United States
Environmental Protection
Agency

Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams

EPA/903/R-03/003
March 2003

Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams

Prepared for:
U.S. Environmental Protection Agency
Environmental Science Center
701 Mapes Road
Ft. Meade, MD 20755-5350
Wayne Davis, Project Officer

Prepared by:

Leska S. Fore
Statistical Design
136 NW 40th St.
Seattle, WA 98107

Under subcontract from:
Technology Planning & Management Corporation
Mill Wharf Plaza, Suite 208
Scituate, MA 02066

Printed on chlorine free 100% recycled paper with
100% post-consumer fiber using vegetable-based ink.

NOTICE

This document has been reviewed and approved in accordance with U.S. Environmental Protection Agency policy. Mention of trade names, products, or services does not convey and should not be interpreted as conveying, official EPA approval, endorsement, or recommendation for use.

Funding was provided by the U.S. Environmental Protection Agency under U.S. Department of Commerce, Commerce Information Technical Solutions Contract No. 50-CMAA-900065 with Technology Planning and Management Corporation.

The appropriate citation for this report is:

Fore, Leska S. 2003. Developing Biological Indicators: Lessons Learned from Mid-Atlantic Streams. Report prepared for EPA under Contract No. 50-CMAA-900065. EPA 903/R-003/003. U.S. Environmental Protection Agency, Office of Environmental Information and Mid-Atlantic Integrated Assessment Program, Region 3, Ft. Meade, MD.

This document can be downloaded from EPA's website for Biological Indicators of Environmental Health:

http://www.epa.gov/bioindicators/

TABLE OF CONTENTS

NOTICE ii

FIGURES iv

TABLES v

ABSTRACT vi

ACKNOWLEDGMENTS vii

I. INTRODUCTION 1

II. SAMPLING DESIGN 3
A probabilistic sampling design was the best choice for MAIA 4
Free information is cost-effective 5
Reference sites did not always meet criteria for reference condition 6

III. THE PERILS OF DATA MANAGEMENT 9
Different "names" for the same site caused confusion 9
Original data must be archived 10
File structure mattered 10
Simple files were best 13

IV. LINKING HUMAN DISTURBANCE TO BIOLOGICAL CHANGE 15
Addressing concerns about circular reasoning 15
Metric testing included safeguards against circular reasoning 18
Patterns of human disturbance were complex 19
Integrated measures of disturbance were better predictors of index values 20

V. METRIC TESTING 23
Simple criteria were used first to eliminate potential metrics 24
Statistical precision was no substitute for correlation with disturbance 24
Watershed features were confounded with metric response to disturbance 26
Metrics from different assemblage types were eliminated for different reasons 26

VI. DEVELOPMENT AND APPLICATION OF MULTIMETRIC INDEXES 29
Biological criteria depend on the definition of reference sites 29
Patterns of index variability were similar across assemblage types 29
Invertebrate and diatom index values were comparable for pool and riffle samples 33
Assemblages differed in their sensitivity to disturbance types 33

VII. CONCLUSIONS 35

REFERENCES 37

FIGURES

FIGURE 1 COMPARISON OF REFERENCE SITE SELECTON CRITERIA 8

FIGURE 2 EXAMPLE DATA FILE FOR DIATOM SAMPLES 12

FIGURE 3 EXAMPLE OF A HORIZONTAL FILE STRUCTURE 14

FIGURE 4 MULTIMETRIC INDEX VALUES BY DISTURBANCE TYPE 17

FIGURE 5 RANGES OF VALUES FOR POOL DEPTH AND EMBEDDEDNESS 25

FIGURE 6 VARIANCE COMPONENTS FOR THE INVERTEBRATE INDEX AND ITS METRICS 31

FIGURE 7 PERCENTAGE CHANGE THAT MULTIMETRIC INDEXES COULD DETECT 32

TABLES

TABLE 1 SPEARMAN'S CORRELATION OF MULTIMETRIC INDEXES AND HUMAN DISTURBANCE MEASURES 18

TABLE 2 SPEARMAN'S CORRELATION MATRIX FOR MEASURES OF HUMAN DISTURBANCE 21

TABLE 3 CANDIDATE METRICS TESTED FOR FISH AND INVERTEBRATE MULTIMETRIC INDEXES AND REASON FOR EXCLUSION 27

TABLE 4 COMPONENTS OF VARIANCE FOR DIATOM, INVERTEBRATE, AND FISH MULTIMETRIC INDEXES 31

TABLE 5 BIOLOGICAL METRICS INCLUDED IN THE FISH, INVERTEBRATE, AND DIATOM INDEXES 34

ABSTRACT

As part of the U.S. Environmental Protection Agency's (USEPA) Environmental Monitoring and Assessment Program (EMAP), a survey of water chemistry, land use, riparian condition, and channel morphology was conducted to understand how human influence alters fish, invertebrate and periphyton assemblages. During 1993-1996, 296 sites were sampled for fish, 583 for invertebrates and 317 for periphyton. A primary goal of the Mid-Atlantic Integrated Assessment (MAIA) study was to define biological indicators for each assemblage that could be used to assess stream condition at the regional level.

During the course of the project, researchers working independently derived different approaches to data analysis and reported different results regarding the relationship between human influence and biological change. To build consensus among the scientists involved, USEPA sponsored a series of workshops to create a consistent approach for testing and selecting biological indicators for fish, invertebrates and periphyton. This document presents some of the issues from those workshops that were most challenging to resolve.

The probabilistic sampling approach was the most efficient method for obtaining an unbiased estimate of regional condition and is a cornerstone of EMAP. The coarse level of resolution has made this design somewhat controversial for application at the state level, but alternative sampling designs fail to yield data that can be applied across the region. Although much attention was given to database design, data management issues often drove the agenda. Major issues not anticipated during the design phase arose in the course of analyzing the data. Final results were delayed, for example, by lengthy discussions of how to count invertebrate taxa with incomplete phylogeny and how to combine periphyton data from soft algae and diatom counts that were derived from different laboratory methods.

One of the greatest challenges for the MAIA project was selecting the best measure of human disturbance from among the hundreds of potential variables. Through discussion and consensus, a set of variables was selected that summarized human influence at multiple spatial scales and included measures related to specific types of disturbance and measures that integrated across the landscape. During the process of testing biological attributes for their association with disturbance, concerns repeatedly arose regarding whether the results would be robust in new contexts. As a consequence, safeguards were introduced at many steps of the analysis to avoid circular reasoning.

Similar statistical tests and criteria were used to select from among the list of candidate metrics for fish, invertebrates and periphyton. All three assemblages showed strong and consistent associations with general measures of disturbance, but showed different sensitivities to individual stressors. Diatoms were more sensitive to water chemistry and invertebrates were more sensitive to riparian condition. Multimetric indexes for each assemblage showed a similar ability to detect both differences in sites and change through time (trend). The indexes for each assemblage represent statistically reliable and biological meaningful monitoring tools for assessing and reporting the biological condition of wadeable streams in the Mid-Atlantic.

ACKNOWLEDGMENTS

This document builds on the work of hundreds of people involved in the sampling design, field work, laboratory analysis, data management and statistical analysis for the Mid-Atlantic Integrated Assessment project. Much of the analysis described here was derived from the workshop discussions and publications of the participants. Editorial reviews by E. Chu and J. Scott improved this report. Additional helpful reviews were provided by K. Blocksom, W. Davis, P. Larsen, L. Reynolds, and J. Stoddard.

I. INTRODUCTION

The Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency (USEPA) was implemented to answer questions about the status and trends of natural resources and to connect observed changes in biological condition to stresses associated with human influence (Stevens, 1994). The program's emphasis on biological indicators of resource condition and its probabilistic sampling design distinguish EMAP from other national monitoring programs (Olsen et al., 1999). One of EMAP's fundamental mandates is to use the probabilistic sampling design to generalize conclusions drawn from a sample of sites to an entire region with a known level of statistical confidence (Stevens, 1994; Urquhart et al., 1998).

