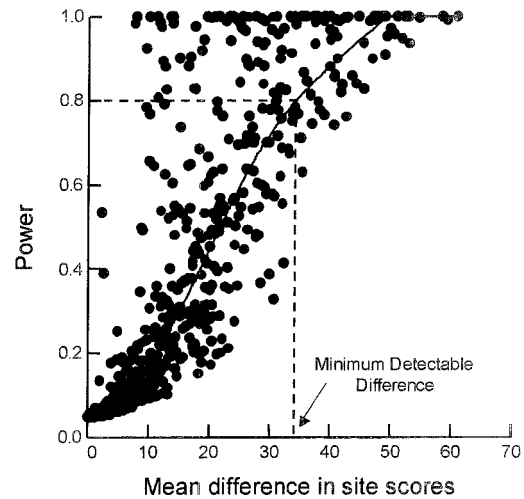
The bootstrapping process was repeated 500 times across 50 samples so that 500 bootstrapped samples were generated from each original sample. The six scoring methods were used to calculate index scores for each of the newly generated samples, and 2.5th and 97.5th percentile values across the 500 bootstrapped samples were determined for each sample and scoring method. The difference in these percentiles is the length of the 95% confidence interval (CI) around the mean score for a sample.

The CI lengths were compared among methods with a repeated measures ANOVA with Tukey multiple comparisons ( $\alpha = 0.05$ ). For a given original sample, variation in composition among the bootstrapped samples can cause variation in metric values, and consequently, in index scores, among the bootstrapped samples. Thus, larger CI lengths indicate more variability in index scores with these relatively minor changes in the composition of a sample (Fore and others 1994) and can be attributed to laboratory subsampling variability. Comparisons of average CI lengths also provide an indicator of the variability contributed by different scoring methods. Smaller CI lengths among methods are more desirable because they indicate that minor changes in composition, as might be expected from laboratory subsampling variability, do not lead to large changes in the final index score or assessment of a site.

The precision of an index can be measured as its

repeatability over a relatively short period of time at the same site. Thus, the second measure of variability was a measure of index precision using within-year revisit data. The precision of an index can be represented as the ability to detect differences among sites (a signal) amid temporal variability within a site (noise). The signal-to-noise (S/N) ratio, used to calculate the precision of habitat attributes by Kaufmann and others (1999, [www.epa.gov/clariton/clhtml/pubtitle.html](http://www.epa.gov/clariton/clhtml/pubtitle.html)), was calculated using revisit, calibration, and validation data for each scoring method. There were 35 sets of within-year revisits in the analysis, with the remainder of sites in the analysis having only one within-year visit. A generalized linear model was calculated with the year as a fixed effect and sampling site nested within year as a random effect. The *F*-statistic for the effect of sampling site, as well as a constant (*c*) varying between 1 and the number of times revisited sites were sampled, were used in the calculation of the S/N ratio  $[(F - 1)/c]$ . All sites were not visited the same number of times within a year, so the value of *c* was estimated using SAS PROC GLM (SAS v.8.2, SAS Institute, Cary, NC). The expected mean square for sites nested within years as a factor in this analysis is a combination of variation due to within-year variability among sites and to within-site sample variation. The coefficient for the variance due to within-year variation among sites, which is provided in the output of PROC GLM, is the value for *c*. Larger S/N ratios among scoring methods are more desirable because they indicate a larger “signal” due to differences in condition among stream sites relative to the “noise” due to temporal variability among samples within sites, and, thus, a higher precision. A S/N ratio of less than 2 is considered very imprecise (Kaufmann and others 1999).

The number of stream condition classes (e.g., good, fair, poor) that can be distinguished by an index provides another measure of index variability (Fore and others 1994, Doberstein and others 2000). This number is obtained by first calculating the minimum detectable difference (MDD) for an index, which is the difference in index scores required to declare that two sites differ in biological condition (Fore and others 1994, Doberstein and others 2000). Dividing the range of the index by the MDD then provides the “maximum number of sites that could be declared different using a given sample size” (Doberstein and others 2000). This is equivalent to calculating the number of condition classes that an index can distinguish (Fore and others 1994, Doberstein and others 2000). This measurement depends on both the number of samples collected and the variability in index scores among samples at individual sites.



