

## A process for creating multimetric indices for large-scale aquatic surveys

**John L. Stoddard<sup>1</sup>**

*Office of Research and Development, National Health and Environmental Effects Research Laboratory,  
Western Ecology Division, US Environmental Protection Agency, 200 SW 35<sup>th</sup> Street,  
Corvallis, Oregon 97333 USA*

**Alan T. Herlihy<sup>2</sup>**

*Department of Fisheries and Wildlife, Oregon State University, 200 SW 35<sup>th</sup> Street, Corvallis, Oregon 97333 USA*

**David V. Peck<sup>3</sup>**

*Office of Research and Development, National Health and Environmental Effects Research Laboratory, Western  
Ecology Division, US Environmental Protection Agency, 200 SW 35<sup>th</sup> Street, Corvallis, Oregon 97333 USA*

**Robert M. Hughes<sup>4</sup> AND Thomas R. Whittier<sup>5</sup>**

*Department of Fisheries and Wildlife, Oregon State University, 200 SW 35<sup>th</sup> Street, Corvallis, Oregon 97333 USA*

**Ellen Tarquinio<sup>6</sup>**

*Office of Water, Office of Wetlands, Oceans, and Watersheds, US Environmental Protection Agency, 1200  
Pennsylvania Avenue, NW, 4501T, Washington, DC 20460 USA*

**Abstract.** Differences in sampling and laboratory protocols, differences in techniques used to evaluate metrics, and differing scales of calibration and application prohibit the use of many existing multimetric indices (MMIs) in large-scale bioassessments. We describe an approach to developing MMIs of ecological condition that is applicable to a variety of biological assemblage types and to spatially extensive (regional, national) aquatic resource surveys. The process involves testing the performance characteristics of candidate metrics in several categories that correspond to key dimensions of biotic condition. The performance characteristics include: information content (range), reproducibility, calibration for natural gradients, responsiveness to stressor gradients, and independence from other metrics. The best-performing metric from each category is included in the final MMI. The consistency of the process enables development of separate MMIs in different regions that can be combined in a national assessment and that are more comparable across regions and taxonomic groups than a set of independently developed MMIs would be. We provide an example of the process applied to macroinvertebrate data from the US Environmental Protection Agency's Wadeable Streams Assessment (WSA) from 3045 sites (of which 1390 were WSA probability sites). The MMIs developed for the WSA demonstrate the feasibility of conducting bioassessments at continental scales and provide a basis for interpreting existing MMIs from regional- and national-level perspectives.

**Key words:** multimetric index, metric evaluation process, bioassessment, ecological indicators, benthic macroinvertebrate assemblage, EMAP, Wadeable Streams Assessment, aquatic resource surveys, large scale, IBI.

<sup>1</sup> E-mail addresses: stoddard.john@epa.gov

<sup>2</sup> herlihy.alan@epa.gov

<sup>3</sup> peck.david@epa.gov

<sup>4</sup> hughes.bob@epa.gov

<sup>5</sup> whittier.thom@epa.gov

<sup>6</sup> tarquinio.ellen@epa.gov

Multimetric indices (MMIs) have become standard tools for bioassessment of macroinvertebrate (Barbour et al. 1995, Klemm et al. 2003, Hering et al. 2004) and fish (Simon and Lyons 1995, Hughes and Oberdorff

1999, Roset et al. 2007) assemblages. From their beginning as semiquantitative measures applied at small scales (Karr 1981), MMIs have evolved into highly quantitative measures used to assess ecological condition at regional (McCormick et al. 2001, Bramblett et al. 2005), national (USEPA 2006), and continental (Harris and Silveira 1999, Hering et al. 2004) scales. As MMIs are used at increasingly larger scales, the need for a standardized process for developing them has become apparent. For example, how does one ensure that an MMI used to assess streams in the mountains of the western US will be comparable to one developed to assess low-gradient streams in the central plains?

We present the current version of the approach that the US Environmental Protection Agency (EPA) Environmental Monitoring and Assessment Program (EMAP) has been developing during the last decade for selecting metrics and combining them into MMIs when the data are collected across spatially extensive areas. Early versions of this approach were used to develop MMIs for fish (McCormick et al. 2001) and macroinvertebrate (Klemm et al. 2003) assemblages in wadeable streams in the Mid-Atlantic Highlands of the US. We added steps to develop MMIs for both aquatic vertebrate (fish and amphibians; Whittier et al. 2007) and macroinvertebrate assemblages (Stoddard et al. 2005a) in streams and rivers in 12 western US states (Stoddard et al. 2005b). We apply this approach to benthic macroinvertebrate data collected during the Wadeable Streams Assessment (WSA) (USEPA 2006), a national assessment of wadeable ( $\sim 1^{\text{st}}\text{--}4^{\text{th}}$  order) streams in the 48 conterminous states (Paulsen et al. 2008).

### Background

One of the challenges of conducting bioassessments at regional and national scales is that, to our knowledge, procedures do not exist for determining whether MMIs developed for small areas can be extended beyond the area for which they were developed. Moreover, to our knowledge, no attempts have been made to identify which metrics in existing indices will be responsive across very large geographic areas. Attempts to aggregate independently developed indices into a regional assessment have met with mixed success (e.g., Herbst and Silldorff 2006, Snook et al. 2007). When faced with the challenge of developing an MMI to assess US wadeable streams, we thought the best strategy was to build a new MMI and to apply the most quantitative approach possible to assure comparability across a large and diverse landscape.

Use of statistical methods to test and select metrics

for use in MMIs has increased steadily in the last decade, but the processes and tests used remain highly variable. Roset et al. (2007) reviewed the most recent methods for developing fish MMIs and concluded that rigorous testing and statistical evaluation of metrics are relatively rare despite the desirability of these steps and that many developers persist in using best professional judgment (BPJ) to select metrics. Work by Fore and Grafe (2002) and Bramblett et al. (2005) are good examples of the general movement away from the BPJ approach and toward use of statistical principles to build MMIs at large scales (in these cases, statewide and multicoregion scales, respectively). Hering et al. (2006) responded to European assessment needs presented by the Water Framework Directive (European Commission 2000) and developed a process that uses quantitative techniques for selecting metrics and building MMIs at the international (continental) scale. The process we describe was used for the WSA in the US and should be regarded as another step in the evolution of MMI development.

The core of our approach is repeatability. The process can be followed with any appropriate data set, in any region, at almost any scale, and will produce results that are internally consistent. MMIs developed with our process can be reproduced by other researchers using the same approach and can be compared to MMIs developed by the same process for other regions. Our approach focuses on defining a set of desirable metric characteristics that incorporate a minimum number of inherent assumptions. We use an iterative series of steps to separate metrics that have these characteristics from metrics that do not. Desirable metric characteristics include: 1) sufficient variability in data values among sites (data range), 2) reproducibility (temporal stability), 3) responsiveness to stressor gradients, and 4) independence from other metrics (Herrick and Schaeffer 1985, Kurtz et al. 2001, Hering et al. 2004). We quantify these characteristics with measured data (Roset et al. 2007). Below, we describe in detail each metric characteristic and the statistical tests we use to evaluate them.

Few researchers, if any, have extensive experience with biological assemblage characteristics over very large geographic areas. Therefore, we decided to consider a very long list of potential metrics to find those that perform well at this scale. When combined with the (now) common practice of using the relative abundance of specific taxa to calculate multiple similar metrics—i.e., numbers of taxa (richness), percentage of taxa, and percentage of individuals (abundance)—the list of candidate metrics can be daunting (e.g., Whittier et al. [2007] evaluated 237 candidate metrics). Therefore, the process to produce a national-scale MMI must

be tailored to the challenge of evaluating hundreds of metrics collected from thousands of sites located across vast and differing geographic areas.

In general, the order in which we apply statistical tests is determined by the ease with which each test can be made (easiest steps first). We use the rationale that more time-intensive tests (e.g., redundancy) should be conducted on the shortest-possible list of candidate metrics. Most of the steps can be viewed as filters through which the metric must pass. Metrics that fail to display any one of the characteristics quantified early in the process are not considered further. Only those metrics that display all of the desired characteristics are used to build a final MMI. Given a choice among multiple good metrics, the process selects those that have the best ability to discriminate good sites from bad. Our philosophy is to develop an MMI that is responsive to a wide array of stressors, some of which might not be well quantified or even known. The process does not focus explicitly on assessing specific stressors or on diagnosing those stressors responsible for producing low MMI scores.

