JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 44, No. 3 AMERICAN WATER RESOURCES ASSOCIATION June 2008

A COMPARISON OF SINGLE AND MULTIPLE HABITAT PROTOCOLS FOR COLLECTING MACROINVERTEBRATES IN WADEABLE STREAMS[1]

Karen A. Blocksom, Bradley C. Autrey, Margaret Passmore, and Lou Reynolds[2]

ABSTRACT: In 2003, we compared two benthic macroinvertebrate sampling methods that are used for rapid biological assessment of wadeable streams. A single habitat method using kick sampling in riffles and runs was compared to a multiple habitat method that sampled all available habitats in proportion of occurrence. Both methods were performed side-by-side at 41 sites in lower gradient streams of the Piedmont and Northern Piedmont ecoregions of the United States, where riffle habitat is less abundant. Differences in sampling methods were examined using similarity indices, two multimetric indices [the family-level Virginia Stream Condition Index (VSCI) and the species-level Macroinvertebrate Biotic Integrity Index (MBII)], their component metrics, and bioassessment endpoints based on each index. Index scores were highly correlated between single and multiple habitat field methods, and sampling method comparability, based on comparison of similarities between and within sampling methods, was particularly high for species level data. The VSCI scores and values of most of its component metrics were not significantly higher for one particular method, but relationships between single and multiple habitat values were highly variable for percent Ephemeroptera, percent chironomids, and percent Plecoptera and Trichoptera (Hydropsychidae excluded). A similar level of variability in the relationship was observed for the MBII and most of its metrics, but Ephemeroptera richness, percent individuals in the dominant five taxa, and Hilsenhoff Biotic Index scores all exhibited differences in values between single and multiple habitat field methods. When applied to multiple habitat samples, the MBII exhibited greater precision, higher index scores, and higher assessment categories than when applied to single habitat samples at the same sites. In streams with limited or no riffle habitats, the multiple habitat method should provide an adequate sample for biological assessment, and at sites with abundant riffle habitat, little difference would be expected between the single and multiple habitat field methods. Thus, in geographic areas with a wide variety of stream types, the multiple habitat method may be more desirable. Even so, the variability in the relationship between single and multiple habitat methods indicates that the data are not interchangeable, and we suggest that any change in sampling method should be accompanied by a recalibration of any existing assessment tool (e.g., multimetric index) with data collected using the new method, regardless of taxonomic level.

(KEY TERMS: method comparability; environmental indicators; assemblage sampling; rapid bioassessment protocols; invertebrates; low gradient streams.)

Blocksom, Karen A., Bradley C. Autrey, Margaret Passmore, and Lou Reynolds, 2008. A Comparison of Single and Multiple Habitat Protocols for Collecting Macroinvertebrates in Wadeable Streams. *Journal of the American Water Resources Association* (JAWRA) 44(3):577-593. DOI: 10.1111/j.1752-1688.2008.00183.x

[1] Paper No. J06094 of the *Journal of the American Water Resources Association* (JAWRA). Received July 13, 2006; accepted September 13, 2007. © 2008 American Water Resources Association. **Discussions are open until December 1, 2008**.

[2] Respectively, Statistician, National Exposure Research Laboratory, U.S. Environmental Protection Agency, 26 W. Martin Luther King Drive, Cincinnati, Ohio 45268; Research Biologist, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, Ohio 45268; Environmental Scientist, Region 3, U.S. Environmental Protection Agency, Wheeling, West Virginia 26003; and Biologist, Region 3, U.S. Environmental Protection Agency, Wheeling, West Virginia 26003 (E-Mail/Blocksom: blocksom.karen@epa.gov).

INTRODUCTION

In the United States (U.S.), each State is required to report on the condition of waters within its boundaries to the U.S. Environmental Protection Agency (USEPA) on a biannual basis pursuant to section 305(b) of the Clean Water Act (CWA) (USEPA, 2005). For each water body, the State must designate beneficial uses (aquatic life, recreation, drinking water, industrial, etc.), as well as numeric and narrative criteria to protect those uses in their State water quality standards. The status of water bodies with respect to these designated uses (i.e., attainment or nonattainment) must then be evaluated. In compliance with section 303(d) of the CWA, States must also provide a listing of water bodies where the designated uses are impaired or threatened and where possible, determine what pollutant or stressor is causing the impairment.

Biological assessment (bioassessment) of resident organisms is typically used for evaluating attainment of the aquatic life use (ALU) of streams and rivers. Bioassessment is conducted by obtaining a representative sample of resident aquatic life at a test site, comparing that sample to what is expected at reference sites where attainment of ALU is achieved, and making a determination of use support or impairment based on that comparison. If a State determines the ALU is impaired, the biological, habitat, and water quality data available at the site are further scrutinized to determine the cause of impairment. Often, further sampling is required to identify the source of impairment.

In an effort to provide tools to perform bioassessments, USEPA led a work group of State and USEPA Regional biologists in the 1980s and developed the rapid bioassessment protocols (RBPs) for sampling and assessing wadeable streams (Barbour *et al.*, 1999). The first edition of the RBP document included protocols for sampling fish and benthic macroinvertebrate assemblages, as well as for habitat and physicochemical parameters (Plafkin *et al.*, 1989). The RBPs for macroinvertebrates included three levels of rigor in sampling, with the two more rigorous of these approaches focused on sampling riffle/run habitats because these are typically the most productive habitats available in streams with the highest macroinvertebrate diversity and abundance (Plafkin *et al.*, 1989). Many State agencies adopted or adapted these protocols for sampling in their own streams as they developed biomonitoring and bioassessment programs, thus allowing the protocols themselves to be tested widely across the U.S. (Barbour *et al.*, 1999). During this time, it became evident that a single habitat approach to sampling would be problematic in regions having lower-gradient streams with limited riffle/run habitats and more sandy or silty substrates. As a result, an approach was introduced in the second edition of the RBPs that sampled multiple habitats rather than a single habitat (Barbour *et al.*, 1999). The intention of introducing the multiple habitat approach was to allow States to collect a more representative sample across all stream types by sampling available habitats in proportion to their abundance at a site.

In higher gradient streams, the predominant habitat tends to be riffles. As a result, a sample that is collected using the single habitat method that focuses on riffle/run habitats should be very similar to the sample that is collected using the multiple habitat method, in which habitats are sampled according to their proportional occurrence in the stream. However, in low gradient streams, which tend to have more pool/glide habitats, the two field methods could produce vastly different results.

The second RBP document provided no guidance or recommendations for how baseline data collected using the single habitat method might be used in conjunction with data collected using the multiple habitat method. This issue is important because some States developed indicators (e.g., mulitmetric indices) for bioassessment using data collected with the single habitat field method but acknowledge that the multiple habitat approach may obtain more representative samples from some lower gradient streams. States questioned whether the new multiple habitat field method resulted in a more effective assessment for low gradient streams, meaning that it would result in a more accurate determination of use support. State agency personnel wondered whether they could continue to use the single habitat field method at high gradient sites, but also adopt the new multiple habitat field method in streams that lacked riffle habitat. If the States used both field methods, could the multiple habitat data be incorporated using the old assessment tools (e.g., an index developed from the data collected using the single habitat field method)? In this case, before changing from the single habitat to the multiple habitat field method, or adopting the multiple habitat field method at a subset of sites, the State must determine whether it can use the existing indicator (e.g., by converting index scores with an algorithm) or if it must develop a brand new indicator from data collected solely with the new field method.

Several studies have examined the issue of comparability of data collected with different methods from varying perspectives. Both Barbour *et al.* (1999) and Diamond *et al.* (1996) described a framework using a performance-based methods system (PBMS) to compare bioassessment methods based on the quality of the data collected with each method. This approach assumes that if the performance characteristics (i.e., precision, bias, sensitivity, performance range, and interferences) are similar among different methods or programs, then the methods themselves will produce comparable data that can be combined into a single dataset (Flotemersch *et al.*, 2006a). However, this is typically only feasible at some higher level of data organization, such as metric or assessment level. As an example, Houston *et al.* (2002) described a comparison of methods among five States in the southeastern U.S. using a strictly PBMS approach. Southerland *et al.* (2006) followed a similar approach in comparing assessment endpoints for Maryland, Virginia, and West Virginia but actually compared probabilistic data among methods in addition to characteristics of the sampling programs. For these two studies, the primary goal was to determine whether, and at what organizational level, bioassessment data collected by different states could be integrated to produce a regional assessment of stream condition. In both cases, the authors concluded that some integration of assessment level data was possible, despite differences in methods and data at lower levels of organization (e.g., taxonomic abundances, metric values). Other studies focusing more on direct comparisons of methods in the field have had similar aims of integrating data across studies (Cao *et al.*, 2005; Herbst and Silldorff, 2006) or of determining the effect of sampling method on data at various levels (Gerth and Herlihy, 2006; Wang *et al.*, 2006). Only Ostermiller and Hawkins (2004) actually conducted a side-by-side field comparison of a single habitat, riffle-based sampling approach to one that sampled available habitats in proportion to their availability. However, the authors of that study focused on the effects of sampling error on River Invertebrate Prediction and Classification System (RIVPACS)-type predictive model bioassessments (Clarke *et al.*, 2003), and differences between the two sampling methods were only a small part of the study. In our study, we focused specifically on differences between the data collected using the RBP single and multiple habitat field methods for the purpose of providing recommendations for States that are considering changing from the single to the multiple habitat approach, or are interested in adopting the multiple habitat approach for a subset of low gradient streams in the State.