**Figure 2.** Example of calculation of the minimum detectable difference (MDD) based on power analysis. Data points represent the calculation of power and mean difference in index scores for each possible pair of sites. Using a LOWESS curve through the points, the mean difference at which the power is 0.80 is identified as the MDD.

In this study, I used power analysis to calculate the MDD for a two-sample *t*-test with a Type I error rate of 0.05 and a power of 0.80. In this case, power was an estimate of the likelihood of detecting differences in condition of a specific magnitude among sites. Power was estimated in SAS/ANALYST (SAS v.8.2, SAS Institute, Cary, NC) using the mean difference in MBII scores between each possible pair of revisit sites (35 sites, 595 pairs of sites), the mean estimate of variance from each pair of sites, and sample sizes of 2 and 3 per site. For each pair of sites, I plotted the power of the test against the mean difference in index scores between those sites. I used robust locally weighted scatterplot smoothing (LOWESS, Cleveland 1979) in SYSTAT (SYSTAT v. 8, SPSS Inc., Chicago, IL) to generate a curve through each plot of 595 points (one plot per scoring method), and the mean difference in index scores that achieved a power of 0.80 (MDD) was estimated from each curve (Figure 2). This approach to power analysis measures the MDD in index scores required to distinguish conditions at two different sites. A different approach that provides similar information to the procedure described above but does not account for the Type II error rate has also been used to calculate an index MDD (Barbour and others 1996). Both of these techniques are in contrast to the use of power analysis to examine the ability of an index to detect regional changes through time, an analysis that relies on a much larger number of sites for a regional assessment of condition (Hughes and others 1998).

An overall evaluation of each scoring method relative to the others was determined using a ranking procedure. The outcomes of each test were ranked from the most desirable values of each characteristic (ranked 1) to the least (ranked 6). In the case of a tie, the average rank was assigned to each of the tied methods.

## Results

### Site Condition

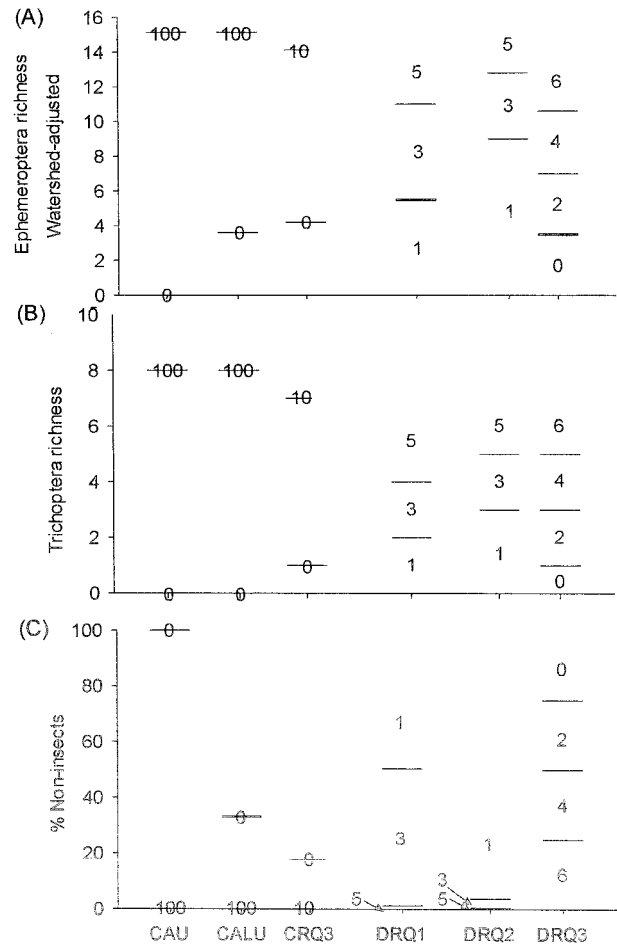
The scoring criteria for each method for watershed-adjusted Ephemeroptera richness, Trichoptera richness, and percent non-insects are provided in Figure 3 as examples of how differences among methods affect scoring ranges.