EMAP was conceived to include all of the nation's natural resources. To match that scope, resources were partitioned into groups: agro-ecosystems, rangeland, forests, the Great Lakes, estuaries, inland surface waters, and wetlands. To demonstrate how monitoring and assessment at the regional scale could be achieved, USEPA implemented the Mid-Atlantic Integrated Assessment (MAIA) pilot study for surface waters. Data were collected from wadeable streams in the eastern United States, from portions of Pennsylvania, Maryland, Delaware, Virginia, West Virginia, and southeastern New York (Herlihy et al., 2000). During the project's first four years (1993-1996), hundreds of sites were sampled for water chemistry, riparian condition, channel morphology, land cover, and fish tissue contaminants; samples were also taken of fish, invertebrate, and periphyton assemblages (Davis and Scott, 2000).

Although EMAP's goals have shifted somewhat since the program began in the late 1980s, its commitment to assessing resource condition in terms of resident biota has been constant (NRC, 1995). For surface waters, one purpose of EMAP is to support the goals of the Clean Water Act (Paulsen et al., 1998; Hughes et al., 2000). Under the act, states are required to assess their surface waters periodically and report the condition of those waters to Congress (Karr, 1991; Ransel, 1995). For biological monitoring and reporting, this mandate has typically been met by assessments based on multimetric biological indexes (Hughes et al., 1998; Hughes et al., 2000; O'Connor et al., 2000; Jackson et al., 2000; USEPA, 2002). Such indexes consist of biological metrics, defined as attributes of the biological assemblage that respond in predictable and measurable ways to human disturbance (Karr et al., 1986; Karr and Chu, 1999; Barbour et al., 1999). Metrics describe and measure a sampled population in terms of its taxonomic composition, community structure, trophic structure, and presence of tolerant and intolerant taxa.

One of the primary outcomes expected from the MAIA study was a set of biological indicators for fish, invertebrates, and periphyton that could be used to assess the biological condition of surface waters across the region, either by USEPA or by states within the region. Using the MAIA data, numerous studies have examined the relationships among biological assemblages, human land use, and geography (Pan et al., 1996; O'Connell et al., 1998; Kaufmann, et al.1999; Pan et al., 1999; Hill et al., 2000; Hughes et al., 2000; Pan et al., 2000; Waite et al., 2000, Hill et al., 2001; McCormick et al., 2001; Fore, 2002b; Klemm, et al., 2002; Blocksom, 2003). Over time, the methods for selecting biological indicators have evolved from individual approaches associated with different taxonomic assemblages and research groups to a general approach, developed during a series of USEPA-sponsored workshops, which could be applied more or less uniformly across different assemblages and would yield similar results when applied by different investigators. Through discussion and meetings, a consistent approach emerged for selecting biological indicators. Results of this process have been documented for fish and invertebrates and will also be applied to periphyton (Davis and Scott, 2000; McCormick et al., 2001; Klemm et al., 2003). The process included identifying potentially meaningful metrics, selecting from among the hundreds of

measures of site condition and human disturbance for metric testing, choosing appropriate statistical tests for metric analysis, and constructing a multimetric index.

Lessons learned from the Mid-Atlantic pilot study were related to sample design and data management as well as data analysis. Random site selection is a fundamental feature of any EMAP project, yet has been controversial as a sampling design at the state level. Other national monitoring programs and their sampling designs were considered by the EMAP designers but failed to meet the needs of a regional assessment program (Olsen et al., 1999). Alternative sampling designs cannot reliably provide unbiased estimates of regional condition, the statistics of greatest concern for USEPA. During the MAIA project, data management and storage decisions both supported and hindered the scientific analysis at every step. The best formats for different types of data were not obvious from the outset, and several versions of the taxonomic data files were created, corrupted and corrected over a period of months or years before the data were finalized and analysis could proceed.

Linking human disturbance to biological response generated concerns about circular reasoning, which were addressed at multiple stages. The complexity of human land use and the difficulty involved in quantifying site condition were recurrent issues. The enormous amount of data available for evaluating site condition did not necessarily make quantifying human disturbance any easier. During the workshops, much discussion centered on defining statistical criteria and logical decision rules for selecting metrics. The consensus was that metrics should be correlated with several different types of disturbance measured at multiple spatial scales. Finally, the development and application of the multimetric indexes revealed the robust ability of these indicators to detect change through time.

Multimetric indexes for fish and invertebrates were finalized before this report was completed, but analysis for the periphyton index is not yet complete (McCormick et al., 2001; Klemm et al., 2002). Diatoms represent a subset of the periphyton assemblage. Where appropriate, results from a previous analysis of diatom metrics are presented (Fore, 2002b).

This document presents issues, discussed in the MAIA workshops, that were controversial, confusing, or time-consuming in the development of biological monitoring tools for fish, invertebrates, and periphyton. It is aimed at agency scientists or managers tasked with implementing regional monitoring programs. The hope is that the lessons learned in the Mid-Atlantic pilot study may serve as guideposts in the development of indicators or management plans. The document's structure reflects the process of developing the biological monitoring tools. Major headings represent general concepts, and subheadings represent lessons learned from the Mid-Atlantic. Although the focus of the workshops and all the scientific analysis that they inspired was the development of biological monitoring tools, the reality was that the sampling design and data management issues defined the boundaries of what was possible at every step of the analysis (see, for example, Ward et al., 1986). Consequently, they are considered first.

II. SAMPLING DESIGN

The only way to know the true condition of a regional resource is to continuously monitor every site. This is impractical, and a subset of sites must be selected. A major strength of EMAP's monitoring strategy was the conscious choice of a sampling design before data were collected (Ward et al., 1986). The strengths and weaknesses of EMAP's probabilistic sampling design are debatable, but that controversy is secondary to the fact that the program's goals along with the expected outcomes and products were defined in advance, and the sampling design was chosen to support those goals (Stevens, 1994; Olsen et al., 1999). In short, there was a plan from the beginning for summarizing and applying the sample data to the larger region.

EMAP is a federal program with a national perspective and a mandate to estimate resource condition over a large spatial area (Stevens, 1994). A probabilistic survey sampling design randomly selects sites from all possible sites. An important advantage of random site selection is that results can be applied to similar sites in the region that were not sampled. In addition, estimates of regional condition derived from a probabilistic survey will have a known level of uncertainty associated with them. Without the error bars around a summary statistic, there is no way to know whether an observed change represents simple variability or a true change in resource condition. This type of information allows management actions to be prioritized according to, for example, regions or resource types that show the greatest decline or poorest status.

EMAP's pilot studies along with numerous regional EMAP, or REMAP, studies conducted at smaller spatial scales were designed to demonstrate how a probabilistic survey design could support emerging state monitoring programs that had previously lacked adequate assessment programs. Because EMAP was envisioned as a template for states, it is somewhat ironic that a defining aspect of the program, probabilistic sampling, has become controversial at the state level. State resource agencies are expected to manage surface waters at the local level, not just assess the general condition of the resource (Yoder and Rankin, 1998). Therefore, from a localized viewpoint, the coarse resolution of data associated with sites selected at a regional scale may not provide information on particular areas of interest. Moreover, sites of particular interest, such as permitted sites or sites with known problems, are unlikely to be selected randomly. Thus, very degraded sites may avoid the TMDL process which is designed to rehabilitate those sites by identifying their causes of degradation (NRC, 2001). From a state's point of view, then, probabilistic sampling can seem like an additional monitoring requirement rather than a design to simplify current monitoring.

Alternative sampling approaches select sites according to other specific (i.e., nonrandom) criteria. These approaches may target sites associated with specific management activities or types of disturbance. Many state programs adopt this approach over probabilistic sampling. Other states combine the two approaches by using a probabilistic design to identify problem areas and a targeted design to focus within those areas (USEPA, 2003). A similar conflict between assessments at the regional scale vs. evaluation of a set of sites with known properties arose for the MAIA study within the context of testing biological indicators. Random sampling tends to miss sites with little or no human disturbance as well as sites with very high levels of human disturbance; therefore, the concern arose that randomly selected sites would not provide a broad enough range of conditions to test biological indicators. As a consequence, targeted sampling of additional reference and impaired sites was used to supplement the MAIA probabilistic data. In the end, this supplemental sampling was largely unnecessary for MAIA because many of the targeted reference sites failed to meet the criteria for reference condition while the probabilistic survey included a sufficient number of minimally disturbed and severely degraded sites for metric testing.