In this study, we carried out a side-by-side comparison of the RBP single and multiple habitat methods for macroinvertebrates in lower gradient streams of the Piedmont and Northern Piedmont Ecoregions of the U.S. (Omernik, 1995), where riffle habitat and cobble substrates can be limited. Two indices that are relevant to the majority of the geographic area encompassed in this study are a family-level index developed specifically for Virginia and a species-level index originally developed for the Mid-Atlantic Highlands region using data from the USEPA's Environmental Monitoring and Assessment Program (EMAP) Mid-Atlantic Highlands Assessment (MAHA). Virginia has evaluated the family-level Stream Condition Index using an independent dataset derived from a State probabilistic monitoring network, and is working toward adopting this index for determining ALU support. Both indices were generated from macroinvertebrate data collected primarily from riffle habitats in streams. The purpose of examining both a family-level and a species-level index was not to directly compare the results at the two taxonomic levels. Rather, it was to provide results at two relevant levels of taxonomic data. While identification of samples to the species or lowest possible taxonomic level may be the ideal, some States currently identify some or all macroinvertebrate taxa to only the family level or have past data recorded at this level (USEPA, 2002).

Comparisons among methods can be made at several levels of data organization: taxonomic composition (relative abundances), metrics, indices, and bioassessment endpoints (e.g., good-fair-poor or impaired-unimpaired). For the purposes of evaluating whether or not streams meet their designated uses for CWA section 305(b), the bioassessment endpoint of attainment/nonattainment of ALU may be the most important level of data for method comparison. If this determination is consistent across methods, other differences could be considered unimportant. However, if a state agency places a stream on the 303(d) list of impaired waters, that agency must be able to determine cause of impairment. To investigate the possible causes of impairment, state biologists often consider the raw data (the taxa lists and counts), as well as associated water quality and physical habitat data of the site. For this reason, it is also important to understand how the choice of field method impacts the taxa lists and the relationship of the macroinvertebrate data to stressor/human disturbance gradients (diagnostic capability or sensitivity). Thus, we must examine the data at finer levels of organization beyond a simple pass/fail designation for ALU support. For multimetric indices, which are commonly used for bioassessment, the underlying metrics comprise this finer level. In this case, the taxonomic composition data can provide further information on the basis of observed differences in metric and index values. For predictive (i.e., RIVPACS-type) models, the taxonomic composition data directly represent this finer level of data organization. If data are similar at the lowest level of organization (i.e., relative abundances), higher levels of data organization should also yield similar results. However, although taxonomic composition may differ strongly between samples from two different methods, those differences may not translate into differences at the metric or index level. Depending on how taxonomic composition differs between the field sampling methods, assessments based on multimetric indices may or may not be affected by such differences.

Our overall goal for this study was to provide information that could be used by States to determine under what conditions one can interchange or combine data obtained by the single and multiple habitat field methods to make ALU assessments. Specifically, we wanted to know how data collected using the two field methods would differ at the level of taxa presence and associated relative abundances, individual metrics, and assessment indicators like multimetric indices that are currently in use by States to determine ALU support. It is important to look at the differences between methods at all levels because State biologists consider all levels of information. Thus, we began our comparison at the level of taxonomic composition data and progressed to successively higher levels of organization, including metrics, indices, and bioassessment endpoints. For metrics and indices, we compared not only values but also variability and stressor relationships, because differences in these characteristics influence comparability of methods. Although there are statistical tests for the significance of the difference between the two methods, no criteria exist to determine if this difference will matter at the level of program-specific goals of the user (Diamond *et al.*, 1996). Thus, at each step, we evaluated the degree and types of differences between the single and multiple habitat field sampling methods.

FIELD METHODS

Study Area and Sample Collection

From 1993 through 1996 and from 1997 through 1998, the USEPA's EMAP conducted the MAHA and Mid-Atlantic Integrated Assessment (MAIA), respectively. These assessments sampled a total of 868 randomly chosen sites in the Mid-Atlantic Region of the USEPA's Region 3. Fifty-four of those sites were located in the Piedmont and Northern Piedmont Ecoregions (Omernik, 1995) and were considered as potential sites for this study. Landowners granted permission for site access at 41 of those sites, and we sampled them between April 1 and May 7, 2003. Watershed size at sampled sites ranged from 1.4 to 107 km^2 (based on EMAP data, available at http://www.epa.gov/emap/html/data/surfwatr/data/index.html) (Figure 1).

We defined each site as a 100-m reach of stream within which all sampling was confined and identified its location using latitude and longitude coordinates to mark the midpoint of the reach. When necessary, we shifted the entire reach so that it was at least 100 m upstream from any road, bridge crossing, or major tributary. From each site, we collected single and multiple habitat macroinvertebrate samples, habitat information, *in situ* water chemistry data, and a water sample to be analyzed later for nutrient concentrations.

Macroinvertebrates. We applied the two RBP macroinvertebrate sampling methods during the same visit in the same reach so as to directly compare field methods side-by-side. For both methods, sampling began at the downstream end of the reach and proceeded upstream. For the single habitat method, we performed four kicks at various velocities in the riffle portion of the reach (or the fastest flowing water if riffles were not present) using a 0.5-m kick net with 595-μm mesh. Each kick consisted of positioning the net and using the toe or heel of the boot to disturb the upper layer of substrate and scrape the underlying bed over an area of 0.25 m^2 upstream of the net. We picked up larger substrate particles and rubbed them by hand to remove any attached organisms. We then composited the material collected from the four kicks into a single sample and rinsed it with stream water. This method is slightly modified from the original RBP single habitat method in that it samples a total of 1 m^2, rather than 2 m^2 as suggested in the first edition of the RBP (Plafkin *et al.*, 1989). This reflects the current approach used by USEPA Region 3 biologists performing stream assessments.

For the multiple habitat method, we sampled habitat types in proportion to their relative surface area within the sampling reach. We performed a total of 20 jabs and/or kicks over the length of the reach using a 0.3-m wide D-frame net with 595-μm mesh. Each jab consisted of forcefully thrusting the net into a particular habitat for a linear distance of 0.5 m. Each kick consisted of positioning the net and using the toe or heel of the boot to disturb the upper layer of substrate and scrape the underlying bed in a 0.25 m^2 area upstream of the net. The categories of habitat types sampled included cobble, snags, vegetated banks, submerged macrophytes, and sand. We composited the material collected during the 20 jabs and/or kicks into a single sample and rinsed it with stream water. Again, we removed large debris and inspected it for organisms. The total estimated area sampled for the multiple habitat method was approximately 3 m^2. All macroinvertebrate samples were preserved with 95% ethanol.

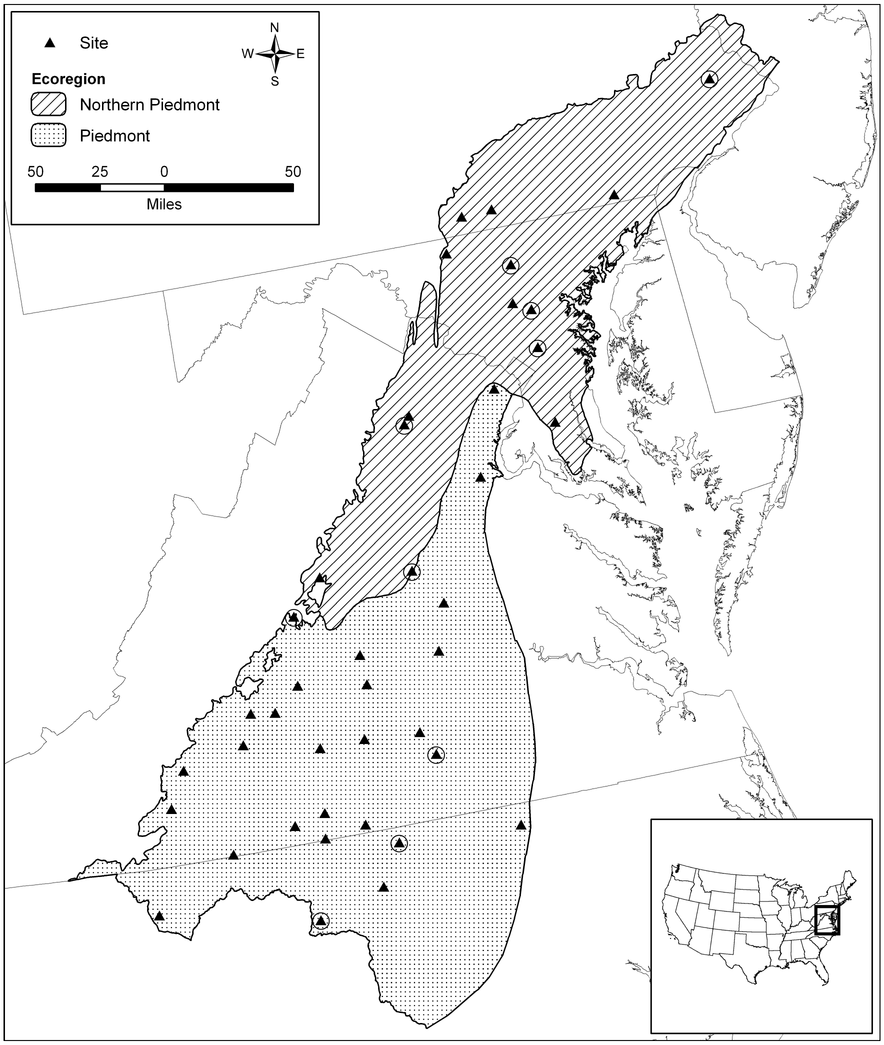

FIGURE 1. Locations of Sites Sampled for This Study. Circled sites were randomly chosen for collection of duplicate samples.