Our approach to MMI development is intended for implementation at regional and national scales. It might be possible to select one set of metrics that possess all of the desired characteristics at the scale of a large region (e.g., the western US) or a nation, but the most discriminating metrics are likely to differ among geographic areas. Therefore, we used classification tools to divide large areas into more homogeneous ecoregions and carried out the metric selection process separately within each ecoregion. For example, in the case of the macroinvertebrate MMIs developed for the WSA, the US was divided into 9 ecoregions that were based on aggregations of Omernik's Level III ecoregions (Omernik 1987). The choice of the number of subdivisions (regions) was based primarily on sample size. A sufficient number of least- and most-disturbed sites was required in each ecoregion to permit the metric selection process to be conducted independently in each ecoregion. Considerable variability will be present in the data, even those from reference sites (Herlihy et al. 2008), used to construct indices at this scale (9 aggregated ecoregions across 48 states). Our approach to MMI development includes elements that deal with regional variability, but future assessments in which the scale of assessment regions is not dictated strongly by sample size probably will produce more-precise indices (e.g., Southerland et al. 2007).

### Methods

The data used to demonstrate our approach to MMI development are drawn from the WSA, a probability

survey of wadeable streams that was conducted from 2000 to 2004 in the US (USEPA 2006). Most sites were sampled synoptically once during a summer baseflow index period. Approximately 10% of the sites were sampled multiple times (usually twice during a single index period and twice during the next index period) to allow assessment of metric variability. Our approach uses reference sites, which we define as streams in least-disturbed condition (Stoddard et al. 2006). Our experience has been that the probability surveys will include a reasonable number of disturbed streams, but that streams at the less-disturbed end of stressor gradients are less common. Therefore, most of the sites in the WSA were selected on a probability basis (Olsen and Peck 2008), but targeted sampling of sites likely to be in least-disturbed condition also was used (Stoddard et al. 2006, Herlihy et al. 2008). Our approach also uses data from sites in most-disturbed condition. We identify these sites by examining sites along a set of chemical and physical-habitat gradients in each ecoregion to determine which sites occupy the least- or most-disturbed ends of these stressor gradients. The criteria for deciding when sites exceed any of the most-disturbed thresholds are listed in Stoddard et al. (2005a).

The WSA (Olsen and Peck 2008) and EPA EMAP (Herlihy et al. 2000, Stoddard et al. 2005a) surveys included collection of ancillary data that were used in the metric selection process. Details of the collection of water-quality and physical-habitat data can be found in Kaufmann et al. (1999), Hughes et al. (2000), and Stoddard et al. (2005a, b).

In the western US, 2 types of macroinvertebrate samples were collected from most of the wadeable stream sites (Peck et al. 2006): 1) reach-wide samples (reach length =  $40 \times$  wetted stream width, minimum length = 150 m) that were a composite of 11 D-frame kick net (500- $\mu$ m mesh, 0.9 m<sup>2</sup> area) samples, one from each of 11 standard transects used to characterize a reach; and 2) targeted-riffle samples that were a composite of 8 D-frame kick net (500- $\mu$ m mesh, 0.9 m<sup>2</sup> area) samples taken randomly from riffles in the same reach as the reach-wide sample. The 8 kick net samples were allocated among the available riffle habitats. The dual sampling protocols were implemented, in part, to allow inclusion of macroinvertebrate data, especially those from reference sites, from other programs (e.g., from the Utah State University Science to Achieve Results program; Stoddard et al. 2005a). In practice, metric values based on the different approaches were nearly indistinguishable at a site (Gerth and Herlihy 2006, Rehn et al. 2007), and we use them interchangeably. Where available, we used metric values estimated from reach-wide samples.

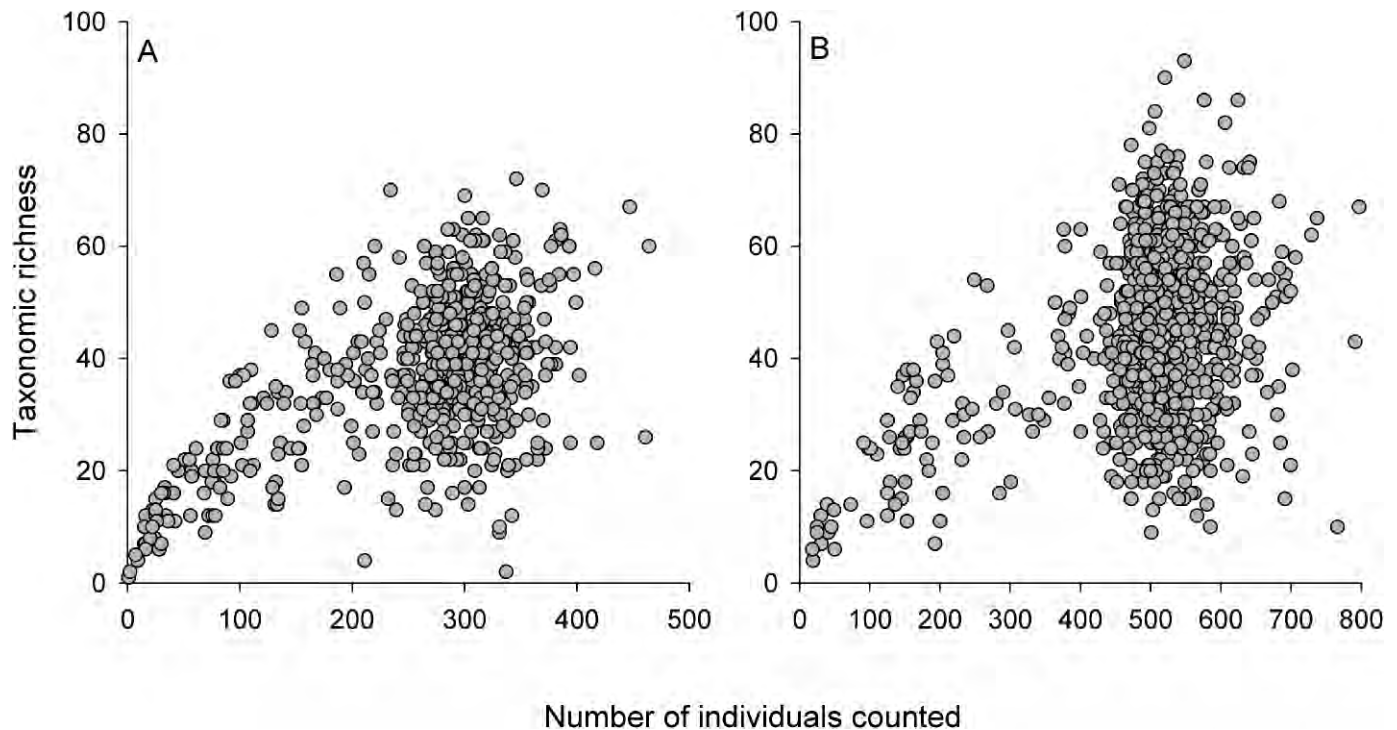


FIG. 1. Dependence of taxonomic richness (number of genera) on number of individuals counted for studies with fixed-count targets of 300 (A) and 500 (B) individuals. Data were collected for the Mid-Atlantic Integrated Assessment (Stoddard et al. 2006a) and the Wadeable Stream Assessment (WSA) (Stoddard et al. 2005), respectively.

Otherwise, we used metric values estimated from targeted-riffle samples. Only the reach-wide protocol was implemented in other parts of the US (Hughes and Peck 2008).

Composite macroinvertebrate samples were preserved in the field with ethanol and transported to the laboratory for processing. A fixed-count protocol was used in which either 500 (for WSA data) or 300 (for the Mid-Atlantic EMAP data) ( $\pm 10\%$ ) individuals were counted and identified to the lowest practical taxonomic level (in most cases, genus level). If the composite sample contained fewer individuals than the fixed-count target, all individuals in the sample were counted and identified.