A probabilistic sampling design was the best choice for MAIA

Given the size of the sampling area and the scope of the questions asked for the MAIA study, most agreed that randomization of site selection was a necessity. The MAIA pilot was designed to answer questions about the overall condition of surface waters within a large geographic area. These types of questions are of greatest interest to USEPA for planning programs and allocating funds to support state and regional programs. Other sampling designs were not appropriate for an EMAP project because they fail to yield an unbiased estimate of conditions across the entire region (Chart 1). Unless one samples every site (census sampling), selecting sites randomly is the only method for inferring regional condition from a smaller set of sites (Overton and Stehman, 1996).

Chart 1: Types of sampling.

Census sampling: samples every site, no inference needed.

Probability sampling: randomly selects sites from the set of all possible sites within a region.

Targeted sampling: selects sites based on some known aspect, for example sites might be sampled downstream of a point source to evaluate the extent of its influence.

Judgment sampling: similar to targeted sampling, sites selected according to some scientific or observational criteria. (See example in Stoddard et al., 1998).

Convenience sampling: samples selected according to where sampling is possible, e.g., due to private access issues. (See Olsen et al. [1999] for further discussion).

Synoptic sampling: typically implies sampling within a short period of time, perhaps to assess variables that are time-dependent, for example, chemical concentration along a stream after a toxic release. Also implies breadth across a large area or concurrent sampling of multiple measures.

Haphazard sampling: Selection of sampling units may be casually described as “random”; however, if the selection process is not formally random, that is, based on the use of random numbers table or randomization algorithm, the method of selection is actually haphazard.

Census sampling, that is, visiting every site, was not possible on as large a scale as the MAIA study area. This is true of most regional monitoring programs. Selecting sites according to scientific criteria or professional judgment was not relevant because regional, rather than local, estimates of all indicators and stressors were desired. Any design that limited the set of sample sites to easily accessible locations (convenience sampling) or sites with known human impacts (targeted sampling) risked producing a biased picture of current regional condition and were rejected (Stevens, 1994; Olsen et al., 1999).

From a statistical point of view, any sampling approach other than census or random sampling will be technically "haphazard," that is, potentially biased and not representative of the larger unsampled population. Haphazard or targeted sampling designs are not inherently bad or wrong. Their primary drawback for monitoring is that the results from sampled sites cannot be extended to any unsampled sites.

Free information is cost-effective

Why sample randomly? The key point is free information. If site selection is random and all sites are sampled with an equal or known probability, then information from the sampled sites can be used to infer the condition of sites not sampled. Thus, results based on a random sample of sites can be scaled up to the entire population of sites within a region, as long as each site in the region could have been included in the sample. The knowledge obtained through random sampling about the unsampled sites is essentially free information. The only other way to know something about every site in a region would be to sample each one. Sampling all the sites is much more expensive than randomly sampling a few.

If sites are selected according to any method other than randomization, then statistics and conclusions derived for those sites can be safely applied only to those sites sampled. Generalizing from a nonrandom sample to the larger unsampled population of sites typically yields biased conclusions. The statistical truth of this fact may not be compelling; in fact, it seems somewhat counterintuitive that a large enough sample size would fail to be representative of general conditions in a region. Nonetheless, numerous large-scale studies based on real data have demonstrated the inevitable bias in nonrandom sampling schemes, even when hundreds of sites were sampled (Paulsen et al., 1998; Stoddard et al., 1998; Peterson et al., 1999).

In the controversy over how probabilistic sampling fails to meet state needs, the coarse scale of sampling is often emphasized as inadequate at the state level (White and Merritt, 1998; Yoder and Rankin, 1998). One relevant concept that is often missed in the discussion is that the scale of probabilistic sampling can be altered to fit a state's more local assessment needs. Both USEPA and state programs ask the same types of questions, but they ask them at different scales. USEPA wants to know if surface waters in a state or region are getting better or worse. A state manager or citizen group wants to know if a specific stream, or basin, is getting better or worse. Same question, different scale. The EMAP concept of random sampling and free information could be applied at a different spatial scale such as a basin or a watershed to infer the condition of other sites or reaches not sampled within the basin or watershed. The EMAP method used to define and select sampling units is flexible and was intentionally designed to be applicable at different spatial scales including the conterminous U.S., geographic regions covering multiple states, or smaller regions defined within state boundaries (Herlihy et al., 2000).

Our statistical ability to detect change in regional condition is also independent of the size of the region when regional condition is summarized in terms of the percentage of sites meeting (or failing

to meet) specific criteria. For example, suppose in the first year of sampling that 50 sites are randomly selected and sampled. Of these, 50% (25 sites) fail to support their designated uses and are considered impaired. If 50 new sites are sampled the next year, a change of greater than 12% in either direction represents a significant change (at 90% confidence) in stream condition within the sample region (USEPA, 2003). These types of summary statistics based on proportions along with their estimates of precision would apply to a probabilistic survey of any 50 sites, whether sampled within a county or within a state.

Reference sites did not always meet criteria for reference condition

Expectations for biological indicators such as multimetric indexes are based on observed conditions at undisturbed or minimally disturbed locations. These reference sites represent a standard for what the biological assemblage would look like in the absence of human influence (Hughes, 1995). Given a probabilistic sampling design, the concern developed during the course of the MAIA pilot that only moderately disturbed sites would be included in the data because a moderate level of human influence was typical throughout the region. As a consequence, examples from the ends of the spectrum, that is, sites with minimal human influence and extreme degradation, would be missing from the data. Therefore, to supplement the EMAP design data, local biologists helped select reference and impaired sites based on their experience and knowledge of the region, i.e., according to their best professional judgment (Gerritsen et al., 1994). The idea was to use these "hand-picked" sites to test metric response to disturbance and then use the probabilistic sites to estimate regional status and trends.

In the meantime, researchers developed a concept of reference condition and defined specific criteria for the Mid-Atlantic (Hughes, 1995; Waite et al., 2000; Klemm et al., 2002). Ironically, most of the hand-picked reference sites (44 out of 60, or 73%) failed to meet the independently-established criteria for reference condition. The lesson learned was that best professional judgment should always be confirmed with objective criteria.

The criteria for reference condition included acid-neutralizing capacity (ANC) > 50 μeq/L, total P < 20 μg/L, total N < 750 μg/L, chloride (CL-) < 100 μeq/L, SO_4 < 400 μeq/L, and mean RBP habitat scores >15 (based on USEPA's Rapid Bioassessment Protocol; Barbour et al., 1999). Reference sites had to satisfy all criteria. Five of the six criteria were based on water chemistry and selected for their close association with specific human activities in the watersheds. Chloride increased along with development, total N and P were associated with agricultural intensity, low levels of ANC indicated acid rain, and SO_4 was related to mine drainage (Herlihy et al., 1990; Herlihy et al., 1993; Herlihy et al., 1998). Therefore, extreme values of these chemical measures also indicated other potential stressors and disturbances associated with these human activities.

The "hand-picked" reference sites selected according to best professional judgment included sites with strong indications of intense human disturbance (Figure 1). For example, values for total N > 3000 μeq /L are usually considered high and are typical of urban or agricultural land use. Chloride values > 300 μeq /L also represented some of the highest values in the entire data set and were associated with relatively high levels of urban development. Similarly RBP values below 12 represented obvious signs of human influence nearby or within the sample reach.

Although the first years of random site selection may not have included a large enough sample of reference sites, successive years of sampling did provide a broad enough range of human influence to test and develop biological indicators. In fact, over time, researchers moved away from the idea of a simple comparison of reference vs. impaired sites and adopted a more sophisticated approach that evaluated metric response across multiple gradients of human disturbance. Thus, the probabilistic survey design yielded an adequate range of conditions for testing metrics. Furthermore, objective definitions of site condition and impairment represented a more reliable approach for testing and selecting biological indicators than did definitions of site condition based on best professional judgment.