To evaluate the variability of each method, we collected macroinvertebrate samples in duplicate for both methods at 10 randomly chosen sites (Figure 1). We took the duplicate samples from the same reach where the original samples were taken for most sites, although we shifted the reach for a few sites to allow adequate sampling area. We collected duplicate samples on the same date as the original samples.

In the laboratory, we placed material from macroinvertebrate samples into gridded sorting pans, then randomly selected a grid square and sorted all macroinvertebrates within the square from debris. We randomly selected and completely sorted additional grid squares until the total number of organisms sorted was at least 270 organisms (within 10% of the target of 300 organisms). Taxonomists identified all organisms to the lowest possible taxon, depending on the condition and life stage (instar) of the specimen, as well as the availability of taxonomic keys. Keys for basic initial identification included Brigham *et al.* (1982), Merritt and Cummins (1996), Peckarsky *et al.* (1990), and Pennak (1989). Numerous additional keys were used for updated information and to identify specific taxonomic groups.

Habitat. At each site, we collected habitat data using the RBP habitat (RH) assessment approach (Barbour *et al.*, 1999). RH scores are based on visual assessment of 10 habitat parameters, with scores ranging from 0 for poor condition to 20 for optimal condition for each component (Table 1). Evaluation and scoring of each component is based on comparison to descriptions provided for four condition categories (optimal, suboptimal, marginal, poor) that each encompass a range of scores. We evaluated each site as either high or low gradient based on whether riffle/run habitat (high gradient) or glide/pool habitat (low gradient) predominated. For most streams, we

TABLE 1. Abbreviations For the Rapid Bioassessment Protocols Habitat Assessment (RH) Approach (Barbour *et al.*, 1999), With High and Low Gradient Variations on Component Parameters.

Abbreviation	High Gradient	Low Gradient
EPISUB (20)	Epifaunal substrate/available cover	Epifaunal substrate/available cover
EMBED (20)	Embeddedness	Pool substrate characterization
VELDEP (20)	Velocity/depth combinations	Pool variability
SEDDEP (20)	Sediment deposition	Sediment deposition
CHANFLW (20)	Channel flow status	Channel flow status
CHANALT (20)	Channel alteration	Channel alteration
FREQBND (20)	Frequency of riffles or bends	Channel sinuosity
BKSTAB (10/bank)	Bank stability	Bank stability
BKVEG (10/bank)	Bank vegetative protection	Bank vegetative protection
RIPVEG (10/bank)	Riparian vegetative zone width	Riparian vegetative zone width

Note: The maximum score for each component, indicating the most optimal condition, is provided in parentheses.

determined only scores for either the high gradient or low gradient RH parameters. However, for streams that were not clearly high gradient or low gradient, we determined scores for both types of streams. In addition to the RH scores, we measured depth, wetted width, and bankfull width using a measuring rod at the upstream and downstream ends and in the middle of the reach. We measured depth in the deepest part of the channel at all three points. We measured bankfull width as the width of the channel at bankfull stage, and wetted width as the actual width of the stream channel that was wet during sampling.

Water Chemistry. We collected data *in situ* for stream water conductivity, temperature, pH, and dissolved oxygen (DO) using either a YSI 85 m or a Corning Checkmate meter. We collected water samples and analyzed them in the laboratory for total phosphorus (P), nitrate plus nitrite nitrogen (NO_x-N), ammonia nitrogen (NH_3-N), and total Kjeldahl nitrogen (TKN). We collected both water samples and *in situ* readings near the middle of the stream near the longitudinal center of the reach. To collect the water sample, we pointed the capped ends of two clean, acid-washed 125-ml bottles in the downstream direction and submerged them to approximately one-half the channel depth. Then we uncapped the bottles, allowed them to fill completely with stream water, and recapped them while submerged. Within 12 hours of obtaining the samples from the stream, we filtered approximately 125 ml of one bottle through a sterile filter using a hand pump. We rinsed the original sample bottle with deionized water and replaced the filtered water in the bottle. Then we added 0.25 ml of H_2SO_4 directly to each of the bottles, recapped them immediately, and chilled them on ice or in a portable refrigeration unit. We stored samples at or below 4°C for no more than 28 days before analyzing them for nutrients using a Segmented Flow Analyzer.

Land Cover. We obtained land cover and land use information for each site from available EMAP-MAHA and EMAP-MAIA data (http://www.epa.gov/emap/html/data/surfwatr/data/index.html). These included site elevation, slope, and watershed area, in addition to road density, population density, and percentages of forest, agriculture, and urban land cover in the watershed of each site. The watershed for each stream site was determined based on the latitude and longitude of the sampling point and Digital Elevation Model data. Elevation and slope were calculated by overlaying the watershed area and the National Elevation Database, available through the U.S. Geological Survey Center for Earth Resources Observation and Science (http://eros.usgs.gov/index.html). Land cover data were derived from leaves-on and leaves-off Landsat satellite thematic mapper scenes acquired from 1991 to 1993, which were projected to Lambert Azimuthal coordinates (Vogelmann *et al.*, 1998). The 30 m^2 picture elements (pixels) of the Landsat scenes were clustered into 100 spectrally distinct classes using an unsupervised clustering algorithm (Kelly and White, 1993). Aerial photographs and supplemental data were used to interpret the classes and assign them one of 15 land cover category labels (Kelly and White, 1993).

DATA ANALYSIS

We focused the statistical analysis on identifying differences in taxonomic data, metric values, index scores, and bioassessment endpoints produced by the two sampling methods when applied to the same set of sites. In addition, we evaluated differences in variability and in relationships with variables associated with human disturbance (stressors). We analyzed the data at the lowest taxonomic level, henceforth referred to as *species* level, and also aggregated and analyzed data at the family level. For all analyses except estimates of similarity and variability, we only included one sample collected with each method from each site. Thus, at sites with duplicate samples, only the first sample collected with each method was used.

Taxonomic Similarity

Cao *et al.* (2005) defined sampling-method comparability (SMC) as a measure of how similarly two methods characterize a biological assemblage. The measure used, referred to as classification strength (CS) SMC in Cao *et al.* (2005), was based on CS as described in Van Sickle (1997), and was calculated from similarities between and within methods. This SMC for an individual site was calculated as

$$\mathrm{SMC} = \frac{2S_\mathrm{b}}{S_\mathrm{ws} + S_\mathrm{wm}} \times 100\%, \tag{1}$$

where S_b is the similarity between methods, and S_ws and S_wm are the similarities between duplicates within the single and multiple habitat field methods, respectively. Following Cao *et al.* (2005), we estimated similarities between methods and between duplicates within a method using two common similarity indices, the Jaccard coefficient (SJ), based on presence-absence data, and the Bray-Curtis (BC) similarity index, an extension of Sorensen similarity based on taxonomic relative abundances (McCune and Grace, 2002). Both similarity indices were calculated according to McCune and Grace (2002). To standardize the level of taxonomy among samples, we set the operational taxonomic unit for each group of taxa by examining the data and determining the taxonomic level at which the smallest amount of information was lost for that group (Ostermiller and Hawkins, 2004; Flotemersch *et al.*, 2006a). This often meant consolidating data to a higher taxonomic level or dropping observations identified to a higher taxonomic level. To reduce the effect of dominant taxa, abundances were transformed using ln(x + 1) for the BC index (van Tongeren, 1995). For SMC estimates, we used only sites with duplicate samples so that similarities both within and between methods could be calculated. From each of these sites, we calculated similarities between methods using the first sample collected for each method. The SMC measures similarity between methods relative to that between duplicates within a method, rather than using just the similarity between sampling methods to evaluate how closely two methods characterize a given assemblage. In addition to the SMC, we were interested in directly comparing within-method similarities (based on duplicate samples) between the two methods. Thus, we performed a nonparametric paired *t*-test, the Wilcoxon signed rank test (Hollander and Wolfe, 1999), on each index at the family and species levels. This test essentially compares similarities for the two field methods at each site and then tests whether there is an overall trend for the similarities within one method to be higher.