Laboratory protocols always called for a fixed count, but obtaining an exact fixed count is impractical. Our experience is that, regardless of what fixed-count target is used, the number of taxa identified is strongly dependent on the number of individuals counted (Vinson and Hawkins 1996, Larsen and Herlihy 1998), and this relationship has a potentially strong effect on metrics (especially richness metrics) (Fig. 1A, B). As a partial solution, we resampled WSA data a posteriori to extract a true fixed count of 300 individuals, drawn at random (without replacement) from the data at each site. We also eliminated any sites from consideration as

least-disturbed sites (but not from the assessment data set) if their samples contained  $< 250$  individuals. Metrics that were not based solely on taxonomic information (e.g., number of Ephemeroptera genera) were based on published autecological information (e.g., Merritt and Cummins 1996, Barbour et al. 1999, Klemm et al. 2002, Carlisle et al. 2007).

Our article focuses on development of a reproducible process for selecting metrics and combining them in an MMI. Many other issues that are critically important to developing MMIs are not within the scope of our article. These issues include plot-scale sampling design (Hughes and Peck 2008), taxonomic resolution (Waite et al. 2004, Chessman et al. 2007), ambiguous taxa (Cuffney et al. 2007), incomplete autecological information, effects of different fixed-count targets on metric variability, and variable reference-site quality (Bailey et al. 2004).

#### *Steps in metric selection*

*1. Classification of metrics.*—The intent of the original MMIs was to use metrics to characterize inherent qualities of aquatic assemblages that would capture key elements of biotic condition (Karr 1981, Karr and Dudley 1981). For example, in early fish indices of biotic integrity (IBIs; Karr and Chu 2000), metrics were

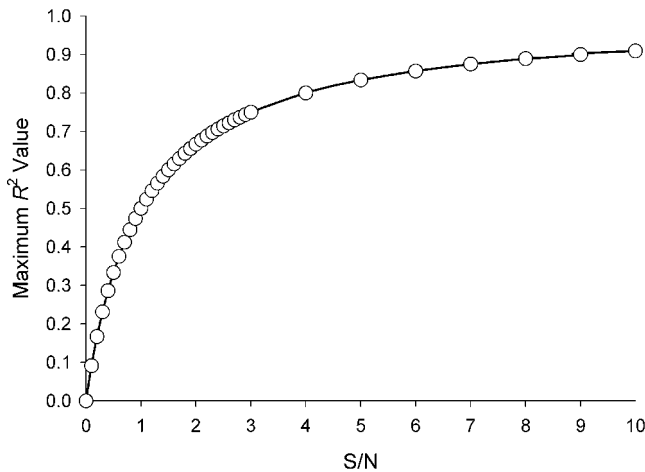


FIG. 2. Effect of the signal:noise ratio (S/N) of a metric on the maximum possible  $R^2$  value of its association with a single stressor.

included to characterize taxonomic richness, tolerance/intolerance, trophic structure, and individual health (anomalies). MMIs developed later often included other ecological attributes, such as age structure, reproductive guilds, life history/behavioral guilds, habitat guilds, or alien or target-species population size (Hughes et al. 1998). Barbour et al. (1999) recommended grouping macroinvertebrate metrics by categories, such as taxonomic richness, taxonomic composition, tolerance/intolerance, feeding group (e.g., predators, scrapers, filter feeders), and habit type (e.g., clingers, burrowers). Our approach is designed to maintain the original intent of MMIs, and we classify candidate macroinvertebrate metrics into 1 of 6 categories: 1) richness, 2) evenness/diversity, 3) composition (relative abundances), 4) functional feeding groups, 5) habit (predominant behavior of each taxon), and 6) tolerance based on published and unpublished pollution tolerance values (Hilsenhoff 1987, Barbour et al. 1999, Carlisle et al. 2007, M. R. Vinson, National Aquatic Monitoring Center, unpublished data). We select metrics independently from within each category, with the goal of including an equal number of metrics from each class in the final MMI. The creation of candidate metrics and classifying those metrics into appropriate metric categories are the 2 aspects of our approach for which biological information and expertise are critical.

**2. Metric range.**—The first filter through which candidate metrics must pass is the range test. The range is the distribution of metric values across all of the available data. The goal is to eliminate metrics that have very small ranges (e.g., richness metrics based on only a few taxa) or that have similar values at most sites (e.g., most sites have values = 0). A small range

could indicate that a metric might not vary sufficiently across sites to discriminate among sites in different conditions. Assigning scores to metrics (see *Metric scoring and calculation of final MMI* below) with small ranges also is problematic. For example, we calculated metrics for Megaloptera (richness, % of taxa, and % of individuals) for the WSA, but Megaloptera were rarely found in the samples (>½ of the sites had no taxa, 90% of sites had either 0 or 1 taxon, maximum Megaloptera richness = 2). We have no fixed threshold below which we eliminate metrics. However, we generally eliminate metrics if their range is <4 or if >⅓ of samples have values = 0. In practice, very few macroinvertebrate metrics are eliminated by this test, but it does eliminate a large number of potentially poor metrics for assemblages with fewer taxa (e.g., fish).

**3. Reproducibility.**—The use of metrics that have relatively stable values at individual sites helps ensure that between-site differences in individual samples are caused by differences in stream condition rather than by sampling variation within a site. Sampling variation is estimated from repeat visits to individual sites. Available measures of sampling variation reflect several sources of variability (i.e., short-term inter-period temporal variability, spatial variability within the reach, and laboratory variability). Low sampling variation is necessary if a metric is to have a high probability of discriminating between sites in good and poor condition, and sampling variation should be small relative to the size of the among-site differences to be discriminated.

We quantify metric reproducibility with a variant of the signal:noise ratio (S/N). S/N is the ratio of the variance among all sites (signal) to the variance of repeated visits to the same site (noise) (Kaufmann et al. 1999). Metrics with high S/N values are more likely to show consistent responses to stressors than are metrics with low S/N values. A metric that is perfectly correlated with a hypothetical stressor and that has no sampling variability will have an  $R^2 = 1.0$  for that stressor. As S/N decreases, the maximum possible  $R^2$  value of the regression decreases because the sampling variability of the metric increases. When  $S/N = 4$ , a perfect correlation between the metric and the stressor would produce  $R^2 = 0.5$  (Fig. 2). We have no fixed threshold below which we eliminate metrics based on S/N. However, S/N values  $\leq 1$  indicate that visiting a single site twice yields as much metric variability as visiting 2 different sites. In practice, the threshold depends on the inherent level of variability in the assemblages being assessed and might depend on other factors, such as generation times of the organisms in the assemblages (Stoddard et al. 2005b). Our experience has been that fish metrics commonly have

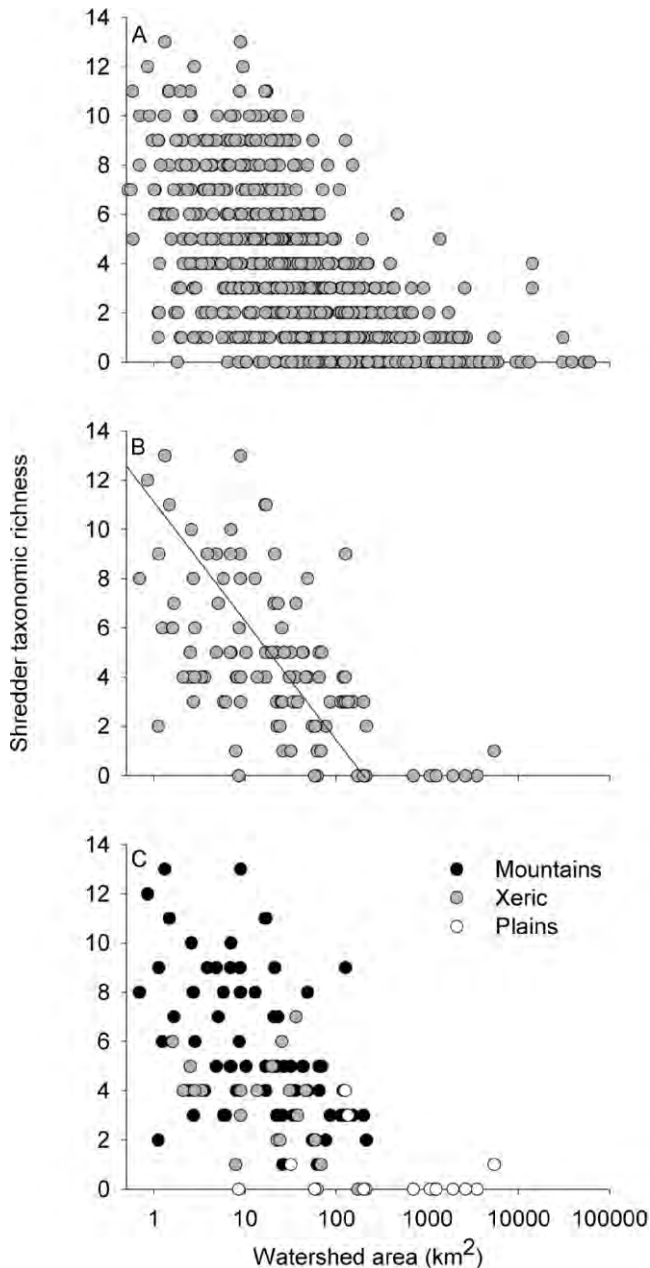


FIG. 3. Example of correlation between a metric (shredder taxonomic richness) and a natural gradient (watershed area) at all Environmental Monitoring and Assessment (EMAP) Western Pilot Study (EMAP-West) sites (A), EMAP-West reference sites (B), and EMAP-West reference sites in each climatic region (C). The regression line in panel B represents the true dependence of shredder taxonomic richness on watershed area in the absence of anthropogenic disturbance.

high S/N values and, therefore, a high threshold for rejection (4 or 5), whereas periphyton metrics have low S/N values and a low threshold for rejection (1 or 1.5). Macroinvertebrate metrics have intermediate S/N values, and we often use a threshold for rejection = 2.0.