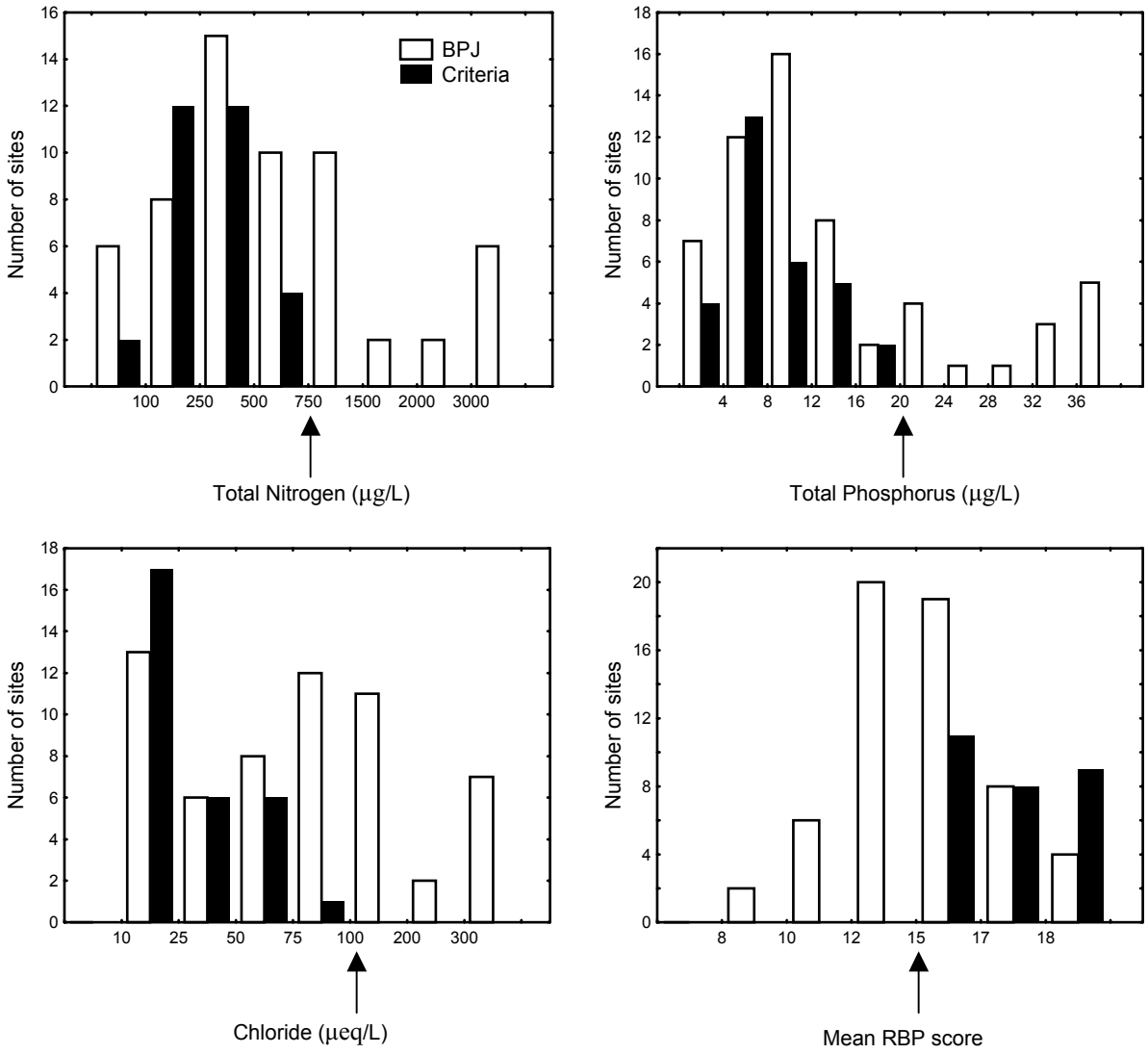

Figure 1. Comparison of reference sites selected on the basis of "best professional judgment" (BPJ; light bars) with sites selected according to reference condition criteria for total N, total P, chloride, and mean RBP score (dark bars). Arrows indicate criteria defined for reference condition; higher values to the right of the arrows indicate greater disturbance; for RBP scores higher values indicate better condition. The broad range of values for BPJ sites indicate that highly disturbed sites were initially included as reference sites.

III. THE PERILS OF DATA MANAGEMENT

Every scientific analysis is founded on the format and reliability of its underlying data. Initial decisions about how to store and format data will influence how or if data are ever used for scientific analysis. What agency doesn't have data languishing on diskettes in file drawers? The lesson learned from the MAIA pilot was that time spent on data management was never wasted and that the time required was usually more than expected.

During the course of the MAIA study, several hundred sites were sampled for hundreds of variables over six years. Some variables were sampled once per site (e.g., land cover and use); other variables were sampled on every possible occasion (e.g., water chemistry); and still others were sampled on only some site visits (e.g., fish and habitat). For invertebrates and algae, two samples were collected during many site visits: one from pools and the other from riffles. Most sites were visited only once, but some sites were sampled repeatedly, either during one year or in successive years, in order to estimate the variability associated with site scale data. As a result, the data set was huge and complex, and data management was an issue that was carefully considered from the beginning (Hale et al., 1998; Hale et al., 1999). However, in spite of careful planning, data analysis was slowed by months due to mishandling of data files that resulted in incomplete and corrupted data.

Many agencies struggle with data management choices made early in their programs that restrict their current ability to analyze and interpret those data. Although many state water monitoring programs are smaller in scale than MAIA, the amount of data collected increases each year and databases are much larger than they were ten or even five years ago. Data management takes on a life of its own when more than about 200 sites are sampled for more than 50 variables. Most state sampling programs now exceed this amount of data for monitoring under the Clean Water Act. Data management lessons learned in the Mid-Atlantic will become increasingly relevant to state programs as data accumulate.

Different "names" for the same site caused confusion

The MAIA data set was large and complex; therefore, it was not possible to put all the data in a single file. Multiple files of related data meant that site names were extremely important for matching data across files. Unfortunately, different types of data required different site information to uniquely identify data from a single site visit. Thus, it was not possible to create a single unique code for matching site visits across files that would correctly match all variables.

For land use data, only the site name was needed because satellite data were only recorded once for each site during the course of the project. For fish and chemistry data, the site name, year, and visit number within year were needed to identify data from a unique site visit. For diatom and invertebrate data, the site name, year, visit number and habitat type (pool vs. riffle) were needed. As a consequence, files were easily corrupted when smaller files were merged without using all the variables necessary to match site visits correctly. An incorrectly merged file could have twice as many entries for some variables or data from one site visit assigned to another. This issue will be complicated for any monitoring program when the same site has different types of data associated with it and the data are collected at different times. The lesson learned was that more information should have been included with the data to identify unique sampling occasions.

Original data must be archived

Biological data evolve from field sheets for fish or laboratory bench sheets for invertebrates and algae with coded taxonomic identifiers; to files with taxa names, counts and natural history information; to calculated metrics. With each new generation of the data file, the original files may appear crude in retrospect and the tendency was to lose track of original files with confusing formats when newer versions were created. For the MAIA pilot, original files were the only way to catch major errors in later versions of the data. The lesson learned was to archive the original field or bench sheets along with the first generation of electronic data in a way that the data will not be changed or lost.

Different research groups tended naturally to follow different paths in evolving their data files. When the different researchers came together, their results sometimes disagreed either due to errors or different decision rules. Checks against the first generation of the data were often the only way to resolve discrepancies between researchers. Errors crept into data sets, for example, when files were merged incorrectly. Other problems arose when large files were truncated without warning by a spreadsheet or statistical program because the number of variables exceeded the number of columns permitted by the software.

During the course of the project, rules for combining invertebrate taxa evolved independently among different researchers. Similarly for periphyton, rules for calculating the relative abundance of soft algae and diatoms also evolved. In both cases, the first generation data files were required to create a correct master file and to calculate metric values consistently. For invertebrates, discrepancies can arise if one mayfly in a sample is identified to genus while another mayfly can only be identified to the family that includes that same genus (either due to damage or being an early instar). Should the two mayflies be counted as two distinct taxa or only one? If the family typically has only one genus, then it's an easy decision, they should be counted as one taxon. But if the family has many genera or the genus identified is known to be rare, then perhaps they should be counted as two taxa. For the MAIA data, much time was spent defining consistent rules and creating a master file that used the taxonomic rules to combine taxa. For periphyton, soft algae and diatoms are identified with different laboratory methods; consequently, the taxa counts were originally saved in separate files. In order to combine the taxonomic data for both types of algae, the relative abundance of each type in the original sample had to be considered. Bench data related to sample volumes was retrieved from the earliest data files to combine the information.

Without access to the original archived data, much of the data would have been irretrievably lost. Although the archived data prevented this, a larger lesson learned from MAIA was that an information manager should be an integral part of any scientific team for projects of this scale. Furthermore, the hours dedicated in the beginning of the project to ensure data were correctly stored were dwarfed by the hours spent after the fact trying to repair or retrieve corrupted data.