There were not strong differences among methods with respect to measures of how well site condition was represented. The discrimination efficiency was very high for all six scoring methods, but the 25th percentile values of the index varied greatly (Table 3, Figure 4). The percentage of impaired sites in the validation data set scoring below this cutoff value were very similar among methods and were greater than 94% (of  $N = 38$ ) for all six methods. The discrete methods using the 75th and 25th percentiles of reference sites (DRQ3 and DRQ1, respectively) produced the highest values for the 25th percentile among reference sites (Figures 4E and 4F).

All six methods reflected the physical habitat and water chemistry gradient to a similar degree. The first PCA axis was correlated significantly with the index scores for each of the scoring methods (all  $r > 0.550$ ,  $p < 0.0001$ ,  $n = 547$ ). Pearson correlations between index scores and the first PCA axis were highest for the continuous method with upper and lower thresholds based on all sites, CALU, and the continuous method using the 75th percentile of reference sites, CRQ3 (Table 3).

### Variability

Measures of variability differed more among scoring methods than those measuring the representation of site condition. Bootstrapping the original samples provided an evaluation of the effect of method on variability of the mean score for each original sample due to laboratory subsampling (Table 4). The mean length of the 95% CI was statistically smaller for the continuous method with only upper thresholds based on all sites (CAU) than for all other methods. The discrete method using the median of reference sites, DRQ2, had a statistically larger mean CI length than all other methods, as well as the largest range of CI lengths among the six methods.



**Figure 3.** Scoring criteria for each method for the (A) watershed-adjusted Ephemeroptera richness, (B) Trichoptera richness, and (C) percent non-insects metrics. Note that scoring for percent non-insects is reversed because this metric increases in response to disturbance. Method names are abbreviated as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

The S/N ratio was largest for the CALU method and slightly smaller for the CAU and CRQ3 methods (Table 4), indicating less noise in index scores (variability due to within-year sampling error) relative to the signal (variability attributable to sampling site). The DRQ2 method had the smallest S/N ratio, indicating higher relative noise in the index scores and lower precision.

The power analysis demonstrated that the CAU method produced an index that could distinguish the largest number of condition categories (Table 5). The original scoring method used for the MBII resulted in one of the highest MDD values, indicating that only two

Table 3. Discrimination efficiency and Pearson correlation with stressor PCA axis for each method

Scoring method <sup>a</sup>	25th percentile value of index at reference sites	Discrimination efficiency <sup>b</sup>	Correlation with PCA axis 1
Continuous			
CRQ3	71.2	97.4	0.638
CAU	76.5	94.7	0.585
CALU	69.7	94.7	0.640
Discrete			
DRQ2	65.7	97.4	0.619
DRQ3	81.0	94.7	0.552
DRQ1	88.6	100.0	0.578

<sup>a</sup>Method codes set as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

<sup>b</sup>Percent of impaired sites (N = 38) falling below 25th percentile value of index at reference sites.

condition categories could be distinguished statistically with two samples per site. Increasing the sample size from two to three increased the number of condition classes across methods by between 0.9 and 2.2 categories from that for two samples. The largest increase was for the CAU method.

#### Overall Ranking

Overall, the CALU and CAU methods performed the best based on the ranking procedure (Table 6), with the DRQ1 method also performing better than the CRQ3, DRQ3, and DRQ2 methods. The DRQ2 and DRQ3 methods performed the worst overall. Although the CAU method ranked relatively poorly for measures of the relationship between the index and condition, the actual values for these measures showed little variation among methods. However, this method had the lowest ranks, indicating better performance, for measures of index variability, which varied widely among the methods.