4. *Adjusting for natural gradients.*—Metric values often vary with both the stressor gradients being assessed and natural gradients (e.g., elevation, slope, stream size; Fig. 3A). The stressors themselves might vary along the same natural gradients. Thus, knowledge of how to apportion the variability in metric values between natural and anthropogenic gradients is important. We try to avoid selecting metrics that appear to respond strongly to some stressor but, in fact, are merely correlated with the same natural gradient with which the stressor is correlated.

One simple technique for normalizing metrics for natural gradients is to remove the stressor gradient from the data by focusing solely on reference-site data and to quantify the remaining correspondence between the metric value and the natural gradient (Fig. 3B). The regression line in Fig. 3B represents the true dependence of shredder taxonomic richness on watershed area in the absence of anthropogenic disturbance. We use this disturbance-free (reference) relationship to predict the expected metric score at every probability site (based on its position on the natural gradient) and then use the differences between observed and expected metric values (i.e., residuals) as an adjusted metric.

Calibrating metrics for natural gradients appears to be more important for aquatic vertebrate MMIs than for macroinvertebrate MMIs. For example, Whittier et al. (2007) evaluated 12 stream-size-corrected metrics for aquatic vertebrate assemblages for the western US, but ultimately, none were retained in the final MMIs. In the case of the WSA macroinvertebrate data, none of the most responsive metrics required calibration for stream size or slope.

5. *Responsiveness.*—The ultimate test of the effectiveness of a metric is its ability to distinguish degraded from relatively undisturbed streams. This responsiveness can be tested in a number of ways. For example, metrics can be chosen on the basis of their correlation with specific stressors (e.g., nutrients, organic pollution, sedimentation). McCormick et al. (2001) used scatter plots to assess whether individual metrics showed predictable relationships with individual stressors. Some of the original metrics used by Karr (1981) were chosen on the basis of their hypothesized responses to specific aquatic stressors (e.g., darters and benthic disturbance). However, several difficulties arise when metrics are evaluated in terms of their relationships with specific stressors. First, many stressors are highly correlated with one another, and attributing metric response to any particular stressor is problematic. Second, not all stressors are well quantified (e.g., short-lived pesticides or herbicides), or even known, at all sites.

A more general approach is to base the evaluation of

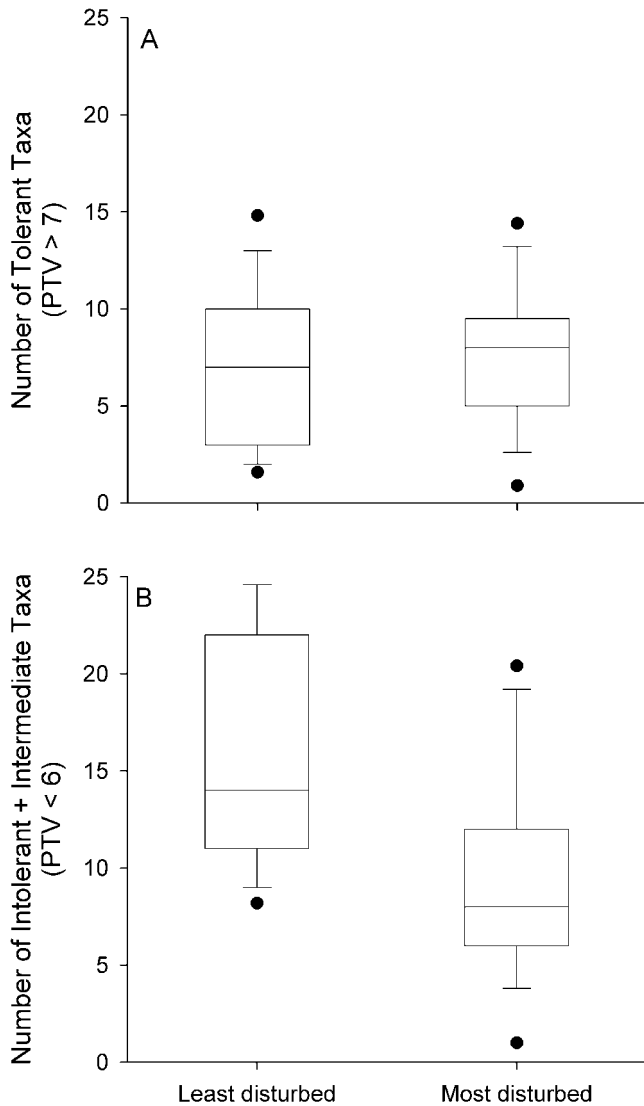


FIG. 4. Example of the use of least-disturbed and most-disturbed sites to test metric responsiveness (Northern Plains streams). A.—Metric with poor responsiveness (number of tolerant taxa,  $t$ -score = -0.9). B.—Metric with good responsiveness (number of intolerant or intermediate taxa,  $t$ -score = 5.7). PTV = pollution tolerance value. Boundaries of boxes indicate interquartile range (25<sup>th</sup>–75<sup>th</sup> percentiles) and median value (midline); whiskers indicate 10<sup>th</sup> and 90<sup>th</sup> percentiles, and symbols indicate 5<sup>th</sup> and 95<sup>th</sup> percentiles.

responsiveness on the ability of a metric to distinguish least-disturbed (reference) from most-disturbed sites. Metric scoring and threshold selection typically are based on a set of least-disturbed sites. For the WSA, least-disturbed sites were chosen by filtering available data to eliminate sites with values above regional thresholds for multiple stressors (Herlihy et al. 2008). We choose most-disturbed sites by filtering available data to eliminate sites with values below regional

thresholds for multiple stressors. We use  $t$ -tests to compare mean values of each metric within each aggregated ecoregion between least- and most-disturbed sites as a test of responsiveness (Fig. 4A, B). We use the  $t$ -value to score responsiveness regardless of its statistical significance. At very large spatial scales, such as those in WSA, sample sizes and gradients in site quality are very large, and even very small values of  $t$  tend to be statistically significant. We identify the metrics that have the highest responsiveness ( $t$ -scores) within each class and aggregated ecoregion and begin selection of final metrics for inclusion in the MMI.

6. *Final metric selection and check for metric redundancy.*—We consider all metrics that pass the filters for range and reproducibility for inclusion in the MMI. We choose the candidate metric that is the most discriminating (i.e., highest  $t$ -score) first. We then proceed iteratively and add the most responsive metric from each metric category until all categories are represented. We base this iterative process on the assumption that choosing the most responsive individual metrics will yield the most robust MMI, provided that each metric provides unique information (i.e., that the metrics are not redundant).

One of the most difficult issues at this step is deciding what constitutes redundancy. Metric redundancy can be defined in at least 2 very different ways: 1) metrics provide very similar biological information, or 2) metrics are highly correlated with other metrics. Philosophically, we might think of the 1<sup>st</sup> of these definitions as most important. We do not want to include 2 metrics that are based on identical (or broadly overlapping) biological or taxonomic information. For example, Ephemeroptera, Plecoptera, Trichoptera (EPT) taxonomic richness, % EPT taxa, and % EPT individuals all rely on information from an identical list of taxa. We might avoid including  $\geq 2$  of these metrics in an MMI because other nonredundant metrics would contribute more new or unique information to the final index even if they are less responsive. However, apparently redundant metrics might, in fact, be relatively uncorrelated. In our experience, the EPT metrics respond to different elements of environmental disturbance and are not universally correlated despite the fact that they are crafted from an identical list of taxa.