File structure mattered

Three key points contributed to the usefulness of the MAIA data and the successful analysis of the data in so many publications. Simple files were shared via an Internet server, file formatting was consistent across files, and metadata files were provided to describe the variable names. In addition, the data were organized into files that lent themselves to statistical analysis. File structure logically anticipated the type of analysis that was likely to be applied to a particular type of data.

When selecting the file structure for different data types, EMAP database designers distinguished between “vertical” and “horizontal” files (Hale and Buffum, 2000). A horizontal file, has a single row, or case, for each unique site visit and a single column for every variable recorded for that site visit. In contrast, a vertical file lists different types of data in a single column and may use multiple rows for a single site visit. Vertical files save space when many of the variables contain no information for a particular site visit. Both are appropriate in different situations.

For the MAIA study, vertical files were used to list fish, invertebrate, diatom, and soft (non-diatom) algal taxa collected at each site (Figure 2). The taxon code and name were listed along with the number of individuals in that taxon. A horizontal file structure would have headed each column with the name of one taxon and yielded a file with most of the cells blank, because not all taxa are found at a particular site. For diatoms, the number of taxa averaged 29 per site out of approximately 950 possible. Most spreadsheet programs do not allow so many columns, so rather than split the data across multiple files, a vertical file structure was used. Taxonomic information was also too cumbersome to include in the data file and was instead recorded in a companion file of metadata that listed phylum, class, order, family and other taxonomic details once for each species code.

Data: List of taxa found at each site

SITE CODE	VISIT NO	YEAR	SAMPLE TYPE	TAXA CODE	TAXON	COUNT
DE750S	1	1994	Pool	BAACBIO	Achnanthes bioreti	7
DE750S	1	1994	Pool	BAACMNU	Achnanthes minutissima	130
DE750S	1	1994	Pool	BAANVIT	Anomoeneis vitrea	2
DE750S	1	1994	Pool	BACBAMP	Cymbella amphicephaia	2
DE750S	1	1994	Pool	BAEUBIL	Eunotia bilunaris	2
DE750S	1	1994	Pool	BASYRURU	Synedra rumpens v rumpens	16
DE750S	1	1994	Pool	BATAFLO	Tabellaria flocculosa	94
MD003S	1	1995	Riffle	BAACMNGR	Achnanthes minutissima v gracillima	5
MD003S	1	1995	Riffle	BACBAMP	Cymbella amphicephala	1
MD003S	1	1995	Riffle	BASYNAN	Synedra nana	1
MD003S	1	1995	Riffle	BASYULUL	Synedra ulna v ulna	5
MD003S	1	1995	Riffle	BATAFLO	Tabellaria flocculosa	4
ETC.						

Metadata: Taxonomic nomenclature and authority for each taxon

TAXA CODE	PHYLUM	GENUS	SPECIES	V or f	SUBSP	AUTHORITY
BAAC	Bacillariophyta	Achnanthes	spp.			
BAACBIA	Bacillariophyta	Achnanthes	Biasolettiana			(Kützing) Grunow
BAACBIO	Bacillariophyta	Achnanthes	bioreti			Germain
BAACBITH	Bacillariophyta	Achnanthes	biasolettiana	v	thienemannii	(Hustedt) Lange-Bertalot
BAACDAU	Bacillariophyta	Achnanthes	daui			
BAACDEAL	Bacillariophyta	Achnanthes	deflexa	v	alpestris	Lowe & Kociolek
BAACDEF	Bacillariophyta	Achnanthes	deflexa			Reimer
BAACDEL	Bacillariophyta	Achnanthes	delicatula			
BAACDESE	Bacillariophyta	Achnanthes	delicatula	spp	septentrional	(Østrup) Lang-Bertalot
BAACDIS	Bacillariophyta	Achnanthes	distincta			Messikommer
BAACEXA	Bacillariophyta	Achnanthes	exigua	v	elliptica	Grunow
BAACEXEL	Bacillariophyta	Achnanthes	exigua			Hustedt
BAACEXI	Bacillariophyta	Achnanthes	exilis			Kützing
BAALPEL	Bacillariophyta	Amphipleura	pellucida			(Kützing) Kützing
BAAMOVA	Bacillariophyta	Amphora	ovalis			(Kützing) Kützing
BAAMPED	Bacillariophyta	Amphora	peducilus			(Kutzing) Grunow
BAAMSUB	Bacillariophyta	Amphora	submontana			Hustedt
ETC.						

Figure 2. Example data file for diatom samples from MAIA sites and companion metadata file with taxonomic details keyed by taxa codes. The upper table is an example of a vertical file, with more than one row per site visit and taxa names collapsed into a single column rather than one column per taxon.

Once calculated, metrics were stored as horizontal files where a single row of data contained all the information for each site visit (Figure 3). Each data file also had a companion file of metadata that explained the meaning, units, and derivation of the codes naming each variable, represented by one column of data, in the file. Other horizontal files were created that grouped sets of variables by topics such as fish metrics, water chemistry, habitat measures, and stressors (Hale et al., 1998).

Although the metadata files were somewhat minimal, they were a key component to the usefulness of the data across projects and research groups (Hale et al., 2000). Unfortunately, additional important information regarding how data were collected was not archived with the data. This caused some confusion at the time of analysis because sampling protocols changed through time. In short, all the attention paid to file structure and data organization was well worth the effort; in fact, more metadata and data description would have been useful.

Simple files were best

Data analysis for MAIA involved multiple institutions and investigators using different statistical software. Posting files on an Internet server was the most practical approach to sharing files between so many remote users. Hosting a searchable relational database that included all the data was an option, but these are typically slow and difficult for the host to maintain (Hale et al., 2000). Because researchers were typically interested in a subset of data, smaller, simpler files with variables grouped according to topic worked best.

Relational database programs (e.g., Access and Oracle) provide the opportunity to develop complicated data structures composed of multiple linked files. However, this approach assumes that users are working in a similar software environment, that is, using the same programs to analyze data or working from the same central database. Relational databases typically link multiple vertical files, thereby saving space by not repeating information that is identical for each entry. Although relational databases avoid some repetition within files, the unseen program code required to support the relationships between files typically takes up more space than a simple flat file with redundant information. Another advantage of a relational database is that data can be stored in one place and relationships between variables and files are encoded in the program so that individual users are not required to remember or derive these details; however, these relationships can be very complicated to code correctly and to maintain.

The MAIA data had to be accessible to many remote users with the intention of manipulating the data within a variety of software environments. Most statistical software packages expect flat files and cannot import the program information that a relational database uses to link variables across files. Flat files have all the information for a site visit in the same place, that is, in one file on the same row (Hale and Buffum, 2000). Thus, rather than create a complicated database structure from which data would have to be exported for analysis, data files were kept simple from the beginning so that they could be easily downloaded from the EMAP Internet site and quickly entered into the user's own statistical software.