#### Discussion

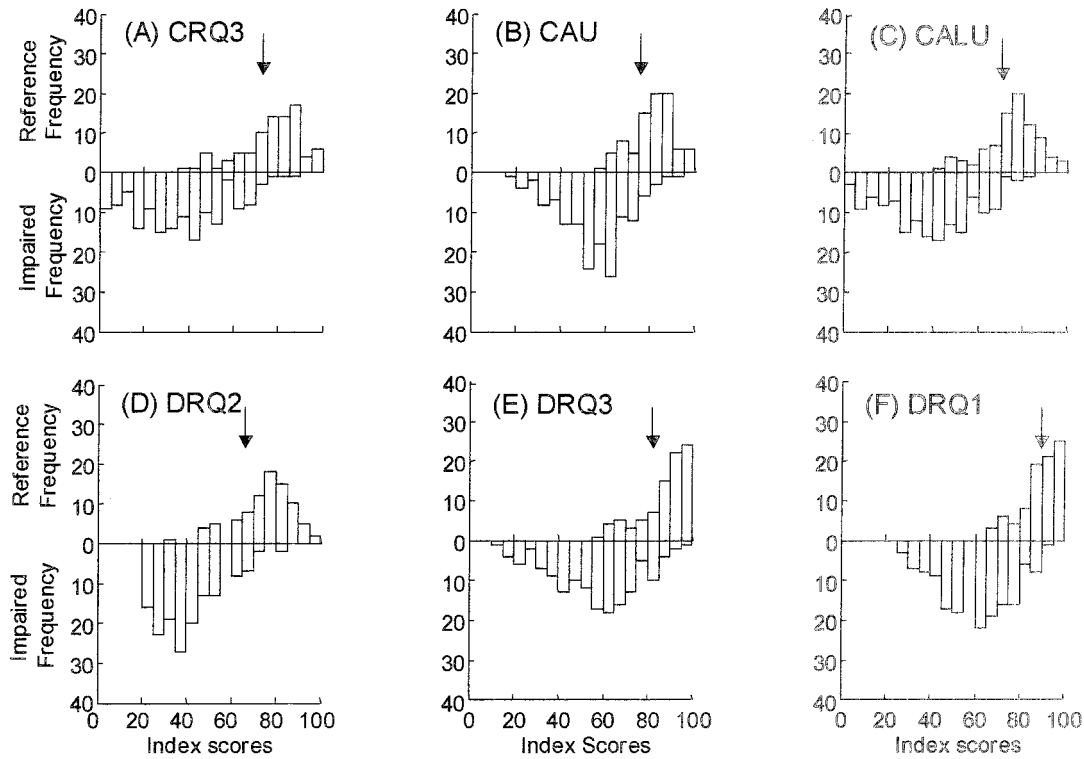
Ideally, a multimetric index reflects site condition accurately, is repeatable through time, and is not sensitive to small changes in composition due to subsampling variability. In this study, the MBII clearly performed differently with respect to these characteristics as a consequence of using different scoring methods. Measures of the relationship of the index to site condition were not particularly informative because the scoring method did not greatly affect discrimination

efficiency or correlations with the stressor gradient represented by the PCA axis. However, measures of index variability were very useful because these values varied greatly among scoring methods.

Although the lack of real differences among methods with respect to site condition might lead some to conclude that all of the scoring methods were equally effective, the ability of an index to consistently measure condition accurately is just as important for use in bioassessment and biomonitoring. The variability of an index can be critical in the ability of state agencies and other groups to create statistically defensible biocriteria. A more variable index results in fewer statistically distinguishable conditions and a reduced ability to detect trends through time or impairment relative to other streams. Thus, controlling and understanding the variability of a given index is necessary for the assessment and protection of water bodies.

Key characteristics of the scoring methods affected index variability to differing degrees. The effects of the type of scoring scale (i.e., continuous or discrete) and the manner in which metric expectations were set were both important. The CAU and CALU continuous methods performed very well overall, but the CRQ3 method had much poorer rankings than the other two continuous methods. In addition, the DRQ1 method, with discrete scaling, performed nearly as well overall as the CAU and CALU methods. These two methods that performed among the best for measures of variability were also the only ones that used the entire distribution of sites for setting scoring expectations. The DRQ1 method also performed well for some, but not all, of these measures, despite using reference site distributions to set expectations. Continuous scoring tended to result in less variable measurements, although using all sites for setting expectations was also advantageous in reducing index variability.

The cause of increased variability in the CRQ3 method was likely the manner in which the maximum and minimum scores were set. For example, for metrics that decrease with increasing disturbance (e.g., number of Ephemeroptera, Figure 3), the upper expectation values did not differ drastically from those of the CAU method, but, by definition, the lower expectations for the CRQ3 method will always be at or above those for the CAU method (Table 1). The range of metric values falling between the upper and lower expectations was then stretched to fill a scoring range from 0 to 10. This means that the range of metric values scoring greater than 0 and less than 10 was smaller for the CRQ3 method than for the CAU method. As a result, a small change in the metric value caused a larger change in the metric score with the CRQ3 method than the CAU



**Figure 4.** Distribution of reference and impaired site scores for each metric scoring method. Downward arrow indicates the 25th percentile of reference sites. The percentage of impaired sites with scores below the 25th percentile of reference sites is the discrimination efficiency.