Indices and Metrics

We applied two multimetric biotic indices developed for streams to data from both sampling methods and all sites. First, we calculated a family-level macroinvertebrate multimetric biotic index that was developed for the State of Virginia (unpublished report by J. Burton and J. Gerritsen, TetraTech, Inc., Owings Mills, Maryland, 2003). The VSCI was developed based on data collected using the single habitat method throughout the noncoastal areas of Virginia. The eight component metrics of the index are as follows: (1) total taxa richness, (2) Ephemeroptera, Plecoptera, and Trichoptera (EPT) taxa richness, (3) percent of individuals in Ephemeroptera, (4) percent of individuals in Plecoptera and Trichoptera, excluding Hydropsychidae, (5) percent scrapers, (6) percent individuals in Chironomidae, (7) percent individuals in the dominant two taxa, and (8) the family-level Hilsenhoff Biotic Index (FBI) (Hilsenhoff, 1987). The FBI is essentially a weighted average tolerance value (TV), based on family-level taxonomic abundance in a sample (Hilsenhoff, 1987). The TVs used in the FBI are taxon-specific and describe the tendency of a particular taxon to occur along a generalized human disturbance gradient, with values ranging from 0 for no tolerance to 10 for high tolerance to disturbance. We used the TVs in the Virginia Department of Environmental Quality (VDEQ) database for calculating the FBI. We based functional feeding group designations at the family level on the VDEQ database used to develop the VSCI. Where these were unavailable, we used information from the RBP document (Barbour *et al.*, 1999) or the EMAP-MAHA database (http://www.epa.gov/emap/html/data/surfwatr/data/mastreams/9396/index.html). A subsample size of approximately 150 organisms was typical of the samples used to develop the VSCI. Therefore, we performed rarefaction (Hurlbert, 1971) to a subsample size of 150 organisms to estimate taxa richness metrics for the VSCI. Scoring or standardization of each metric was based on a threshold derived from the 95th percentile of the distribution of all sites in the Virginia dataset. Scores were transformed to the range 0-100, and the VSCI score is the average of the eight metric scores.

The species level multimetric index for macroinvertebrates used in these analyses, the Macroinvertebrate Biotic Integrity Index (MBII), was developed for the entire Mid-Atlantic Highlands region (including the Piedmont) using lowest possible taxon data from the USEPA's EMAP-MAHA study (Klemm *et al.*, 2003). Data used to develop the MBII were based on single habitat sampling, although the actual field method differed from the RBP single habitat approach. We followed the same procedure for laboratory processing of samples used in the EMAP-MAHA study. Thus, no rarefaction was required. The MBII consists of seven metrics: (1) Ephemeroptera taxa richness, (2) Plecoptera taxa richness, (3) Trichoptera taxa richness, (4) collector-filterer taxa richness, (5) percent noninsect individuals, (6) percent dominant five taxa, and (7) the Hilsenhoff Biotic Index (HBI). We used the TVs associated with the EMAP-MAHA database to calculate the HBI (http://www.epa.gov/emap/html/data/surfwatr/data/mastreams/9396/index.html). Scoring for each metric was based on scaling between two thresholds derived from the 75th percentile of least disturbed sites and 25th percentile of impaired sites. Calculations for Ephemeroptera richness, Plecoptera richness, and collector-filterer taxa richness metrics included an adjustment for watershed area prior to scoring using a simple linear regression (Klemm *et al.*, 2003). The MBII score is calculated as the sum of metric scores multiplied by (100/7) to convert the total to a 100-point scale.

After calculating the two indices and their component metrics, we compared field methods in several ways. We were interested in not only a direct comparison of metric and index values but also in differences in variability and response to potential stressors between methods.

Direct Comparisons. We performed a paired *t*-test to identify significant differences between methods for each index and set of metrics. In addition, we examined bivariate scatter plots of metric and index values for the single habitat method against the multiple habitat method to assess variability qualitatively in the relationship between values for the two methods. For both the VSCI and the MBII, we regressed single habitat index scores on those for the multiple habitat field method to determine the ability of multiple habitat values to predict single habitat values, then used 90% prediction intervals as a measure of the uncertainty associated with predictions. Residuals were then examined to ensure approximate normality and homoscedasticity. Although this regression approach does not take into account variation in the predictor variable (multiple habitat MBII scores) due to measurement or sampling error, it does provide a way to gauge variation in the ability of multiple habitat scores to predict single habitat scores.

Variability. To evaluate differences in variability between field methods, we calculated the root mean square error (RMSE) and the signal-to-noise ratio (S/N) for the two multimetric indices. The RMSE is an estimate of the measurement error associated with a method, and the S/N, a measure of precision, is a comparison of the variance among sites (signal) relative to variance within sites (noise) (Kaufmann *et al.*, 1999). For both analyses, we used only sites with duplicate data.

To obtain the RMSE, we performed a generalized linear model (GLM) with index value as the response and sampling site as a random factor. From this analysis, the square root of the mean square error is the RMSE. Larger values of the RMSE indicate higher measurement error within a method.

From the same GLM model, we calculated the S/N for VSCI and MBII scores according to Kaufmann *et al.* (1999) as

$$\mathrm{S/N} = \frac{\sigma^2_{\mathrm{site}}}{\sigma^2_{\mathrm{rep}}}, \tag{2}$$

where σ^2_{site} is the variance among sites and σ^2_{rep} is the variance among replicates (within method). Using the mean squares table from the GLM output, Equation (2) reduces to

$$\mathrm{S/N} = (F-1)/c, \tag{3}$$

where F is the F-statistic for sampling site and c is a constant representing the number of samples per site for a given method (i.e., $c = 2$ for this study) (Kaufmann *et al.*, 1999). The residuals from the models were examined to ensure normality and homoscedasticity.

Relationships to Potential Stressors. Within each method, we examined the relationships of metrics and indices to abiotic variables that could be considered to represent stressors. As both instream and near-stream characteristics, including water chemistry and habitat variables, and watershed characteristics, including land cover and land use variables, can represent sources of stress to stream macroinvertebrates, these were all considered as potential stressors in our streams. We first carried out a principal component analysis (PCA) on habitat, water quality, and land use variables, then used the resulting axes to represent the disturbance gradient. The abiotic variables included the RBP habitat metrics, *in situ* measurements, water chemistry, and land cover measurements. Of the 40 sites with nutrient data, 34 had total phosphorus (TP) concentrations that were below the detection limit of 0.05 mg/l, and the detection limit was used for these observations. A rule of thumb for normality of variables used in parametric multivariate analyses is to achieve $|\text{skew}| < 1$ (McCune and Grace, 2002), but some variables were so skewed that no transformation could achieve this requirement and were excluded (i.e., TP and NH_3). Others were transformed using natural log or square root, or by squaring values. We dropped DO and temperature, as these could vary by time of day of the sample, and we excluded percent urban land cover, population density, RH sediment deposition, and RH frequency of riffles or bends because of high correlations ($|r| > 0.75$) with other variables. From the resulting PCA, we kept and interpreted only those axes with eigenvalues (λ) larger than their corresponding broken-stick eigenvalues (McCune and Grace, 2002), as provided in PC-ORD (v. 4.25, MjM Software, Gleneden Beach, Oregon). For each macroinvertebrate sampling method, we ran Spearman rank correlations of the two index scores and component metrics with the PCA axes and examined similarities between methods qualitatively. Because of the exploratory nature of these analyses, we did not assess significance of individual correlations (Van Sickle, 2003).

Bioassessment Endpoints

The effect of sampling method on the bioassessment endpoint is at least as important as numerical changes in index scores or metrics. The VSCI has a tentative impairment threshold of 60, based on the distribution of scores among reference (least disturbed) sites. Scores falling below this value would be considered "failing", or in nonattainment, and those above as "passing", or in attainment, for ALU. Two thresholds were set for the MBII to divide the range of scores into Poor, Fair, And Good condition categories based on percentiles of the distribution of scores among reference (least disturbed) sites. Scores of 74 or higher were assigned the condition of Good, scores between 39 and 73 were considered Fair, and scores below 39 were considered Poor (Klemm *et al.*, 2003). We assigned each sample to a condition category and estimated the proportion of site condition assessments that differed between the sampling methods. We ran McNemar's test of symmetry (Agresti, 1996) on each table to determine if one method resulted in higher bioassessment condition ratings than the other. Because there were so few scores with an MBII condition of Good, we combined the Good and Fair categories for this test.

Potential Influence of Physical Factors

To determine the potential effect of physical site characteristics on differences between methods, we calculated Spearman rank correlations of mean thalweg depth, wetted width, and bankfull width with differences between methods in metrics and index scores. We selected these habitat variables because they reflect general physical differences among wadeable streams that might influence method comparisons. We calculated the differences within each site as the single habitat value minus the multiple habitat value for each metric and index score. To account for the large number of correlations, we performed Holm's procedure (Legendre and Legendre, 1998) to adjust *p*-values for VSCI and MBII metrics separately. Then, only correlations with an adjusted *p*-value of 0.05 or less were considered significant.

RESULTS

We encountered a relatively large range of physical and water quality site characteristics in this study. Stream elevation ranged from 44 to 375 m and slope from 0.4 to 16%. The average depth was from 12 cm in very small streams to almost a meter, and wetted width ranged from 1.6 to 15.1 m. Streams were either dominated by riffle/run habitats (primarily runs) or had approximately equal amounts of riffle/run and pool/glide habitats. Therefore, we were able to collect high gradient RH parameter scores at all sites, and only these RH parameters were included in analyses. The RH scores spanned most of the possible range for each parameter, and the water quality parameters ranged from very low to moderate or high readings. Of the 20 jabs or kicks performed at each site for the multiple habitat method, both the number of kicks and the subset of those kicks in sand habitat ranged widely.