Data: Example of a horizontal data file structure

STRM_ID	VISIT #	DATE COL	TEAM ID	PHSTVL	PHEQ	ANC	COND	COLOR	CA	MG	NA	K	NH4	CL	NO3	SO4	ALTD
DE750S	1	5/17/94	3	6.17	7.38	115	102.5	8	303.4	213.1	266.2	67	2.8	229	303	132	54
MD003S	1	5/15/95	3	6.91	7.57	207.2	90.6	13	528.9	147.2	84	20.2	0	44	30.3	443	36
MD005S	1	5/15/95	3	6.66	7.15	82.2	42.4	30	181.6	80.6	82.2	15.6	0	85	9.7	160	47
MD006S	1	5/22/95	3	7.42	7.9	435.2	99	13	391.7	231.2	263.6	27.4	0	310	72.3	77	12
MD008S	1	5/23/95	3	6.39	6.92	37.4	31.9	6	124.3	73.2	37.4	27.6	0	30	1.6	180	19
MD507S	1	5/25/93	1	6.82	7.48	154	53.2	3	272	107.8	53.1	20.7	2.8	79	44.3	200	10
MD508S	1	5/27/93	1	7.14	7.75	317	68.8	4	340.8	176	85.3	34	0.6	36	50.9	243	12
MD510S	1	6/1/93	1	7.08	7.47	153	62.2	7	237.5	179.3	69.6	35.5	0	102	37.4	234	5
MD511S	1	6/10/93	1	7.11	7.73	286	121.8	7	367.3	261.6	394.5	47.8	0	429	16.5	323	8
MD512S	1	6/9/93	1	7.06	7.82	391	90.8	6	409.2	283	102.2	48.1	0	54	7	376	6
MD513S	1	6/7/93	1	8.44	8.66	3550	441	5	3263.5	1299.7	97.4	43	0	204	575	418	6
MD750S	1	5/16/94	3	7.12	8.3	1090	298.5	17	1462.1	804.5	418.5	67.8	2.2	696	375	401	12
MD751S	1	5/19/94	3	7.18	7.93	412	171.2	8	588.8	418.7	425.9	54.5	0	663	248	135	30
MD752S	1	5/31/94	1	6.72	7.77	295	142.3	2	673.7	380.9	124	50.9	0	32	7	851	1
MD753S	1	5/23/94	3	7.55	8.49	1670	251.5	11	1556.9	479.6	391.5	35	1.1	515	61.4	172	20
MD754S	1	5/18/94	3	6.66	7.45	145	62.7	5	183.6	167	151.4	26.3	0	166	7.2	224	20
MD755S	1	5/24/94	5	7.15	7.68	246	78	8	327.3	154.6	160.1	23.5	0	166	28.4	228	12
MD756S	1	5/23/94	5	7.93	8.23	895	142.3	8	728.5	394.8	226.6	27.9	0	242	60.3	168	19
MD757S	1	5/18/94	6	7.24	8.4	1450	279.3	17	1117.8	519.1	648.2	82.6	232.1	327	15.1	862	30
NY001S	1	6/22/95	2	7.4	7.83	385.5	115	5	408.2	168.6	416.3	22.2	1.1	469	4.7	156	13
NY002S	1	6/22/95	2	7.6	8.03	576	129	9	623.8	193.3	305.4	39.4	1.1	326	79.7	191	15
ETC.																	

Metadata: Description of column headings

Variable Name	Description
STRM_ID	Stream ID
VISIT_NO	Visit Number
DATE_COL	MMDDYY Date stream visited
TEAM_ID	Sampling crew
TRANSECT	Transect ID
PHSTVL	Closed System pH
PHEQ	Air-equilibrated pH
ANC	Gran Acid Neutralizing Capacity (μeq/L)
COND	Specific Conductance (uS/cm)
COLOR	Color (PCU)
CA	Calcium (μeq/L)
MG	Magnesium (μeq/L)

Variable Name	Description
NA	Sodium (μeq/L)
K	Potassium (μeq/L)
NH4	Ammonium (μeq/L)
CL	Chloride (μeq/L)
NO3	Nitrate (μeq/L)
SO4	Sulfate (μeq/L)
ALTD	Total Dissolved aluminum (μg/L)
ALDS	PCV reactive (monomeric) aluminum (μg/L)
ALOR	Nonexch. PCV (organic) aluminum (μg/L)
DIC	Dissolved Inorganic Carbon (mg/L)
ETC.	

Figure 3. Example of a horizontal file structure for water chemistry data. Stream ID, visit number and date identify a unique sampling event represented by a single row in the data file. A companion metadata file lists each variable name, its description, and units of measure.

IV. LINKING HUMAN DISTURBANCE TO BIOLOGICAL CHANGE

Like any scientific inquiry, understanding how human disturbance influences, alters, and degrades biological processes is an iterative process. Hypotheses are proposed on the basis of current understanding; theories are modified in the wake of results; and new insights drive subsequent hypothesis testing. As knowledge accrues, connections between predictions and results tighten, and new studies present fewer surprises. A paradigm forms.

The development of biological monitoring tools has followed this type of scientific process. After 20 years of testing, consistent patterns have emerged (see literature review in Barbour et al., 1999; Karr and Chu, 1999; Karr et al., 2000). The same biological measures tend to correlate with human disturbance in very different geographic settings, e.g., the number of taxa decline, tolerant taxa dominate, and taxa with unique habitat requirements disappear. Yet when results closely match expectations, concerns arise regarding "circular reasoning." An argument based on circular reasoning is one in which the conclusion is embedded in the premise, as for example, in the statement, "decline in mayfly taxa richness is a good indicator of biological disturbance because we find many types of mayflies at undisturbed places." Without a direct causal link, the concern is that metrics may be selected as indicators of human disturbance simply because they are correlated with human disturbance (Suter, 2001). That is, biological indicators may be unrelated to specific biological or societal values.

Because biological systems are complex and human disturbance is multidimensional, single causes and mechanisms of impairment are difficult to isolate; as a result, much of the evidence for human degradation of natural resources is correlative. In such situations, although the path to causality is blocked by the inability to perform controlled experiments and use statistical inference, logical argument (or weight of evidence) constructed according to a recognized set of rules can be used instead (Beyers, 1998). In fact, this approach typically yields a stronger case because researchers consider alternative explanations explicitly, rather than assuming they do not operate. Results from the Mid-Atlantic illustrate how a causal argument can be constructed to support the idea that human disturbance causes biological change.

Addressing concerns about circular reasoning

The data for MAIA were not collected with the intention of demonstrating a causal link between human disturbance and biological degradation; however, the expectation of a cause and effect relationship is implicit in EMAP's survey design and project goals. The process used to test and select metrics did, however, support the type of structured logical argument reviewed by Beyers (1998) for establishing a causal connection between human influence and biological change. In fact, results from the MAIA pilot supported six of Beyers' ten criteria (Chart 2).

Researchers in the field of epidemiology face a similar challenge in defining causality when linking a specific disease with its infectious agent. Rules developed within that field can be applied in an environmental context because the situations are parallel (Beyers, 1998). In epidemiology, patients develop a disease; an investigator cannot randomly infect patients with a variety of infectious agents to see which one causes the disease. Analogously for environmental studies, human development cannot be "applied" at will to see how a place will respond. Furthermore, treatments cannot be replicated in either case. Each patient is unique in terms of medical history and life style as is each watershed unique in terms of its geological formation, hydrological structure, size, and climate (Hurlbert, 1984; Heffner et al., 1996).

Chart 2. Ten criteria for constructing causal arguments (modified after Beyers, 1998).

1. **Strength:** a large proportion of sampling units are affected in exposed areas compared with reference areas
2. **Consistency:** the association has been observed at other times and places
3. **Specificity:** the effect is diagnostic of exposure
4. **Temporality:** exposure must precede the effect in time
5. **Dose-response:** the intensity of the observed effect is related to the intensity of the exposure
6. **Plausibility:** a plausible mechanism links cause and effect
7. **Evidence:** a valid experiment provides strong evidence of causation
8. **Analogy:** similar stressors cause similar effects
9. **Coherence:** the causal hypothesis does not conflict with current knowledge
10. **Exposure:** indicators of exposure must be found in affected organisms

Six of Beyers' ten criteria were relevant for constructing a logical argument for the causal connection between human disturbance and biological decline in Mid-Atlantic streams.

- First, the association between proposed cause and effect was strong. The majority of sites with human disturbance in the watershed had lower values for biological indexes than did the reference, or minimally disturbed, sites (Figure 4). In addition, indexes were significantly correlated with independently derived measures of human disturbance (Table 1).
- Second, the observed association with biological metrics and indexes was consistent with the results observed by other scientists in similar situations. Klemm et al. (2002) found that many of the same invertebrate metrics associated with disturbance at the regional level had also been selected for their response to disturbance by state programs at a more local level. McCormick et al. (2001) related fish metrics selected for Mid-Atlantic streams to functionally similar metrics selected in other regions.
- Third, because evidence of human disturbance tends to persist, it was reasonable to conclude that exposure to disturbance preceded biological change.
- Fourth, graphed dose-response relationships illustrated that the biota changed in proportion to the intensity of disturbance.
- Fifth, the hypothesis that human disturbance causes biological degradation does not conflict with existing knowledge or experimental evidence (see examples in Hudson and Cibrowski, 1996; Wallace et al., 1996; Lemly, 2000; Richardson and Kiffney, 2000; Mebane, 2002).
- Sixth, tissue analysis found chemicals associated with human development, e.g., DDT and mercury, in fish. These contaminants were also correlated with an increase in the number of fish species tolerant of chemical pollution.