**Table 4.** Mean 95% confidence interval (CI) length and signal-to-noise ratio for each method

Scoring method <sup>a</sup>	Mean 95% CI length	Range of 95% CI length <sup>b</sup>	S/N ratio
Continuous			
CRQ3	12.8	1.0–26.4 <sup>3</sup>	4.32
CAU	8.4	3.9–14.4 <sup>1</sup>	4.55
CALU	10.5	1.2–22.0 <sup>2</sup>	4.86
Discrete			
DRQ2	15.4	0.0–28.6 <sup>4</sup>	1.99
DRQ3	12.2	0.0–23.8 <sup>3</sup>	3.66
DRQ1	11.2	0.0–22.9 <sup>2,3</sup>	3.14

<sup>a</sup>Method codes set as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

<sup>b</sup>Superscript numbers indicate Tukey multiple comparison groupings (overall  $\alpha = 0.05$ ) based on a repeated measures ANOVA of CI lengths. Grouping numbers represent the ordering of CI lengths from the smallest (1) to the largest (4).

method. Thus, small changes in values across several metrics from year to year can cause relatively large changes in the index score for the CRQ3 method. The

S/N ratio masked this feature to some extent because it included all sites to establish the “signal”. However, in bootstrapping each sample to obtain a 95% CI length, small changes in the composition had a larger effect on the index score for the CRQ3 than for the CAU method. The effect of compressing the range of metric values falling between the minimum and maximum scores is also evident in the power analysis results for the CALU method relative to those for the CAU method.

Similar effects on variability were observed for the DRQ2 method. Although this method did not set specific expectations for a score of 0, it relied on the median of reference distributions to set criteria for the maximum metric scores. The way that the range was divided for the DRQ2 method could lead to very compressed ranges for the middle score of 3 (Figure 3), particularly for reference distributions that are somewhat skewed toward higher values. For this method, the median and 10th percentile values of the reference distribution were used to establish minimum thresholds for scores of 5 and 3, respectively, and the difference between these values could potentially be relatively small. This might result in cases where a slight differ-

Table 5. Number of distinguishable condition classes for each method, based on power analysis

Scoring method <sup>a</sup>	Minimum detectable difference (MDD) (N = 2 per site)	No. condition classes (Index range/MDD)	MDD (N = 3 per site)	No. condition classes
Continuous				
CRQ3	47.5	2.1	28.0	3.6
CAU	34.0	2.9	19.5	5.1
CALU	40.0	2.5	23.5	4.3
Discrete				
DRQ2	45.5	2.2	32.0	3.1
DRQ3	41.2	2.4	25.0	4.0
DRQ1	35.5	2.8	23.0	4.4

<sup>a</sup>Method codes set as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

Table 6. Ranking results<sup>a</sup> for each analysis and scoring method

Scoring method <sup>b</sup>	Discrimination efficiency	PCA correlations	Mean 95% CI length	S/N ratio	Power analysis (MDD)	Sum of ranks
CRQ3	2.5	2	5	3	6	18.5
CAU	5	4	1	2	1	13
CALU	5	1	2	1	3	12
DRQ2	2.5	3	6	6	5	22.5
DRQ3	5	6	4	4	4	23
DRQ1	1	5	3	5	2	16

<sup>a</sup>Results are ranked from most (rank 1) to least (rank 6) desirable conditions.

<sup>b</sup>Method codes set as follows: C = Continuous, D = Discrete, R = Reference sites used to set expectations, A = All sites used to set expectations, Q1, Q2, Q3 = 25th, 50th, 75th percentile of reference sites used for expectations, U = Upper expectation set (all sites only), L = Lower expectation set (all sites only).

ence in metric value from one sample to another could shift the metric score to or from a score of 3. Furthermore, if this was a relatively common occurrence across metrics, these shifts could result in greater variability in the overall index score.