Alternatively, we might avoid including metrics with values that are strongly correlated. If 2 metrics covary because the same taxa are changing in abundance as levels of disturbance rise and fall, then the metrics are clearly correlated. However, if metrics covary because they respond to similar stressors, we might question whether correlated metrics are necessarily redundant. For example, consider a case in

which one metric responds strongly to total P concentrations, whereas another responds to total N. The macroinvertebrate metrics might appear strongly correlated because these 2 stressors covary (in the WSA  $R^2 = 0.41$ ,  $p < 0.0001$ ). We do not regard metrics as redundant simply because their correlation coefficients are large when the entire data set is analyzed.

We assume that metrics that covary in response to natural gradients probably are driven by changes in abundances of the same, or similar, taxa. Therefore, we use only reference site data when we examine metrics for redundancy. By eliminating the stressor gradients from this step, we avoid eliminating metrics that respond to similar stressor gradients but that reflect very different taxonomic information. In practice, we often consider metrics as too strongly correlated when their Pearson correlation coefficients at least-disturbed sites are  $>|0.71|$  ( $R^2 \approx 0.5$ ). Interpreted loosely, metrics with  $r > |0.71|$  share  $>1/2$  of their information content.

7. *Metric scoring and calculation of final MMI.*—The iterative process described to this point identifies a single metric in each metric category and ecological region that has all of the desirable characteristics—sufficient range, reproducibility, responsiveness, and independence. The final steps in building an MMI are to normalize all metrics so that they are scored on a common scale and to combine them in a final MMI.

Numerous methods exist for scoring metrics, but the primary decision is whether to score metrics discretely with values of 1, 3, or 5, based on a subjective assessment of the range of each metric (the original Index of Biotic Integrity approach) or to score the metrics on a continuous, but consistent scale. Discrete scoring can have the effect of increasing the variability of the final MMI (Blocksom 2003) and could limit its ability to differentiate among ecological condition classes, although differences usually are fairly minor (Ganasan and Hughes 1998, Blocksom 2003). We use continuous scoring (0–10) to avoid the subjective nature of discrete scoring (e.g., the assumption that sites with metric scores in the top  $1/3$  of the range of values are all in good condition, whereas sites in the bottom  $1/3$  of the range of values are in poor condition) (Hughes et al. 1998). Key goals of our approach to MMI development are consistency and reproducibility, which require that metric scores be quantified in a way that can be easily explained to and applied by other users of the data. In addition, we think it more useful to maintain the continuous nature of many of the best performing metrics (e.g., for use in examining associations between metric scores and other environmental variables) than to reduce them to discrete values.

Use of continuous scoring requires consideration of how to set ceiling and floor values for each metric, i.e.,

how to decide what values of a metric indicate good biological condition (score = 10) and what values indicate poor condition (score = 0). We used the 95<sup>th</sup> percentile of the reference-site distribution of values for each metric as the scoring ceiling and the 5<sup>th</sup> percentile of the distribution of values at all sites as the scoring floor. This approach is similar to the approach recommended by Blocksom (2003), and it produces an MMI with the best responsiveness and lowest variability (measured by S/N). Metric values between the ceiling and floor are interpolated linearly to yield values between 0 and 10, and the final MMI for a site is calculated as the sum of its scored metrics. For convenience when interpreting final scores, we rescale the final MMI to range from 0 to 100.

## Results and Discussion

Our approach to MMI development is based on experience that has been gained as MMIs have evolved from site-specific uses to larger-scale evaluations of populations of aquatic resources. For large-scale assessments, our approach offers several advantages over development processes that are analysis intensive. Our approach can be applied consistently across multiple regions of interest and improves comparability of results among regions. It incorporates the critical performance issues for MMIs identified by others: reproducibility, responsiveness, natural variability, and redundancy. The process efficiently screens large numbers of candidate metrics and their variants (richness, % taxa, and % individuals) and many data points. For large-scale assessments, such as the WSA, our approach is facilitated by use of consistent sample collection and subsampling protocols, consistent levels of taxonomic resolution across all regions, and a consistent approach to selecting least-disturbed and most-disturbed sites with criteria that are adjusted by region.

### *Metric selection*

Issues remain for each step in the process described above. For example, should metrics be selected to ensure that different assemblage attributes are represented in the final index or should they be selected on the basis of their capability to diagnose specific causes of poor condition? Our approach is similar to the general approach described by Hering et al. (2006). It is an attempt to retain the ecological foundation obtained by including metrics that represent different structural and functional attributes of an assemblage and to provide demonstrated reproducibility and responsiveness to human disturbance.

Metric selection also can be based primarily on

responsiveness to a specific stressor (Hering et al. 2006). In this approach, different metric categories (i.e., composition, richness, sensitivity, and functional) are considered, metrics that are responsive to a stressor are combined into a single value that reflects the severity of the stressor, and values derived for different stressors are then combined into an MMI. Ofenböck et al. (2004) provide an example of this approach, but their regional MMIs were developed to respond to a single major stressor that differed among regions. Responsive metrics also can be selected from groups identified in multidimensional trait space by cluster analysis, as demonstrated by Cao et al. (2007) for diatom assemblages. How groups of metrics identified a posteriori correspond to the a priori metric categories we used or to similar metric categories proposed by others (e.g., Fore and Grafe 2002, Hering et al. 2006) is not known.

#### *Geographic scale and natural variation*

The S/N test used in our approach is straightforward and easy to interpret. However, this test is affected by scale. The signal (among-site variation) is expected to be higher relative to noise (sampling variation) at larger than at smaller geographical scales. Hence, metrics that pass this test for a regional MMI might not be suitable for use at local scales, where among-site variation might be small relative to sampling variation. We had to use a relatively low criterion value to screen benthic metrics ( $S/N = 2$ ), a result that suggests some components of sampling variation were larger than desired. Components of sampling variation that can be controlled are within-site temporal variability (shorten the index period or adjust it to be more consistent within and among regions) and measurement error during sampling (which we attempted to minimize with consistent collection and analysis protocols). Detailed evaluation of residual variance components might prove useful for determining whether improvements can be made. At local scales, more importance might be placed on minimizing sampling variation relative to responsiveness (i.e., one might sacrifice some responsiveness to include a more reproducible metric from a given metric category).

At large geographical scales, natural gradients become large and cannot be ignored when evaluating metrics and resulting MMIs. Evaluation of metrics for natural gradients should be based on values at least-disturbed sites because gradients of human disturbance often covary with natural gradients. We use residuals (calculated from regressions using least-disturbed sites) to account for deviation in metric

values (in both directions) from expected values. This approach differs from the historical approach based on “maximum species richness lines” (e.g., Rankin and Yoder 1999), which assumes that richness increases monotonically along a natural gradient (e.g., stream size) and decreases universally as anthropogenic disturbance increases. Stoddard et al. (2006) discuss an example for fish taxon richness, where, at any point on the stream-size gradient, richness is highest in streams with intermediate levels of disturbance. Our results suggested that some metrics were correlated with natural gradients, but these metrics did not pass other filters and were not included in the final set of metrics used for the MMIs. Cao et al. (2007) and Pont et al. (2006) demonstrated for diatom and fish assemblages, respectively, that developing predictive models that describe metric responses (rather than just taxonomic richness) in the absence of human disturbance could improve the performance of MMIs and comparability across regions. This approach requires a large number of least-disturbed sites, consistent methods, and additional expertise to build and interpret the models, but should be considered further for use in large-scale assessments in the US (Pont et al., in press). A similar approach based on nearest-neighbor analysis improved the discriminatory ability of metrics derived from a regional data set compatible with the WSA (Bates Prins and Smith 2007).