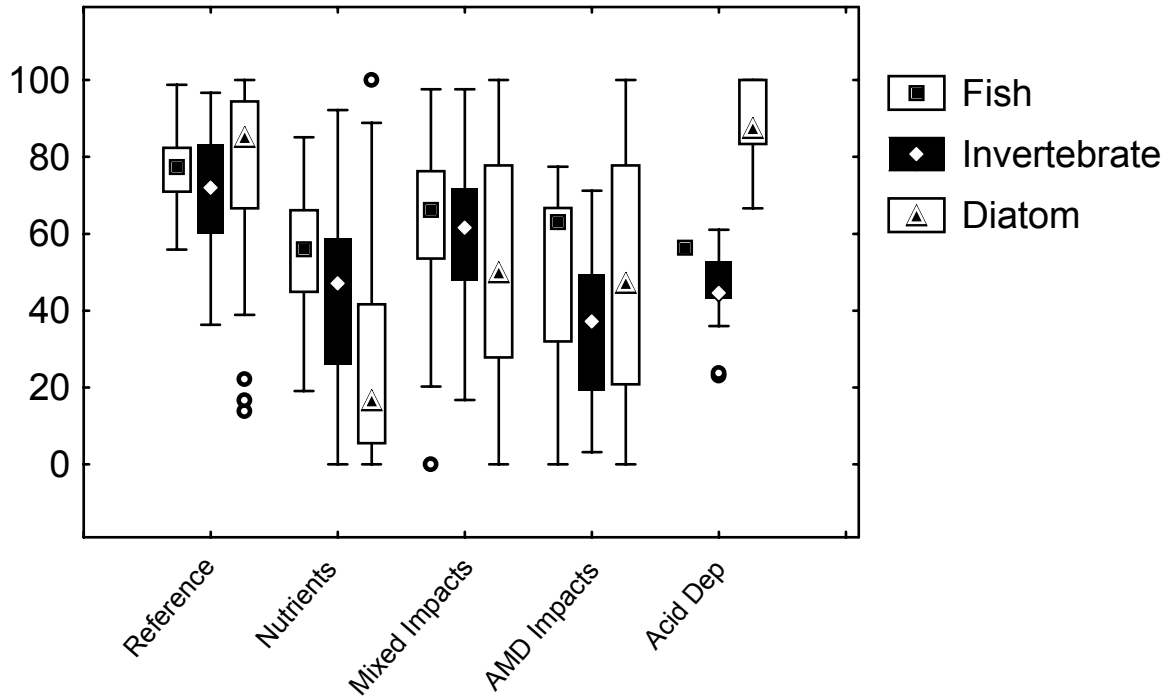

Figure 4. Multimetric index values for fish, invertebrates, and diatoms as a function of types of human disturbance. All index values were higher at reference sites. Diatom index values were lower for sites with high nutrients. Diatom index values were higher than invertebrate index values for sites with acid deposition.

The remaining four criteria (numbers 3, 6, 7 and 8 in Chart 2) require that an observed effect (metric value) be diagnostic of exposure (human disturbance); a plausible mechanism of action exists to link cause and effect; controlled experiments support causation; and analogous responses are associated with similar stressors. A broader survey of the literature could add further examples under these headings to the overall argument for causality. For example, Yoder and DeShon (2002) used metrics to diagnose disturbance type, and Richardson and Kiffney (2000) present experimental evidence for the effect of heavy metals on invertebrates. Nonetheless, the examples above serve to illustrate how the logical construction of an argument for causality represents an alternative to doubts regarding circular reasoning.

Table 1. Spearman's correlation of three multimetric indexes (based on fish, invertebrates, and diatoms) with selected measures of human disturbance. All correlation coefficients were significant; only values > 0.3 (or < –0.3) are shown. Measures of human disturbance were related to (1) nutrients (total nitrogen [N], total phosphorus [P], ammonia [NH_4]); (2) acidity (acid neutralizing capacity [ANC] and sulfate [SO_4]); (3) sediment (turbidity [Turb], percentage of sand and fine sediments [%S_F], and pebble size corrected for stream power [PbSz]); and (4) channel and riparian condition (riparian vegetation [RVeg], sum of all disturbance types within the riparian area weighted by proximity to the stream [RDist], and average of measures from a rapid habitat protocol [RBP]). Measures of general disturbance included chloride (CL); Bryce et al.'s [1999] disturbance categories derived from watershed and riparian measures of disturbance (Bryce); and the sum of urban, agricultural, and mining land use within the watershed (%Dist).

Measure	Fish	Invertebrate	Diatom
N	-0.45	-0.32	-0.54
P			-0.61
NH_4	-0.33	-0.36	-0.32
ANC		-0.33	-0.53
SO_4	-0.34		
Turb		-0.33	-0.39
%S_F		-0.54	-0.39
PbSz		0.43	0.33
RVeg			0.30
RDist		-0.35	
RBP		0.42	0.36
CL	-0.45	-0.31	-0.55
Bryce	-0.39	-0.57	-0.54
%Dist	-0.33	-0.40	-0.54
Total	**6**	**11**	**12**

Metric testing included safeguards against circular reasoning

One risk associated with using correlative data to demonstrate connections between human disturbance and biological degradation is that the observed correlation may be due to spurious correlation with another underlying cause, such as elevation or watershed size that drives both biology and patterns of human settlement. Or, the observed correlation may be due to some additional factor that was not considered. When developing biomonitoring tools, the goal is to select biological indicators that vary only as a function of human disturbance and are immune to variability associated with natural physical or geographic features. Unfortunately, humans are biota, too, and their land use patterns tend to follow landscape features, thus confounding human activities with physical features. The challenge for the MAIA project was to demonstrate that human disturbance was the most likely agent of biological change.

A variety of safeguards helped reduce the probability of drawing unsubstantiated conclusions from the MAIA data analysis. Five approaches were used to isolate the relationship between human disturbance and biological change from other confounding influences.

- First, site selection was randomized across a large geographic area to ensure that the sample was representative of the entire population of possible sites. Unbiased selection of sites provides some protection against confusing the effects of human disturbance with other, natural features (Stewart-Oaten et al., 1986; 1992).
- Second, measures of disturbance were selected independently of the biological metrics. Bryce et al. (1999) present an integrated definition of human disturbance for the Mid-Atlantic region derived without consideration of biological indicators.
- Third, metrics were tested in multiple years, or part of the data set was reserved to test the final indexes as an independent test (McCormick et al., 2001; Fore, 2002b; Klemm et al., 2003). In this way the observed relationships were demonstrated to be consistent across years.
- Fourth, all metrics were tested for correlation with multiple gradients of human disturbance rather than for their simple ability to distinguish between one set of minimally disturbed, or reference, sites and severely degraded, or impaired, sites. Thus, the metrics selected were consistently associated with different aspects of human disturbance.
- Fifth, potential confounding factors such as watershed area or elevation that could underlie patterns of both human influence and biological condition were explicitly tested. Where necessary these effects were removed, for example, when fish taxa richness metrics correlated with watershed area. In this way, metrics were selected for their association with disturbance rather than other natural features.

Patterns of human disturbance were complex

The MAIA pilot had the luxury (and the curse!) of data for practically any variable that has ever been recorded to evaluate water resources. Dozens of variables related to water chemistry, metals, nutrients, fish tissue contaminants, habitat, channel morphology, geographic features, human census data, satellite land cover and use, and specific point sources were included in the data set. Hundreds more were derived from the data collected. The hope was that such a complete record of human activity would provide a clear picture of human influence and disturbance within a watershed. The reality was that the different measures of disturbance tended to tell their own story; that is, different measures were associated with specific types of human activity. Therefore, disturbance measures were not necessarily correlated with each other because not all activities were present in every watershed. As a consequence, one of the primary challenges for the MAIA pilot study was to determine which variables most accurately characterized human influence.

During the process of developing biological indicators, much discussion surrounded the choice of appropriate measures of site condition for metric testing. A primary lesson learned in the Mid-Atlantic was that no simple method existed to quantify human influence in such a complex landscape with such a long and varied history of human activity (Herlihy et al., 1998; Bryce et al., 1999, McCormick et al., 2001). A missing piece from this project was a comprehensive study linking the types of human activities (e.g., mining or agriculture) with their specific stressors (e.g., SO_4 or nutrients). Given the tendency to find multiple types of disturbance in each watershed, a

clear picture may not have been possible. Nonetheless, a better understanding of which of the many measures of disturbance tended to vary together along with a better understanding of which measures were related to natural geographic or landscape features would have helped clarify metric response to disturbance. For example, diatom metrics were correlated with elevation but so was disturbance because towns and farms tended to be found at lower elevations (Fore, 2002b).