The DRQ1 method had power analysis results similar to those for the CAU method at a sample size of two per site, but at three samples per site, the difference between the methods in distinguishable classes increased. The DRQ1 method also had a much larger range of CI lengths and a considerably smaller S/N ratio. The DRQ3 method performed very similarly to the CALU method for the power analysis and minimum detectable difference measures of variability, although these results were still considerably worse than those for the CAU method. The DRQ3 method truncated the original metric range slightly more than the DRQ1 method by setting the upper threshold at the 75th percentile and quadrisectioning the remaining range of metric values. However, both of these methods compressed the original range of metric values to some degree. For these methods, the division of metric values

into scores resulted in scoring ranges that were generally approximately even in size, as long as the original metric distribution was not highly skewed toward one end of the range. This more even division of the range tended to produce less shifting of scores from one visit to the next and from one bootstrapped sample to another and resulted in lower index variability compared with the DRQ2 method. Even though the distribution of scores was more equally divided across the range of metric values for the DRQ1 and DRQ3 methods, the limited number of possible scores for each metric led to slightly more error related to temporal and laboratory subsampling variability than the best-performing continuous CAU method.

Of the six methods tested, the CAU method, which set scoring expectations based on the entire distribution of sites, produced an index that was closest to ideal. Discrimination efficiency was very high, and the relationship of the index with the PCA axis was relatively high. This method resulted in the least variable index and allowed for the largest number of statistically defensible condition classes. Although the number of

distinguishable stream condition classes calculated in this study may seem low, the maximum number of classes observed for sample sizes of three per site (5.1) was very similar to the results of an equivalent analysis on the Ohio EPA fish IBI, which also was based on three samples per site (Fore and others 1994). Doberstein and others (2000) calculated much higher numbers of distinguishable stream classes based on a single sample per site. However, the estimate of variance used in that analysis represented variation among large numbers of bootstrapped samples, which is a measure of laboratory subsampling variability rather than temporal variability. The overall results for the CAU method were not surprising. This scoring method was identified as less variable among reference lakes in Florida than a discrete scaling method based on similar scoring threshold values (Florida DEP 2000). Additionally, continuous scoring was advocated by Minns and others (1994) and Hughes and others (1998) because it is a more accurate depiction of data, creates a less variable index, and avoids gaps in possible scores.

Although the CAU scoring method produced the best overall index for this set of metrics and data, each scoring method has some limitations for implementation. A set of sampling sites chosen specifically to reflect reference and impaired conditions will yield different expectations from one based on randomly selected sites. For a set of sites representing only reference and impaired conditions, use of reference and impaired distributions to set scoring expectations is appropriate. For sites sampled randomly, the entire range of conditions is assumed to be included in the design. Thus, using a percentile of the entire distribution is appropriate. However, if non-biological criteria are used to define reference and impaired sites from among the total set of sites, and the numbers of reference and impaired sites are reasonably large (50–100), either method can be used to set scoring thresholds. This data set was collected using a random probability design, which met the assumption that the range of conditions was captured. There also were over 500 sites in this data set, which helped to avoid having a single site drastically influence the 5th or 95th percentile metric values. Large numbers of reference and impaired sites were identified using abiotic variables, also allowing examination of other scoring methods.

The primary conclusion of this research is not necessarily that a particular method is always better than all others but that the method of scoring individual metrics can affect the performance of the final index significantly. Evaluating different methods of scoring metrics when developing a multimetric index can make the difference between an index that can distinguish ac-

ceptable from unacceptable site conditions and one that can distinguish multiple levels of condition. The measures used in this study were not an exhaustive set of possible analysis tools, but they provided key insight into how different types of scoring affect important characteristics of an index. Although some situations may warrant use of a specific type of scoring method (e.g., use of reference conditions to set expectations), this type of analysis can help provide the justification for choosing a particular scoring method. Selection of a scoring method producing an index with favorable performance characteristics can increase the confidence in bioassessment results and lead to a stronger biomonitoring program.

### Acknowledgments

Thanks to John Hutchens and Florence Fulk (USEPA, NERL) for helpful advice and suggestions on many aspects of this paper. Thanks to Michael Paul (Tetra Tech) for valuable comments on an earlier draft of this paper. Thanks also to Bob Hughes (Oregon State University), and Phil Larsen, John Stoddard, and Dave Peck (USEPA, NHEERL) for helpful discussions regarding the level of variability in the MBII. The United States Environmental Protection Agency through its Office of Research and Development funded the research described here. It has been subjected to Agency review and approved for publication.