Other approaches can be used to account for natural variability. Analysis of data at small geographic scales might remove most of the effects of a natural gradient (e.g., Ode et al. 2008). For example, in the EMAP Western Pilot Study (EMAP-West), small watersheds tend to occur in mountains, and large watersheds occur in the plains (Stoddard et al. 2005a). Shredder taxonomic richness varies systematically along a gradient of watershed area at all sites (Fig. 3A) and at reference sites only (Fig. 3B). However, within each of 3 climatic regions (Mountains, Xeric, and Plains) in the western US, shredder taxonomic richness is not strongly related to watershed area (Fig. 3C), and the metric need not be adjusted for watershed area if the data are analyzed at the geographic scale of these regions. Classification of sites in the Coast Range of the northwestern US by size, lithology, and bed stability reduced the effect of natural variability in an IBI (Kaufmann and Hughes 2006). In Europe, modeling has been used to reduce metric variability by removing the effects of size, air temperature, elevation, slope, and watershed area (Oberdorff et al. 2002, Pont et al. 2006). Classification and regression trees (CART) have been used to remove effects of natural gradients from diatom metrics (Cao et al. 2007). The CART approach

TABLE 1. Metrics selected with our approach to multimetric index (MMI) development and used in the Wadeable Stream Assessment (WSA) (USEPA 2006). Metrics were selected and scored separately for each of 9 aggregated ecoregions used in the WSA. Values are *t*-scores used to distinguish least-disturbed from most-disturbed sites. *t*-scores were calculated for metrics based on data from sites within each ecoregion, for the ecoregional MMIs, and for a national MMI based on all sites in the data set. *t*-scores are shown only for the best metric in each category and ecoregion. Metrics shown in bold were adequate in all ecoregions and were used to create the national MMI. NAP = Northern Appalachians, SAP = Southern Appalachians, CPL = Coastal Plain, UMW = Upper Midwest, TPL = Temperate Plains, NPL = Northern Plains, SPL = Southern Plains, WMT = Western Mountains, XER = Xeric West, EPT = Ephemeroptera, Plecoptera, Trichoptera, PTV = pollution tolerance value.

| Metric category       | Individual metric                | Aggregated ecoregion |      |      |      |     |     |     |      |      |
|-----------------------|----------------------------------|----------------------|------|------|------|-----|-----|-----|------|------|
|                       |                                  | NAP                  | SAP  | CPL  | UMW  | TPL | NPL | SPL | WMT  | XER  |
| Taxonomic composition | % <b>EPT taxa</b>                | 7.8                  |      |      |      |     | 4.2 |     | 15.3 |      |
|                       | % EPT individuals                |                      |      |      |      | 6.7 |     | 3.9 |      |      |
|                       | % noninsect taxa                 |                      |      |      |      |     |     |     |      | 12.1 |
|                       | % noninsect individuals          |                      |      | 8.6  |      |     |     |     |      |      |
|                       | % Ephemeroptera taxa             |                      | 11.4 |      |      |     |     |     |      |      |
| Taxonomic diversity   | % Chironomid taxa                |                      |      |      | 8.2  |     |     |     |      |      |
|                       | <b>Shannon diversity</b>         |                      | 5.0  | 6.1  | 3.3  | 4.7 |     | 2.9 |      |      |
|                       | % individuals in top 5 taxa      | 3.1                  |      |      |      |     |     |     | 3.4  | 7.2  |
| Feeding group         | % individuals in top 3 taxa      |                      |      |      |      | 2.7 |     |     |      |      |
|                       | <b>Scraper richness</b>          | 3.2                  | 6.0  |      |      | 6.7 | 4.3 | 3.6 | 6.4  | 5.2  |
| Habit                 | Shredder richness                |                      |      | 5.9  | 4.4  |     |     |     |      |      |
|                       | % <b>burrower taxa</b>           |                      | 11.6 |      | 7.8  |     | 4.2 | 5.0 |      |      |
| Taxonomic richness    | % clinger taxa                   | 11.3                 |      | 9.8  |      |     |     |     | 13.8 | 10.6 |
|                       | Clinger taxonomic richness       |                      |      |      |      | 6.1 |     |     |      |      |
|                       | <b>EPT taxonomic richness</b>    | 10.5                 | 11.4 | 8.0  | 5.9  |     |     | 4.2 | 11.1 | 12.3 |
| Pollution tolerance   | Ephemeroptera taxonomic richness |                      |      |      |      | 7.0 |     |     |      |      |
|                       | Total taxonomic richness         |                      |      |      |      |     | 3.2 |     |      |      |
|                       | <b>Intolerant richness</b>       |                      |      |      |      |     | 4.4 | 3.4 |      |      |
|                       | % tolerant individuals           |                      | 6.9  | 8.7  |      |     |     |     | 10.7 | 8.4  |
|                       | % taxa with PTV = 0–5.9          | 13.9                 |      |      |      |     |     |     |      |      |
|                       | % taxa with PTV = 6–10           |                      |      |      | 6.3  | 7.1 |     |     |      |      |
| <b>Regional MMI</b>   |                                  | 12.6                 | 14.7 | 12.0 | 10.1 | 9.7 | 6.6 | 5.7 | 15.7 | 14.7 |
| <b>National MMI</b>   |                                  | 10.5                 | 12.5 | 9.6  | 7.1  | 8.2 | 6.6 | 5.8 | 11.5 | 12.0 |

is similar to ours in that CART is used on reference-site data, and residuals are calculated for all sites.

### Responsiveness

Responsiveness to disturbance is the primary criterion used to select metrics in our approach, and we have focused on the ends of the disturbance gradient (as measured by deviation from least-disturbed condition in a region). The range of *t*-scores for metrics selected in each region (Table 1) illustrates some important concepts of our approach to MMI development. The process outlined here selects the most responsive metric, regardless of metric category, for initial entry into the final MMI. No 1 metric category consistently produces the most discriminating metrics. Metrics in taxonomic composition, habit, taxonomic richness, and pollution tolerance categories all have the highest *t*-scores in 1 regions. In general, diversity metrics, followed by feeding group metrics, show the poorest performance. Only 1 metric (% of taxa with Pollution Tolerance Value [PTV] 0–5.9) in 1

aggregated ecoregion (Northern Appalachians) of the 54 possible metric × aggregated ecoregion combinations (Table 1) was more discriminating than the final MMI. In most cases, final MMIs were considerably more responsive than any of the component metrics. For example, in the Upper Midwest, Temperate Plains, and Northern Plains, *t*-scores for final MMIs were 25 to 50% higher than *t*-scores for the most responsive metric in the region. The range of *t*-scores for the MMIs is 5.7 in the Southern Plains to 15.7 in the Western Mountains, and *t*-scores tended to be higher in less disturbed aggregated ecoregions.

As a test of our process, we used the metric screening process with combined data from all 9 aggregated ecoregions to identify metrics within each metric category that did reasonably well everywhere (Table 1). These metrics were scored separately in each aggregated ecoregion, and a national-level MMI was calculated. The national MMI did as well as the ecoregional MMIs in the aggregated ecoregions where the separation between least-disturbed and most-disturbed MMI scores was small (Southern and

Northern Plains). However, the national MMI was substantially less able to discriminate least- from most-disturbed sites in the aggregated ecoregions with the highest *t*-scores (Western Mountains, Xeric West, and Southern Appalachians). This national MMI was not used in the final WSA condition assessment because the regional MMIs were slightly more responsive (USEPA 2006).

Alternatives to the *t*-test approach include testing responsiveness over the entire range of a stressor gradient by examination of scatter plots or by ordination (e.g., Fore et al. 1996, Hering et al. 2006). The *t*-test is one of several alternatives that include the *F*-test (Whittier et al. 2007), its nonparametric equivalent (Mann–Whitney *U*-test; Cao et al. 2007), rank correlation (Böhmer et al. 2004), and discrimination efficiency (Ofenböck et al. 2004). An advantage of the *t*-test is that *t*-scores can be used to rank metrics from responsive to unresponsive and can be interpreted as a measure of effect size. Therefore, the *t*-score is an overall measure of metric responsiveness (i.e., it differentiates superbly responsive metrics from those that are only adequate).

#### *Redundancy*

Our approach selects the single best metric from each of the 6 metric categories we identified for macroinvertebrates. This approach is simple and avoids the problems associated with weighting one category more than another in the final index. Hering et al. (2006) suggest that 3 metrics/category is ideal but do not provide supporting evidence for this assertion. Our experience is that metric redundancy is a limiting step in MMI development. In many cases, it is difficult or impossible to include >1 metric from any given metric category without relaxing the redundancy threshold.