Taxonomic Similarity

The SMCs were relatively high at both the family and species levels when based on both abundances (BC) and presence-absence (Jaccard) (Table 2). In fact, the mean SMC for species level data was generally much higher (approximately 100%) than that for family level data for both similarity indices. This indicates that, relative to the similarity between duplicate samples within a method, the similarity of samples between methods was very comparable. In general, however, similarities themselves were rather

TABLE 2. Mean Jaccard and Bray-Curtis Taxonomic Similarities Among Duplicates and Between Single and Multiple Habitat Field Methods for Family and Species Levels, Including Mean Classification Strength Sampling Method Comparability (SMC).

Taxonomic Level	Index	Between Methods	Within Single	Within Multiple	Mean SMC (%)
Family	Jaccard	0.39	0.47	0.49	78
	Bray-Curtis	0.56	0.65	0.67	83
Species	Jaccard	0.25	0.22	0.27	104
	Bray-Curtis	0.40	0.36	0.44	102

low, and presence-absence-based similarity (SJ) typically was much lower than that based on abundances (BC).

Similarities between duplicate samples within a method did not differ between single and multiple habitat methods, regardless of similarity measure or taxonomic level of data. For family level data, Wilcoxon rank sum tests showed very minimal differences between methods in terms of similarity of duplicate samples for both the Jaccard (S = 5.5, p = 0.625) and BC (S = 0.5, p = 1.000) indices. At the species level, the small differences between methods were nonsignificant (S = −15.5, p = 0.131 for both indices).

Indices and Metrics

Direct Comparisons. Field method did not consistently affect the values of the VSCI and its metrics. Differences in VSCI scores ranged from 4.2 to 14.8 points, with a median difference of 5.2 points on a 100-point scale. This was not a significant

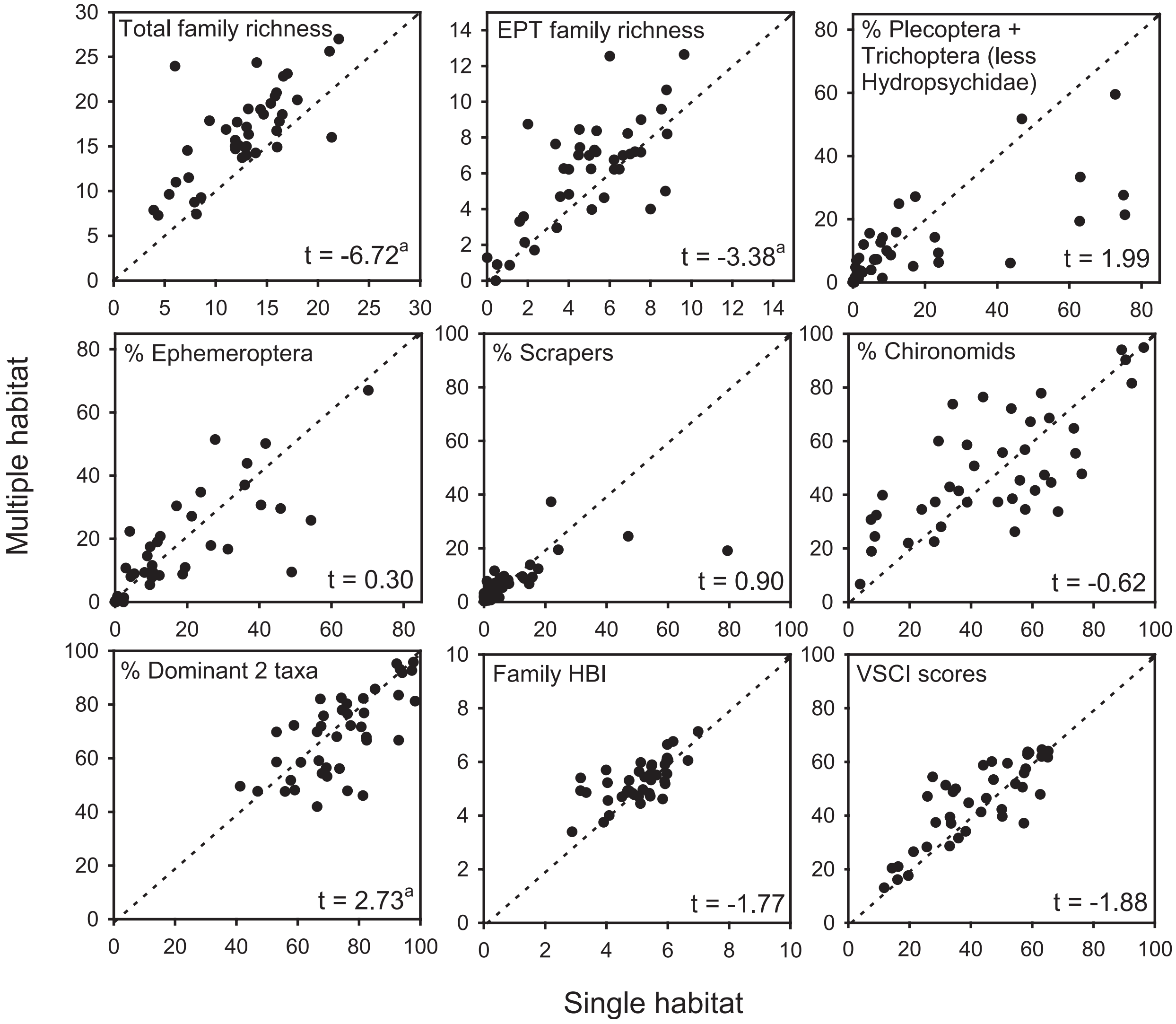

FIGURE 2. Bivariate Plots of Each Metric and Virginia Stream Condition Index (VSCI) Scores of the Single Habitat Method Plotted Against the Corresponding Multiple Habitat Value for Each Site. The dotted diagonal line represents a 1:1 relationship between methods. The paired t-test statistics are provided, with ([a]) indicating significance at the 0.01 level.

difference, but total and EPT family richness were both significantly higher in multiple habitat samples, and percent dominant two families was significantly higher in single habitat samples (Figure 2). For several metrics that did not differ significantly between methods, there was large variation in the relationship between the single and multiple habitat values (Figure 2). Regressions of single habitat VSCI scores on multiple habitat VSCI scores resulted in 90% prediction intervals ranging in width from approximately 31.4 to 33.0 points (Figure 3), almost one-third of the range of the index. The coefficient of determination (R^2) for the regression was approximately 0.68, indicating that about two-thirds of the variation in single habitat index scores could be explained by multiple habitat scores. Residuals were approximately normal with only mild heteroscedasticity.

The effect of field method on the MBII and its component metrics was also inconsistent. The MBII, Ephemeroptera richness, and HBI were all significantly higher in multiple habitat samples, and percent dominant five taxa was significantly higher in single habitat samples. The relationship between values from the two methods was highly variable for most metrics, particularly taxa richness metrics (Figure 4). The regression of single habitat MBII scores on multiple habitat scores produced 90% prediction intervals ranging in size from approximately 45.2-47.0 points (Figure 3), nearly half the 100-point range of the MBII. The R^2 of this regression was approximately 0.37, with just over a third of the variation in single habitat MBII scores explained by multiple habitat scores. Residuals were approximately normal with homogeneous variance.

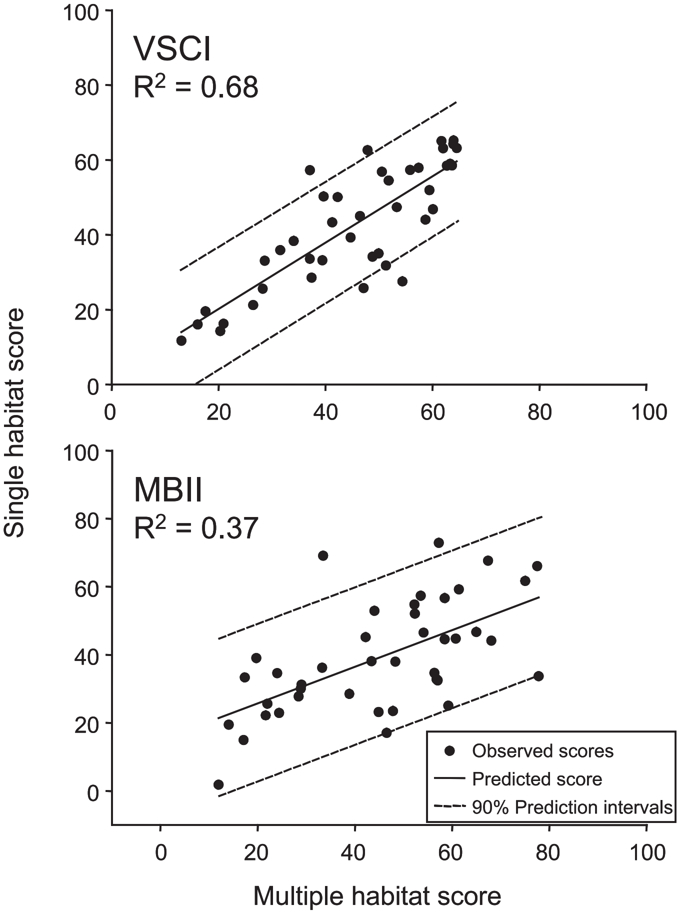

FIGURE 3. Regression of Single Habitat Virginia Stream Condition Index (VSCI) and Macroinvertebrate Biotic Integrity Index (MBII) Scores on Multiple Habitat Values (solid line) with 90% Prediction Intervals (dashed lines).