### Literature Cited

- Barbour, M. T., J. B. Stribling, and J. R. Karr. 1995. Multimetric approach for establishing biocriteria and measuring biological condition. Pages 63–77 in W. S. Davis and T. P. Simon (eds.) *Biological assessment and criteria: Tools for water resource planning and decision making*. Lewis Publishers, Boca Raton, Florida.
- Barbour, M. T., J. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White, and M. L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. *Journal of the North American Benthological Society* 15:185–211.
- Barbour, M. T., J. Gerritsen, B. D. Snyder, and J. B. Stribling. 1999. *Rapid Bioassessment Protocols for Use in Wadeable Streams and Rivers*, 2nd ed. U.S. Environmental Protection Agency, Office of Water, Washington, DC. EPA 841-B-99-002. ([www.epa.gov/clariton/clhtml/pubtitle.html](http://www.epa.gov/clariton/clhtml/pubtitle.html))
- Blocksom, K. A., J. P. Kurtenbach, D. J. Klemm, F. A. Fulk, and S. M. Cormier. 2002. Development and evaluation of the Lake Macroinvertebrate Integrity Index (LMII) for New Jersey lakes and reservoirs. *Environmental Monitoring and Assessment* 77:311–333.
- Cairns, J., P. V. McCormick, and B. R. Niederlehner. 1993. A