Most investigators agree that redundant metrics should be avoided, but approaches to determining redundancy are varied. Most approaches to MMI development that include some type of redundancy check use correlation analysis (Pearson or rank correlation), but the values used to determine redundancy vary, and often no a priori rule exists for deciding which of 2 redundant metrics should be retained. We calculate the correlation matrix for metric scores at reference sites only. We also establish a criterion (0.71) based on the amount of shared information provided by correlated metrics (>50%) and use a stepwise process to select the final metrics to ensure that the most responsive metric from each category that is not redundant with previously selected metrics is retained in the final MMI. Cao et

al. (2007) advocate using cluster analysis of the metric correlation matrix as an objective means to select nonredundant metrics. In their approach, the most responsive metric from each cluster is chosen for inclusion in the final MMI.

#### *Final remarks*

Rigorous metric evaluation has not always been a common practice in MMI construction (Roset et al. 2007), although EMAP scientists have subscribed to quantitative testing of metrics for some time (Hughes et al. 1998, McCormick et al. 2001, Mebane et al. 2003, Whittier et al. 2007). Developers of fish MMIs have begun to evaluate metrics in an increasingly rigorous manner similar to the approach we have described (Roth et al. 1998, Angermeier et al. 2000, Oberdorff et al. 2002, Pont et al. 2006). MMIs developed with our approach can be used to evaluate and combine results from independent data sets (Meador et al. 2008).

We have tried to stress that the MMI approach used for the WSA is 1 step in the long evolution of the process for developing MMIs for use at larger scales. The process has moved away from the use of BPJ toward use of objective performance criteria. More work remains to be done, especially in the critical areas of correcting for natural gradients and the evaluation of redundancy. We hope that this discussion of our approach to MMI development will stimulate others to improve MMIs in other innovative ways and to go beyond our original goal of conducting a very large-scale bioassessment. At a minimum, we hope that our approach will be adopted by others faced with the challenges of evaluating assemblage data at regional, national, and continental scales.

#### **Acknowledgements**

We thank the many people involved with the EMAP-West and the WSA for their insights and work in collecting the survey data. This paper was improved by comments from J. Van Sickle, T. Johnson, M. Barbour, C. Hawkins, P. Silver, and an anonymous referee. C. Hawkins and J. Van Sickle suggested the *t*-test method to evaluate metric responsiveness. The information in this document has been funded wholly (or in part) by the US EPA, and preparation of this manuscript was partially funded under cooperative agreement CR831682–01 to Oregon State University. It has been subjected to review by the National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of

trade names or commercial products constitute endorsement or recommendation for use.

### Literature Cited

- ANGERMEIER, P. L., R. A. SMOGOR, AND J. R. STAUFFER. 2000. Regional frameworks and candidate metrics for assessing biotic integrity in Mid-Atlantic Highland streams. *Transactions of the American Fisheries Society* 129:962–981.
- BAILEY, R. C., R. H. NORRIS, AND T. B. REYNOLDS. 2004. Bioassessment of freshwater ecosystems: using the reference condition approach. Kluwer Academic Publishers, New York.
- BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Rapid bioassessment protocols for use in streams and Wadeable rivers. EPA 841/B-99/002. Office of Water, US Environmental Protection Agency, Washington, DC.
- 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 (editors). *Biological assessment and criteria: tools for water resource planning and decision making*. Lewis Publishers, Boca Raton, Florida.
- BATES PRINS, S. C., AND E. P. SMITH. 2007. Using biological metrics to score and evaluate sites: a nearest-neighbour reference condition approach. *Freshwater Biology* 52:98–111.
- BLOCKSOM, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. *Environmental Management* 31:670–682.
- BÖHMER, J., C. RAWER-JOST, AND A. ZENKER. 2004. Multimetric assessment of data provided by water managers from Germany: assessment of several different types of stressors with macrozoobenthos communities. *Hydrobiologia* 516:215–228.
- BRAMBLETT, R. G., T. R. JOHNSON, A. V. ZALE, AND D. G. HEGGEM. 2005. Development and evaluation of a fish assemblage index of biotic integrity for northwestern Great Plains streams. *Transactions of the American Fisheries Society* 134:624–640.
- CAO, Y., C. P. HAWKINS, J. OLSON, AND M. A. KOSTERMAN. 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. *Journal of the North American Benthological Society* 26:566–585.
- CARLISLE, D. M., M. R. MEADOR, S. R. MOULTON, AND P. M. RUHL. 2007. Estimation and application of indicator values for common macroinvertebrate genera and families of the United States. *Ecological Indicators* 7: 22–23.
- CHESSMAN, B., S. WILLIAMS, AND C. BESLEY. 2007. Bioassessment of streams with macroinvertebrates: effect of sampled habitat and taxonomic resolution. *Journal of the North American Benthological Society* 26:546–565.
- CUFFNEY, T. E., M. D. BILGER, AND A. M. HAIGLER. 2007. Ambiguous taxa: effects on the characterization and interpretation of invertebrate assemblages. *Journal of the North American Benthological Society* 26:286–307.
- EUROPEAN COMMISSION. 2000. Directive 2000/60/EC of the European Parliament and the Council establishing a framework for the community action in the field of water policy. Directive 2000/60/EC. Official Journal of the European Communities L 327:1–72. (Available from: [http://ec.europa.eu/environment/water/water-framework/index\\_en.html](http://ec.europa.eu/environment/water/water-framework/index_en.html))
- FORE, L. S., AND C. GRAFE. 2002. Using diatoms to assess the biological condition of large rivers in Idaho (U.S.A.). *Freshwater Biology* 47:2015–2037.
- FORE, L. S., J. R. KARR, AND R. W. WISSEMAN. 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. *Journal of the North American Benthological Society* 15:212–231.
- GANASAN, V., AND R. M. HUGHES. 1998. Application of an index of biological integrity (IBI) to fish assemblages of the rivers Khan and Kshipra (Madhya Pradesh), India. *Freshwater Biology* 40:367–383.
- GERTH, W. J., AND A. T. HERLIHY. 2006. Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. *Journal of the North American Benthological Society* 25:501–512.
- HARRIS, J. H., AND R. SILVEIRA. 1999. Large-scale assessments of river health using an index of biotic integrity with low-diversity fish communities. *Freshwater Biology* 41:235–252.
- HERBST, D. B., AND E. L. SILLDORFF. 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. *Journal of the North American Benthological Society* 25:513–530.
- HERING, D., C. K. FELD, O. MOOG, AND T. OFENBÖCK. 2006. Cook book for the development of a multimetric index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR projects and related initiatives. *Hydrobiologia* 566:311–324.
- HERING, D., O. MOOG, L. SANDIN, AND P. F. M. VERDONSCHOT. 2004. Overview and application of the AQEM assessment system. *Hydrobiologia* 516:1–20.
- HERLIHY, A. T., D. P. LARSEN, S. G. PAULSEN, N. S. URQUHART, AND B. J. ROSENBAUM. 2000. Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP Mid-Atlantic Pilot Study. *Environmental Monitoring and Assessment* 63:95–113.
- HERLIHY, A. T., S. G. PAULSEN, J. VAN SICKLE, J. L. STODDARD, C. P. HAWKINS, AND L. L. YUAN. 2008. Striving for consistency in a national assessment: the challenges of applying a reference condition approach at a continental scale. *Journal of the North American Benthological Society* 27:860–877.
- HERRICKS, E. E., AND D. J. SCHAEFFER. 1985. Can we optimize biomonitoring? *Environmental Management* 9:487–492.
- HILSENHOFF, W. L. 1987. An improved biotic index of organic stream pollution. *Great Lakes Entomologist* 20:31–40.
- 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.
- HUGHES, R. M., AND T. OBERDORFF. 1999. Applications of IBI concepts and metrics to waters outside the United States and Canada. Pages 79–93 in T. P. Simon (editor). *Assessing the sustainability and biological integrity of water resources using fish assemblages*. Lewis Publishers, Boca Raton, Florida.
- HUGHES, R. M., AND D. V. PECK. 2008. Acquiring data for large aquatic resources surveys: the art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society* 27:837–859.
- HUGHES, R. M., J. L. STODDARD, AND S. G. PAULSEN. 2000. A national, multi-assemblage, probability survey of ecological integrity. *Hydrobiologia* 422/423:429–443.
- KARR, J. R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6(6):21–27.
- KARR, J. R., AND E. W. CHU. 2000. Sustaining living rivers. *Hydrobiologia* 422/423:1–14.
- KARR, J. R., AND D. R. DUDLEY. 1981. Ecological perspective on water quality goals. *Environmental Management* 5:55–68.
- KAUFMANN, P. R., AND R. M. HUGHES. 2006. Geomorphic and anthropogenic influences on fish and amphibians in Pacific Northwest coastal streams. Pages 429–455 in R. M. Hughes, L. Wang, and P. W. Seelbach (editors). *Landscape influences on stream habitat and biological assemblages*. Symposium 48. American Fisheries Society, Bethesda, Maryland.
- KAUFMANN, P. R., P. LEVINE, E. G. ROBISON, C. SEELIGER, AND D. PECK. 1999. Quantifying physical habitat in Wadeable streams. EPA/620/R-99/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- KLEMM, D. J., K. A. BLOCKSOM, F. A. FULK, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. V. PECK, J. L. STODDARD, AND W. T. THOENY. 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., 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* 1:49–60.
- KURTZ, J. C., L. E. JACKSON, AND W. S. FISHER. 2001. Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development. *Ecological Indicators* 1:49–60.
- LARSEN, D. P., AND A. T. HERLIHY. 1998. The dilemma of sampling streams for macroinvertebrate richness. *Journal of the North American Benthological Society* 17:359–366.
- 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.
- MEADOR, M. R., T. R. WHITTIER, R. M. GOLDSTEIN, R. M. HUGHES, AND D. V. PECK. 2008. Evaluation of an index of biotic integrity approach used to assess biological condition in western U.S. streams. *Transactions of the American Fisheries Society* 137:13–22.
- MEBANE, C. A., T. R. MARET, AND R. M. HUGHES. 2003. An index of biological integrity (IBI) for Pacific Northwest rivers. *Transactions of the American Fisheries Society* 132:239–261.
- MERRITT, R. W., AND K. W. CUMMINS (EDITORS). 1996. *An introduction to the aquatic insects of North America*. 3<sup>rd</sup> edition. Kendall/Hunt, Dubuque, Iowa.
- OBERDORFF, T. D., D. PONT, B. HUGUENY, AND J. P. PORCHER. 2002. Development and validation of a fish-based index (FBI) for the assessment of “river health” in France. *Freshwater Biology* 47:1720–1734.
- ODE, P. R., C. P. HAWKINS, AND R. D. MAZOR. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. *Journal of the North American Benthological Society* 27:967–985.
- OFENBÖCK, T., O. MOOG, J. GERRITSEN, AND M. T. BARBOUR. 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. *Hydrobiologia* 516:251–268.
- OLSEN, A. R., AND D. V. PECK. 2008. Survey design and extent estimates for the Wadeable Stream Assessment. *Journal of the North American Benthological Society* 27:822–836.
- OMERNIK, J. M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77:118–125.
- PAULSEN, S. G., A. MAYO, D. V. PECK, J. L. STODDARD, E. TARQUINIO, S. M. HOLDSWORTH, J. VAN SICKLE, L. L. YUAN, C. P. HAWKINS, A. T. HERLIHY, P. R. KAUFMANN, M. T. BARBOUR, D. P. LARSEN, AND A. R. OLSEN. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. *Journal of the North American Benthological Society* 27:812–821.
- PECK, D. V., A. T. HERLIHY, B. H. HILL, R. M. HUGHES, P. R. KAUFMANN, D. J. KLEMM, J. M. LAZORCHAK, F. H. MCCORMICK, S. A. PETERSON, P. L. RINGOLD, T. MAGEE, AND M. R. CAPPAERT. 2006. *Environmental Monitoring and Assessment Program—Surface Waters Western Pilot Study: field operations manual for Wadeable streams*. EPA 600/R-06/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- PONT, D., R. M. HUGHES, T. R. WHITTIER, AND S. SCHMUTZ. A predictive IBI model for fish assemblages of western USA streams. *Transactions of the American Fisheries Society* (in press).
- PONT, D., B. HUGUENY, U. BEIER, D. GOFFAUX, A. MELCHER, R. NOBLE, C. M. ROGERS, N. G. ROSET, AND S. SCHMUTZ. 2006. Assessing river biotic condition at the continental scale: a European approach using functional metrics and fish assemblages. *Journal of Applied Ecology* 43:70–80.
- RANKIN, E. T., AND C. O. YODER. 1999. Methods for deriving maximum species richness lines and other threshold relationships in biological field data. Pages 611–624 in T. P. Simon (editor). *Assessing the sustainability and*