Variability. The level of variability was similar between the two methods for the VSCI, but differed strongly for the MBII. Both measurement error (RMSE) and the ability to distinguish among sites (S/N) were similar between methods when based on the VSCI (Table 3). However, the MBII showed a moderate difference in measurement error but a five-fold difference in S/N, indicating a stronger separation among sites for the multiple habitat field method. A minimum S/N of 2.0 is suggested by Kaufmann *et al.* (1999) for moderate precision, and a ratio of more than 6.0 suggests good precision. Thus, both the MBII and VSCI showed good precision for the multiple habitat method. For the single habitat method, the MBII had poor precision (S/N = 1.82) and the VSCI showed good precision. Residuals of both GLM models showed approximate normality and homoscedasticity.

Relationships to Potential Stressors. We used the first two PCA axes as measures of composite disturbance gradients (Table 4). Due to missing pH data for six sites and missing nutrient data for one site, the PCA and subsequent correlations were based on 34 sites. The first PC axis explained about 27.6% of variation [eigenvalue (λ) = 3.59] and was strongly positively associated with high quality vegetative cover on banks and in riparian areas and to a lesser degree with optimal water flow and minimal channelization. At the same time, this axis was also positively associated with road density and conductivity. The second axis explained about another 20% of variation (λ = 2.64) and was most strongly correlated with favorable epifaunal substrates for colonization and low embeddedness, indicating a general gradient of sedimentation.

Based on qualitative examination of correlations with PC axes, single habitat metrics tended to be more closely related to disturbance gradients (Table 5). This was particularly true for the sedimentation gradient (PC 2), with the VSCI metrics percent

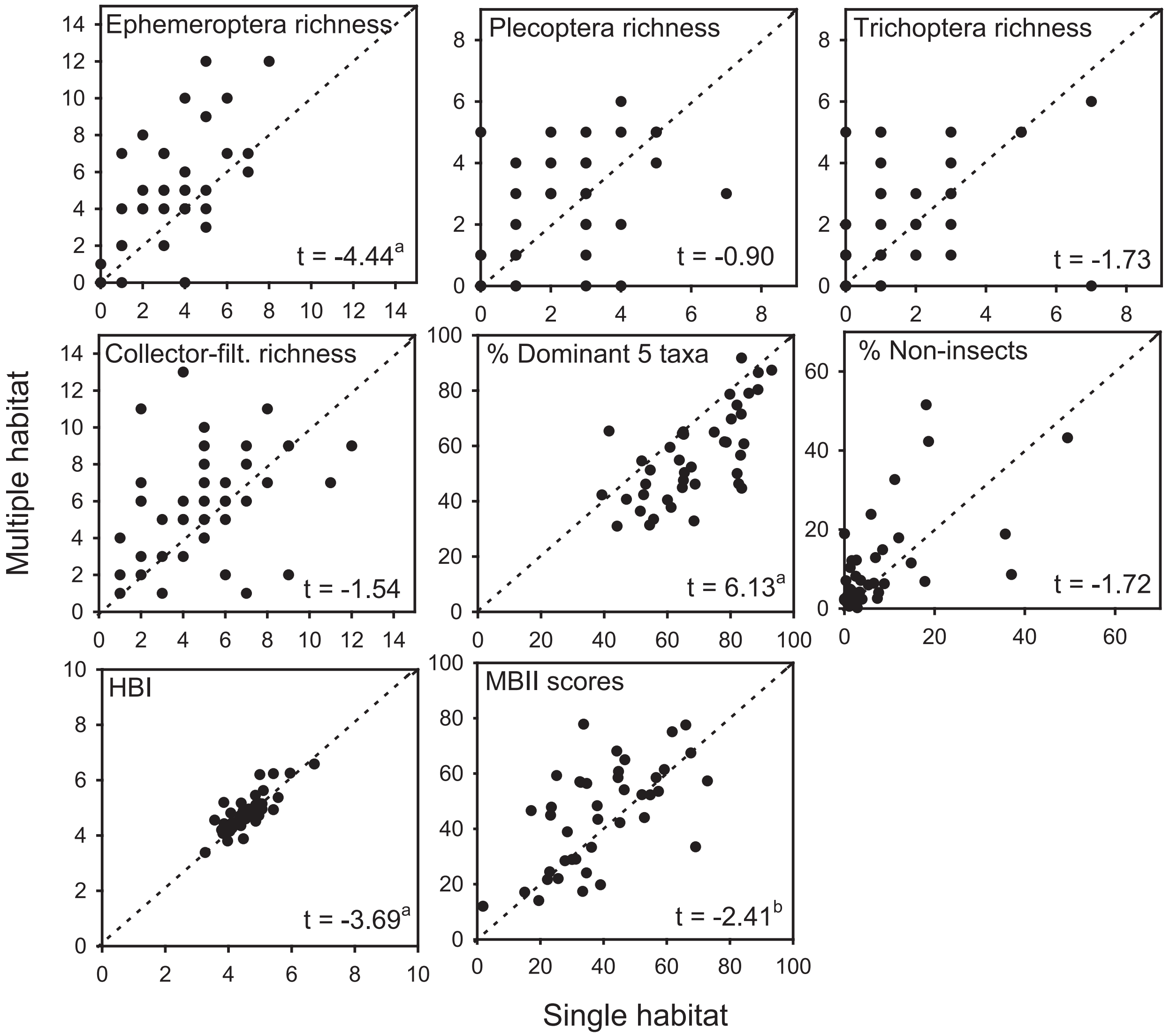

FIGURE 4. Bivariate Plots of Each Metric and Macroinvertebrate Biotic Integrity Index (MBII) Scores of the Single Habitat Method Plotted Against the Corresponding Multiple Habitat Value for Each Site. The dotted diagonal line represents a 1:1 relationship between methods. The paired *t*-test statistics are provided, with ([a]) indicating significance at the 0.01 level and ([b]) at the 0.05 level.

TABLE 3. Variability of VSCI and MBII Scores By Method, Measured as Root Mean Square Error (RMSE) and Signal-to-Noise Ratio (S/N).

Index	Field Method	RMSE	S/N
VSCI	Single	5.54	7.23
	Multiple	4.84	8.94
MBII	Single	8.30	1.82
	Multiple	4.99	9.14

Chironomidae and percent dominant two families having much stronger correlations for the single habitat data. The gradient related to riparian condition and bank vegetation (PC 1) only showed relationships with the FBI and the HBI, but they were similar between single and multiple habitat data. The MBII and its component metrics were very similar between methods with respect to relationships with potential stressor gradients.

Bioassessment Endpoints

Based on the VSCI, the assessment category (pass or fail) differed between the single and multiple habitat samples at only 5 of 41 sites (12%) (Table 6). At three of these sites, the difference in VSCI scores between methods was five points or less, indicating that both samples scored close to the cutoff value. At the other two sites, the difference was more than 13 points. There was no significant tendency for one method to lead to a

TABLE 4. Pearson Correlations of First Two Principal Components (PC) (% variance explained) With Abiotic Variables Included in Analysis (n = 34).

Variable (transform, if any)	PC 1 (27.6%)	PC 2 (20.3%)
RH EPISUB (x^2)	0.38	0.83
RH EMBED (x^2)	0.14	0.79
RH VELDEP	0.21	0.52
RH CHANFLW (x^2)	0.64	−0.11
RH CHANALT	0.61	−0.02
RH BKSTAB	0.42	0.56
RH BKVEG	0.76	−0.10
RH RIPVEG	0.70	−0.34
TKN (ln(x))	0.45	−0.12
Conductivity [ln(x)]	0.60	−0.52
pH	0.48	−0.38
% Agriculture (sq. root)	0.48	−0.25
Road density [ln(x + 1)]	0.58	−0.32

TABLE 5. Spearman Rank Correlations of First Two Principal Components (PC) With Two Indices and Their Component Metrics (n = 34).

Taxonomic Level	Metric/ Index	PC 1		PC 2	
		Multiple	Single	Multiple	Single
Family	VSCI	−0.30	−0.19	0.45	0.59
	Total family richness	−0.11	−0.01	0.44	0.49
	EPT family richness	−0.09	−0.10	0.51	0.51
	% Ephemeroptera	−0.31	−0.38	0.32	0.18
	% Plecoptera + Trichoptera − Hydropsychidae	−0.33	−0.21	0.41	0.54
	% Scrapers	0.07	0.15	0.24	0.33
	% Chironomidae	0.27	0.22	−0.30	−0.52
	% Dominant two families	0.11	−0.06	−0.36	−0.54
	FBI	0.55	0.35	−0.33	−0.46
Species	MBII	−0.07	0.07	0.47	0.40
	Ephemeroptera taxa richness	−0.24	−0.04	0.42	0.35
	Plecoptera taxa richness	−0.15	−0.04	0.44	0.51
	Trichoptera taxa richness	0.28	0.08	0.32	0.32
	Coll.-filterer taxa richness	0.28	0.36	0.17	0.17
	% Dominant five taxa	−0.05	−0.35	−0.29	−0.30
	% Non-insects	0.29	0.26	−0.16	0.13
	HBI	0.42	0.39	−0.21	−0.10

higher assessment category (McNemar's test, S = 1.80, p = 0.3750).