- proposed framework for developing indicators of ecosystem health. *Hydrobiologia* 263:1–44.
- Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74:829–836.
- Davis, W. S., B. D. Snyder, J. B. Stribling, and C. Stoughton. 1996. Summary of state biological assessment programs for streams and rivers. U.S. Environmental Protection Agency, Office of Planning, Policy, and Evaluation, Washington, D.C. EPA 230-R-96-007.
- Dixon, P. M. 1992. The bootstrap and the jackknife: describing the precision of ecological indices. Pages 290–318 in S. M. Scheiner and J. Gurevitch (eds.) *Design and analysis of ecological experiments*. Chapman and Hall, London, England.
- Doberstein, C. P., J. R. Karr, and L. L. Conquest. 2000. The effect of fixed-count subsampling on macroinvertebrate biomonitoring in small streams. *Freshwater Biology* 44:355–371.
- Florida DEP. 2000. Development of Lake Condition Indexes (LCI) for Florida. Florida Department of Environmental Protection (Florida DEP), Nonpoint Source Bioassessment Program, Orlando, Florida.
- Fore, L.S., J. R. Karr, and L. L. Conquest. 1994. Statistical properties of an index of biological integrity used to evaluate water resources. *Canadian Journal of Fisheries and Aquatic Sciences* 51:1077–1087.
- Herlihy, A. T., D. P. Larsen, S. G. Paulsen, N. S. Urquhart, and B. J. Rosenbaum. 2000. Designing a spatially balanced, randomized reach selection process for regional stream surveys: The EMAP Mid-Atlantic pilot study. *Environmental Monitoring and Assessment* 63:95–113.
- Hilsenhoff, W. L. 1987. An improved biotic index of organic stream pollution. *The Great Lakes Entomologist* 20:31–39.
- Hughes, R. M., P. R. Kaufmann, A. T. Herlihy, T. M. Kincaid, L. Reynolds, and D. P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. *Canadian Journal of Fisheries and Aquatic Sciences* 55:1618–1631.
- Karr, J. R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6:21–27.
- Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assessing biological integrity in running waters: a method and its rationale. Illinois Natural History Survey, Urbana, Illinois. Special Publication no. 5.
- Kaufmann, P. R., P. Levine, E. G. Robison, C. Seeliger, and D. V. Peck. 1999. Quantifying Physical Habitat in Wadeable Streams. U.S. Environmental Protection Agency, Office of Research and Development, Washington, D.C. EPA 620-R-99-003. ([www.epa.gov/clariton/clhtml/pubtitle.html](http://www.epa.gov/clariton/clhtml/pubtitle.html))
- Klemm, D. J., K. A. Blocksom, W. T. Thoeny, F. A. Fulk, A. T. Herlihy, P. R. Kaufmann, and S. M. Cormier. 2002. Methods development and use of macroinvertebrates as indicators of ecological conditions for streams in the Mid-Atlantic Highlands Region. *Environmental Monitoring and Assessment* 78: 169–212.
- Klemm, D. J., K. A. Blocksom, F. A. Fulk, A. T. Herlihy, R. M. Hughes, P. R. Kaufmann, D. V. Peck, J. L. Stoddard, W. T. Thoeny, M. B. Griffith, and W. S. Davis. 2003. Development and Evaluation of a Macroinvertebrate Biotic Integrity Index (MBII) for Regionally Assessing Mid-Atlantic Highlands Streams. *Environmental Management* 31:656–669.
- Klemm, D. J., and J. M. Lazorchak. 1994. Environmental Monitoring and Assessment Program, Surface Waters and Region 3 Regional environmental Monitoring and Assessment Program: Pilot Laboratory Methods Manual for Streams. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring Systems Laboratory, Cincinnati, Ohio. EPA 620-R-94-003.
- Lazorchak, J. M., D. J. Klemm, and D. V. Peck. 1998. Environmental Monitoring and Assessment Program-Surface Waters: Field operations and methods for measuring the ecological condition of wadeable streams. U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory and National Health and Environmental Effects Research Laboratory, Research Triangle Park, North Carolina. EPA 620-R-94-004F.
- Margalef, R. 1958. Information theory in ecology. *General Systems* 3:36–71.
- Maryland DNR. 1998. Development of a Benthic Index of Biotic Integrity for Maryland Streams. Maryland Department of Natural Resources (Maryland DNR), Monitoring and Non-Tidal Assessment Division, Annapolis, Maryland. Report number CBWP-EA-98-3.
- Maxted, J. R., M. T. Barbour, J. Gerritsen, V. Poretti, N. Primrose, A. Silvia, D. Penrose, and R. Renfrow. 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. *Journal of the North American Benthological Society* 19:128–144.
- McCormick, F. H., R. M. Hughes, P. R. Kaufmann, D. V. Peck, J. L. Stoddard, and A. T. Herlihy. 2001. Development of an index of biotic integrity for the Mid-Atlantic highlands region. *Transactions of the American Fisheries Society* 130:857–877.
- Minns, C. K., V. W. Cairns, R. G. Randall, and J. E. Moore. 1994. An index of biotic integrity (IBI) for fish assemblages in the littoral zone of Great Lakes areas of concern. *Canadian Journal of Fisheries and Aquatic Sciences* 51:1804–1822.
- Ohio EPA. 1987. Biological Criteria for the Protection of Aquatic Life, Volume II: Users Manual for Biological Field Assessment of Ohio Surface Waters. Ohio Environmental Protection Agency (Ohio EPA), Ecological Assessment Section, Division of Water Quality, Columbus, Ohio.
- Overton, W. S., D. White, and D. L. Stevens, Jr. 1990. Design report for EMAP Environmental Monitoring and Assessment Program. U.S. Environmental Protection Agency, Environmental Research Laboratory, Corvallis, Oregon. EPA 600-3-91-053.
- Plafkin, J. L., M. T. Barbour, K. D. Porter, S. K. Gross, and R. M. Hughes. 1989. Rapid bioassessment protocols for use in streams and rivers: Benthic macroinvertebrates and fish. U.S. Environmental Protection Agency, Office of Water Regulations and Standards, Washington, D.C. EPA 440-4-89-001.
- Tetra Tech. 2000. A Stream Condition Index for West Virginia Wadeable Streams. Unpublished report. (Available at: [www.tetra-tech.com](http://www.tetra-tech.com))

- [dep.state.wv.us/item.dep?ssid=11&sslid=192](http://dep.state.wv.us/item.dep?ssid=11&sslid=192)). U.S. EPA Region 3 Environmental Services Division and Office of Water, Wheeling, WV.
- Urquhart, N. S. 1982. Adjustment in covariance when one factor affects the covariate. *Biometrics* 38:651–660.
- Waite, I. R., A. T. Herlihy, D. P. Larsen, and D. J. Klemm. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. *Journal of the North American Benthological Society* 19:429–441.