- biological integrity of water resources using fish communities. CRC Press, Boca Raton, Florida.
- REHN, A. C., P. R. ODE, AND C. P. HAWKINS. 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. *Journal of the North American Benthological Society* 26:332–348.
- ROSET, N., G. GRENOUILLET, D. GOFFAUX, D. PONT, AND P. KESTEMONT. 2007. A review of existing fish assemblage indicators and methodologies. *Fisheries Management and Ecology* 14:393–405.
- ROTH, N. E., M. T. SOUTHERLAND, J. C. CHAILLOU, R. J. KLAUDA, P. F. KAZYAK, S. A. STRANKO, S. B. WEISBERG, L. W. HALL, AND R. P. MORGAN. 1998. Maryland biological stream survey: development of a fish index of biotic integrity. *Environmental Monitoring and Assessment* 51:89–106.
- SIMON, T. P., AND J. LYONS. 1995. Application of the index of biotic integrity to evaluate water resource integrity in freshwater ecosystems. Pages 245–262 *in* W. S. Davis and T. P. Simon (editors). *Biological assessment and criteria: tools for water resource planning and decision making*. Lewis Publishers, Chelsea, Michigan.
- SNOOK, H., S. P. DAVIES, J. GERRITSEN, B. K. JESSUP, R. LANGDON, D. NEILS, AND E. PIZZUTO. 2007. The New England Wadeable Stream Survey (NEWS): development of common assessments in the framework of the biological condition gradient. New England Regional Laboratory, US Environmental Protection Agency, Chelmsford, Massachusetts. (Available from: [http://www.epa.gov/NE/lab/pdfs/NEWSfinalReport\\_August2007.pdf](http://www.epa.gov/NE/lab/pdfs/NEWSfinalReport_August2007.pdf))
- SOUTHERLAND, M. T., G. M. ROGERS, M. J. KLINE, R. P. MORGAN, P. F. KAZYAK, R. J. KLAUDA, AND S. A. STRANKO. 2007. Improving biological indicators to better assess the condition of streams. *Ecological Indicators* 7:751–767.
- STODDARD, J. L., D. P. LARSEN, C. P. HAWKINS, R. K. JOHNSON, AND R. H. NORRIS. 2006. Setting expectations for the ecological condition of running waters: the concept of reference condition. *Ecological Applications* 16:1267–1276.
- STODDARD, J. L., D. V. PECK, A. R. OLSEN, D. P. LARSEN, J. VAN SICKLE, C. P. HAWKINS, R. M. HUGHES, T. R. WHITTIER, G. LOMNICKY, A. T. HERLIHY, P. R. KAUFMANN, S. A. PETERSON, P. L. RINGOLD, S. G. PAULSEN, AND R. BLAIR. 2005a. Environmental Monitoring and Assessment Program (EMAP): western streams and rivers statistical summary. EPA 620/R-05/006. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- STODDARD, J. L., D. V. PECK, S. G. PAULSEN, J. VAN SICKLE, C. P. HAWKINS, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. P. LARSEN, G. LOMNICKY, A. R. OLSEN, S. A. PETERSON, P. L. RINGOLD, AND T. R. WHITTIER. 2005b. An ecological assessment of western streams and rivers. EPA 620/R-05/005. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2006. Wadeable Streams Assessment: a collaborative survey of the Nation's streams. EPA 841/B-06/002. Office of Water, US Environmental Protection Agency, Washington, DC.
- VINSON, M. R., AND C. P. HAWKINS. 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among streams. *Journal of the North American Benthological Society* 15:392–399.
- WAITE, I. R., A. T. HERLIHY, D. P. LARSEN, N. S. URQUHART, AND D. J. KLEMM. 2004. The effects of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, U.S.A. *Freshwater Biology* 49:474–489.
- WHITTIER, T. R., R. M. HUGHES, J. L. STODDARD, G. A. LOMNICKY, D. V. PECK, AND A. T. HERLIHY. 2007. A structured approach for developing indices of biotic integrity: three examples from streams and rivers in the western USA. *Transactions of the American Fisheries Society* 136:718–735.

*Received: 25 March 2008*  
*Accepted: 29 August 2008*