The assessment categories based on the MBII differed more strongly between the two methods, with 14 of 41 sites (34%) rated differently (Table 6). Of these, 12 moved to a higher category from the single to the multiple habitat samples. Sites with different assessment categories between methods differed by an average of 15 points. Using the combined categories of Fair and Good, 10 sites were assessed to higher categories with the multiple habitat and two with the single habitat ratings. In this case, the multiple habitat data showed a significant tendency to result in a higher assessment category than single habitat data (McNemar's test, S = 5.33, p = 0.0386).

TABLE 6. Correspondence of VSCI Condition Ratings Between Single and Multiple Habitat Samples.

	Multiple Habitat	
Single Habitat	Fail	Pass
VSCI (0.3750)		
Fail	31	4
Pass	1	5
	Poor	Good/Fair
MBII (0.0386)		
Poor	13	10
Good/Fair	2	16

Note: Values in parentheses indicate p-values for McNemar's test of significant tendency of one method to result in higher condition rating.

Potential Influence of Physical Factors

Mean thalweg depth was the only physical factor related to differences between methods. Differences (single-multiple habitat) in both percent dominant two families and HBI increased with increasing depth (Figure 5), and all other correlations were nonsignificant after adjusting p-values. For percent dominant two families, there tended to be slightly larger values for multiple habitat samples in shallower sites, but in deeper sites, single habitat values tended to be much

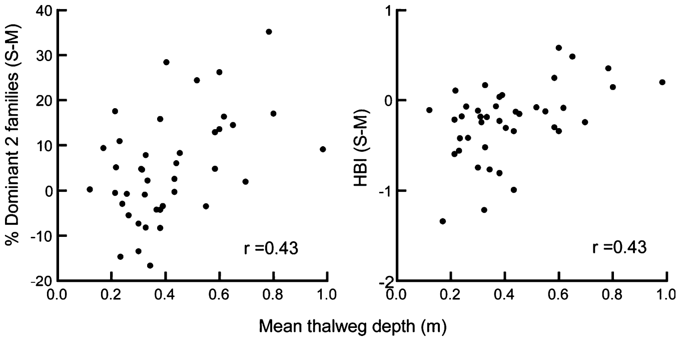

FIGURE 5. Relationship of Mean Thalweg Depth to Differences Between Single and Multiple Habitat Methods for Percent Dominant Two Taxa, and Species-level Hilsenhoff Biotic Index (HBI). Spearman correlation coefficients are also shown and are significant at the 0.05 level (N = 41).

larger. HBI values tended to be higher in multiple habitat samples at shallow sites, switching to slightly higher values in single habitat samples at deeper sites.

DISCUSSION

Taxonomic Similarity, Indices, and Metrics

Differences between the single and multiple habitat field methods were more evident at the metric and index levels than at the level of taxonomic similarity, even more so for the MBII and its component metrics than for the VSCI. There were few significant differences based on *t*-tests, largely due to the level of variation in differences between methods (i.e., more scatter in bivariate plots). However, many richness metrics tended to be significantly higher and dominance metrics significantly lower in multiple habitat samples. This might be expected because the multiple habitat method by definition is sampling more habitats in streams lacking abundant riffle habitat, thus increasing the likelihood of picking up new taxa relative to a single habitat sampling method (Parsons and Norris, 1996). By collecting more taxa, the percent of individuals in the most abundant taxa becomes more diluted, as long as the additional taxa are not rare. Weaker differences for the VSCI scores and metrics as compared to MBII are logical because data are aggregated to the family level, and collection of additional species should have less effect on metric values at this level. In addition, the MBII is based on a 300-organism count, and this would likely lead to more of the less common taxa in a sample being detected compared to the 150-organism subsample size used for the VSCI (Flotemersch *et al.*, 2006b). In general, regression for both indices showed some ability of single habitat scores to predict multiple habitat index scores, thus allowing for an algorithm to convert from one method to the other. However, wide prediction intervals indicate that predicted values are associated with high variability, and consequently, that index scores are not directly interchangeable between the two field methods.

Strong differences in variability between methods became apparent only at the species level. Although the SMC values were very high, indicating high agreement between methods, similarities in general were very low for the species level data. Similarities between method duplicates tended to be slightly lower for single habitat than multiple habitat samples. This discrepancy could be caused by single habitat samples containing more taxa that are rare within samples, leading to detection of a given taxon in one duplicate sample but not the other. This would account for the even lower Jaccard similarity, which is based only on presence-absence and not abundance of taxa. The multiple habitat samples also distinguished among sites better than the single habitat method (higher S/N), but only for the MBII. These results are contrary to what we expected, given the wider variety of habitats sampled for the multiple habitat field method and the results of previous research. Ostermiller and Hawkins (2004) found that precision of RIVPACS-type predictive models was higher for single habitat samples when rarer taxa were excluded from the list of expected taxa. Parsons and Norris (1996) found that inconsistent sampling effort across habitats (as occurs in multiple habitat sampling) from site to site contributes to assessment variability and may confound detection of impairments.

We did not observe strong differences in relationships with potential stressors between methods except in a few cases. The index scores were similarly correlated with the sedimentation gradient between methods, and only differed strongly for percent Chironomidae and percent dominant two taxa. These results do not indicate a strong advantage of using one method over the other for the purposes of detecting changes in water quality. Our results correspond with other studies that have concluded that multiple habitat sampling does not improve on the ability to distinguish impairment in streams (Parsons and Norris, 1996) and is unnecessary for broad-scale biological monitoring (Hewlett, 2000).

The first PC axis provided a conflicting result, positively correlated with conductivity and road density but also with several habitat variables. Anecdotally, some of the sites with better riparian habitat were located between developed areas, and this might account for the relationship between these two types of measures of abiotic condition. Given that HBI and FBI were most strongly correlated with this PC axis, and these metrics tend to increase with increasing levels of disturbance, it is likely that conductivity and road density were more representative of the disturbance gradient than the habitat variables.

Bioassessment Endpoints

There were some differences in assessment endpoints based on both the VSCI and the MBII, and the differences were more pronounced for the MBII. Multiple habitat sampling tended to result in a higher MBII score and hence, a higher assessment category for a given site. Given the tendency toward higher taxa richness in multiple habitat samples, this result makes sense because the MBII is made up of more richness-related metrics than the VSCI. Over a number of metrics, small differences in scores at the metric level can translate into much larger differences in index scores, potentially leading to assignment of a site to a different assessment category. This seems to contrast with other studies that have concluded that multiple habitat sampling generally does not result in a different assessment from single habitat sampling (Parsons and Norris, 1996; Hewlett, 2000; Ostermiller and Hawkins, 2004). However, these other studies based this comparison on predictive models developed using the data from each sampling method, whereas our comparison relies on existing multimetric indices developed largely from single habitat data. Development of a multimetric index specific to multiple habitat data is not feasible for this study because a much larger dataset would be required, so a comparison similar to these other studies cannot be made.

Potential Influence of Physical Factors

The relationship of method differences with depth for two metrics may be associated with the similarity of sampled habitats between methods at shallow and deep sites. In deeper sites, the habitats sampled by multiple and single habitat methods will tend to differ more, and we would expect to observe greater differences in metric and index values between methods. In shallower sites, with more riffle habitats, we expect both sampling methods to be focused on riffle habitat, and this should result in more similar metric and index values between methods. For the percent dominant two families, this appears to be the case, with the largest differences occurring at deeper sites. For this metric, lower values, indicating better condition, were observed in single habitat samples at shallow sites and in multiple habitat samples in deeper sites. Conversely, the HBI exhibited the largest differences between methods at shallower sites, with multiple habitat samples indicating poorer condition (higher values). This result indicates that the other habitats sampled in the multiple habitat method tended to contain more tolerant macroinvertebrate taxa. In any case, there does appear to be some influence of the depth of a site on the comparability of the single and multiple habitat methods, and this corresponds with a comparison of several sampling methods in nonwadeable streams (Blocksom and Flotemersch, 2005b).

Conclusions and Recommendations

Determining which of the two methods to use for sampling macroinvertebrates depends in large part on the history of the program for which the samples are collected, the types of streams to be assessed, and the objectives of the organization assessing the streams. If a program has a history of collecting samples using the single habitat method, there is little evidence here to compel that program to change its method to the multiple habitat method, particularly if riffle/run habitat is available in the streams to be sampled. If a change is preferable, however, our results support those of Wang *et al.* (2006), who concluded that directly merging datasets collected using different sampling methods could be problematic at any taxonomic level and should be avoided. Even using the VSCI, which showed minimal differences between methods, metric scoring would require recalibration specifically using multiple habitat data because some component metrics did differ between methods. For example, EPT richness tended to be higher in multiple habitat samples, and scoring of this metric should reflect this higher expectation. Such recalibration of metrics would likely require the collection of extensive additional data using the new method.

The type of streams being assessed is an important factor in deciding which of these methods to use. As Ostermiller and Hawkins (2004) concluded, the multiple habitat field method may be preferable in streams with limited riffle/run habitats. This would allow the major nonriffle habitats to be sampled in these streams and is of particular interest for deeper wadeable streams or lower gradient streams dominated by pools and glides. For streams with predominantly riffle habitat, the multiple habitat method may not provide much additional information. The streams of the Piedmont and Northern Piedmont ecoregions sampled in this study exhibited a wide range of physical site characteristics, including flat and steep slopes, low and higher elevations, narrow and wide wetted widths, and shallow and deep thalwegs. An area with this level of variation may be conducive to the multiple habitat method because it tends to be applicable across a wide variety of streams types (Barbour *et al.*, 1999), even if it might not be the most efficient sampling method in some streams.

Finally, the objectives of the program assessing streams may influence method choice. The typical goal of bioassessment and monitoring is not to collect all of the taxa at a site but to collect a representative biological sample that reflects the water quality and habitat conditions at a site. For the streams sampled in our study, it is not clear that one method really does this better than the other at either taxonomic level, but the limited evidence presented here and in previous studies (Blocksom and Flotemersch, 2005a) does suggest that the ability of some individual metrics to detect specific disturbance gradients may vary somewhat depending on sampling method. In general, however, both the multiple and single habitat methods were able to provide effective samples across the wide range of stream types.

ACKNOWLEDGMENTS

This work would not have been possible without the numerous landowners who granted us permission to cross their properties to access the streams that were sampled for this project. For their thoughtful reviews of earlier drafts of either this manuscript or an earlier report, we thank Larry Willis, Jason Hill, George Devlin and Mary Dail of the VDEQ, and Brent Johnson of the USEPA National Exposure Research Laboratory, Cincinnati, Ohio, as well as three anonymous reviewers. This paper has been subjected to Agency review and approved for publication. The views expressed in this paper are those of the authors and do not necessarily reflect the views and policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Some of the data described in this article were produced by the U.S. Environmental Protection Agency through its Environmental Monitoring and Assessment Program (EMAP), http://www.epa.gov/emap/.

LITERATURE CITED

Agresti, A., 1996. An Introduction to Categorical Data Analysis. John Wiley and Sons, Inc., New York, New York, 290 pp.

Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling, 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates, and Fish (Second Edition). EPA-841-B-99-002, U.S. Environmental Protection Agency; Office of Water, Washington, D.C.

Blocksom, K.A. and J.E. Flotemersch, 2005a. Field and Laboratory Performance Characteristics of a New Sampling Protocol for Riverine Macroinvertebrate Assemblages. EPA/600/R-05/097. U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Cincinnati, Ohio.

Blocksom, K.A. and J.E. Flotemersch, 2005b. Comparison of Macroinvertebrate Sampling Methods for Nonwadeable Streams. Environmental Monitoring and Assessment 102:243-262.

Brigham, A.R., W.U. Brigham, and A. Gnilka, 1982. Aquatic Insects and Oligochaetes of North and South Carolina. Midwest Aquatic Enterprises, Mahomet, Illinois, 837 pp.

Cao, Y., C.P. Hawkins, and A.W. Storey, 2005. A Method for Measuring the Comparability of Different Sampling Methods Used in Biological Surveys: Implications for Data Integration and Synthesis. Freshwater Biology 50:1105-1115.

Clarke, R.T., J.F. Wright, and M.T. Furse, 2003. RIVPACS Models for Predicting the Expected Macroinvertebrate Fauna and Assessing the Ecological Quality of Rivers. Ecological Modeling 160:219-233.

Diamond, J.M., M.T. Barbour, and J.B. Stribling, 1996. Characterizing and Comparing Bioassessment Methods and Their Results: A Perspective. Journal of the North American Benthological Society 15:713-727.

Flotemersch, J.E., K.A. Blocksom, J.J. Hutchens, Jr., and B.C. Autrey, 2006b. Development of a Standardized Large River Bioassessment Protocol (LR-BP) for Macroinvertebrate Assemblages. River Research and Applications 22:775-790.

Flotemersch, J.E., J.B. Stribling, and M.J. Paul, 2006a. Concepts and Approaches for the Bioassessment of Non-Wadeable Streams and Rivers. EPA/600/R-06/127, U.S. Environmental Protection Agency, Cincinnati, Ohio.

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.

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.

Hewlett, R., 2000. Implications of Taxonomic Resolution and Sample Habitat for Stream Classification at a Broad Geographic Scale. Journal of the North American Benthological Society 19:352-361.

Hilsenhoff, W.L., 1987. An Improved Biotic Index of Organic Stream Pollution. The Great Lakes Entomologist 20:31-39.

Hollander, M. and D.A. Wolfe, 1999. Nonparametric Statistical Methods. John Wiley and Sons, Inc., New York, New York, 787 pp.

Houston, L., M.T. Barbour, D. Lenat, and D. Penrose, 2002. A Multi-Agency Comparison of Aquatic Macroinvertebrate-Based Stream Bioassessment Methodologies. Ecological Indicators 1:279-292.

Hurlbert, S.H., 1971. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52:577-586.

Kaufmann, P.R., P. Levine, E.G. Robison, C. Seeliger, and D.V. Peck, 1999. Quantifying Physical Habitat in Wadeable Streams. EPA/620/R-99/003, U.S. Environmental Protection Agency, Washington, D.C.

Kelly, P. M. and J.M. White, 1993. Preprocessing Remotely Sensed Data for Efficient Analysis and Classification, Knowledge-Based Systems in Aerospace and Industry. Proceedings of SPIE 1993:24-30.

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.

Legendre, P. and L. Legendre, 1998. Numerical Ecology. Elsevier, Amsterdam, 853 pp.

McCune, B. and J.B. Grace, 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, Oregon, 283 pp.

Merritt, R.W. and K.W. Cummins, 1996. An Introduction to the Aquatic Insects of North America (Third Edition). Kendall/Hunt Publishing Company, Dubuque, Iowa, 862 pp.

Omernik, J.M., 1995. Ecoregions: A Framework for Environmental Management. *In:* Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making, W.S. Davis, and T.P. Simon (Editors). Lewis Publishers, Boca Raton, Florida, pp. 49-62.

Ostermiller, J.D. and C.P. Hawkins, 2004. Effects of Sampling Error on Bioassessments of Stream Ecosystems: Application to RIVPACS-Type Models. Journal of the North American Benthological Society 23:363-382.

Parsons, M. and R.H. Norris, 1996. The Effect of Habitat-Specific Sampling on Biological Assessment of Water Quality Using a Predictive Model. Freshwater Biology 36:419-434.

Peckarsky, B.L., P.R. Fraissinet, M.A. Penton, and D.J. Conklin, Jr, 1990. Freshwater Macroinvertebrates of Northeastern North America. Cornell University Press, Ithaca, New York, 442 pp.

Pennak, R.W., 1989. Freshwater Invertebrates of the United States: Protozoa to Mollusca (Third Edition). John Wiley and Sons, New York, New York, 656 pp.

Plafkin, J.L., M.T. Barbour, K.D. Porter, S.K. Gross, and R.M. Hughers, 1989. Rapid Bioassessment Protocols for Use in Rivers and Streams: Benthic Macroinvertebrates and Fish. EPA-440-4-89-001, U.S. Environmental Protection Agency, Office of Water Regulations and Standards, Washington, D.C.

Southerland, M., J. Volstad, L. Erb, E. Weber, and G. Rogers, 2006. Proof of Concept for Integrating Bioassessment Results from Three State Probabilistic Monitoring Programs. EPA/903/R-05/003, U.S. Environmental Protection Agency, Office of Environmental Information and Mid-Atlantic Integrated Assessment Program, Ft. Meade, Maryland.

van Tongeren, O.F.R, 1995. Cluster Analysis. *In:* Data Analysis in Community and Landscape Ecology, R.H.G. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren (Editors). Cambridge University Press, Cambridge, United Kingdom, pp. 174-212.

U.S. Environmental Protection Agency (USEPA), 2002. Summary of Biological Assessment Programs and Biocriteria Development for States, Tribes, Territories, and Interstate Commissions: Streams and Wadeable Rivers. EPA-822-R-02-048. U.S. Environmental Protection Agency, Washington, D.C.

U.S. Environmental Protection Agency (USEPA), 2005. Guidance for 2006 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act. U.S. Environmental Protection Agency, Office of Water, Washington, D.C. http://www.epa.gov/owow/tmdl/2006IRG/, *accessed* March 10, 2008.

Van Sickle, J., 1997. Using Mean Similarity Dendrograms to Evaluate Classification. Journal of Agricultural, Biological, and Environmental Statistics 2:370-388.

Van Sickle, J., 2003. Analyzing Correlations Between Stream and Watershed Attributes. Journal of the American Water Resources Association 39:717-726.

Vogelmann, J.E., T.L. Sohl, P.V. Campbell, and D.M. Shaw, 1998. Regional Land Cover Characterization Using Landsat Thematic Mapper Data and Ancillary Data Sources. Environmental Monitoring and Assessment 51:415-428.

Wang, L., B.W. Weigel, P. Kanehl, and K. Lohman, 2006. Influence of Riffle and Snag Habitat Specific Sampling on Stream Macroinvertebrate Assemblage Measures in Bioassessment. Environmental Monitoring and Assessment 119:245-273.