Examination of a correlation matrix of all the variables related to site condition (too large to show here) revealed patterns of correlation among related variables. Sets of variables could be identified that seemed to measure similar or underlying processes. Variables related to water chemistry, nutrients, and water quality tended to be significantly correlated with each other. Other sets of variables related to channel structure, fish cover, or the condition and extent of riparian (streamside) vegetation showed similar patterns of higher correlation among related sets of variables. In contrast, some groups of variables showed little correlation across groups—for example, measures of riparian vegetation did not tend to correlate with measures of water chemistry.

Integrated measures of disturbance were better predictors of index values

In general, specific stressors tended to be more highly correlated with integrative measures of human disturbance than they were with similar measures that measured only a single aspect of disturbance. For example, turbidity, percentage of sand and fine, pebble size, riparian vegetation condition and riparian disturbance were correlated with one or two of each other, but all five were correlated with a habitat index developed to integrate measures of site condition at the reach scale (Table 2). A similar pattern was observed for water chemistry measures such as total N, total P, NH_4, and SO_4 that showed fewer significant correlations with each other than with integrative measures that summarized human influence at the watershed scale. Bryce et al.'s (1999) condition classes were derived from an analysis of patterns of human land use within the watershed and the observed riparian condition at 102 sites. All the listed measures correlated with Bryce et al.'s index of disturbance. The percentage of disturbed land was the sum of land in the upstream watershed used for agriculture, urbanization or mining; all but one of the uni-dimensional measures correlated with this measure.

Similarly for biological indicators, multimetric indexes for all three assemblages showed a higher correlation with integrative measures of disturbance than with specific stressors (see Table 1 on page 18). One chemical measure, chloride, was a strong indicator of general disturbance and also highly correlated with all three biological indexes (Herlihy et al., 1998).

Thus, measures of disturbance that integrated measures of site condition over multiple spatial scales tended to better capture the cumulative effects of human influence. This result supports the idea that much of the scatter observed in plots of biological measures against human disturbance gradients is in fact associated with the x-axis: one-dimensional measures of disturbance simply fail to capture the cumulative influence of human activities on the biota (Karr et al., 2000).

Table 2. Spearman's correlation matrix for measures of human disturbance that were used to test biological metrics for the MAIA study; only correlation coefficients > 0.3 (or < –0.3) are shown. (See Table 1 for description of variables.)

Measure	N	P	NH_4	ANC	SO_4	Turb	%S_F	PbSz	RVeg	RDist	RBP	CL	Bryce	%Dist
N	—	0.50	0.34	0.39								0.51	0.38	0.67
P	0.50	—	0.36	0.31		0.57	0.44	-0.34				0.35	0.30	0.50
NH_4	0.34	0.36	—									0.41	0.36	0.32
ANC	0.39	0.31		—	0.35					0.33		0.45	0.47	0.54
SO_4				0.35	—							0.45	0.39	
Turb		0.57				—	0.47	-0.38			-0.32		0.32	0.33
%S_F		0.44				0.47	—	-0.75			-0.39		0.55	0.49
PbSz		-0.34				-0.38	-0.75	—			0.31		-0.36	-0.34
RVeg									—	-0.65	0.60		-0.60	
RDist				0.33					-0.65	—	-0.35		0.45	0.30
RBP						-0.32	-0.39	0.31	0.60	-0.35	—		-0.77	-0.41
CL	0.51	0.35	0.41	0.45	0.45							—	0.58	0.68
Bryce	0.38	0.30	0.36	0.47	0.39	0.32	0.55	-0.36	-0.60	0.45	-0.77	0.58	—	0.47
%Dist	0.67	0.50	0.32	0.54		0.33	0.49	-0.34		0.30	-0.41	0.68	0.47	—
Total	**6**	**9**	**5**	**7**	**3**	**6**	**6**	**6**	**3**	**5**	**7**	**7**	**13**	**11**

V. METRIC TESTING

Candidate metrics are selected for inclusion in a multimetric index if they are biologically meaningful, consistently associated with human disturbance, not redundant with other metrics, and reliably and easily quantified from field samples (Karr and Chu, 1999; Jackson et al., 2000). Typically the approach selected for metric testing is limited by the type and variety of data available for quantifying human disturbance. For the MAIA study there were virtually no limitations because of the variety of variables measured. For any study relating biological change to the degradation associated with human activities, the most difficult step is identifying the independent measure of human disturbance. Independent measures of disturbance are needed to test for a consistent biological response across the range of possible conditions. Measures of human disturbance must be independently derived from biological data to avoid simply choosing aspects of disturbance or measures of biology that match our expectations. Because human influence is complex and human activities are multidimensional, the challenge revolves around how to integrate disparate measures of human influence into a single axis of human disturbance for metric testing.

Statistically, it is simplest to compare measures taken at impaired sites with the same measures taken at minimally disturbed reference sites (Barbour et al., 1996). When only a few measurements of disturbance are made, statistical testing may involve simple tests of differences in means (ANOVA) or association between one-dimensional measures of disturbance (regression or correlation; Miltner and Rankin, 1998). When multiple, related measures of disturbance are made, multiple regression may be used test the disturbance measures together. If, however, these measures are themselves correlated with each other (which they often are), multiple regression's assumption of independence is violated, and the results may not be robust (Loftis et al., 1991; Wang et al., 1998; Olden and Jackson, 2000). When many measures of disturbance are collected, measures may be combined using principal components analysis (Hughes et al., 1998; Norton et al., 2000). Other projects have used a ranking system to summarize information about human disturbance at different spatial scales (Bryce et al., 1999; Fore and Grafe, 2002).

For the MAIA pilot, the wealth of measures associated with human influence necessitated a larger discussion among the researchers involved. The disturbance measures used for metric testing were selected during a series of workshops sponsored by the USEPA. Through discussion and consensus, researchers derived a list that captured multiple aspects of human influence at different spatial scales (McCormick et al., 2001; Klemm et al., 2003). Reference and test conditions were defined to include a subset of least disturbed and most degraded sites (Waite et al., 2000). Measures of nutrient concentration were included to reflect the influence of agriculture and urban land uses; acidity was included to capture the effects of acid rain deposition and acid mine drainage; variables related to sediment and turbidity measured erosion; and measures of riparian condition summarized the physical disturbance near the site. In addition, Bryce et al. (1999) developed an integrative measure of human disturbance at the watershed and reach scale. The percentage of disturbed land in the watershed was also used as a summary measure.

Redundant testing of metrics against multiple measures of disturbance ensured that metrics were selected for their biological meaning rather than statistical chance. With a large list of candidate metrics and a single test for each, candidates may meet the criteria for metric selection because of chance alone. The probability of such chance selection equals the p-value selected for the test. Multiple tests against measures of different aspects of human disturbance avoided this pitfall and insured that the metrics selected represented meaningful and reliable indicators of biological change associated with human influence.

Simple criteria were used first to eliminate potential metrics

For each assemblage, a large number of candidate metrics were identified for testing (Stevenson and Bahls, 1999; McCormick et al., 2001; Klemm et al., 2003). Simple statistical rules were developed to shorten the long list of candidate metrics to a smaller number that were then considered more carefully.

During the first round of elimination, metrics were evaluated for their range of values. Metrics calculated on the basis of a small number of species might not yield an adequate range of values for calculating taxa richness or percentage relative abundance. For example, taxa richness metrics that could take on values of only 0 or 1 or percentage metrics that only had a range of 10% had too few potential values to distinguish between different levels of human disturbance. These candidate metrics were eliminated in favor of metrics with a broader range of values. For Mid-Atlantic streams, candidate metrics such as number of native cottid species and percentages of *Corbicula* were eliminated because their ranges of values were simply too small.

Statistical precision was no substitute for correlation with disturbance

If the variability of a candidate metric within individual sites is higher than its variability between all sites, then the measure is unlikely to detect differences in biological condition among sites (or differences at sites that change through time). Signal-to-noise ratios estimate a measure's ability to distinguish differences among sites from differences observed within individual sites. In this context, "signal" is defined as variability of a metric value across all sites and "noise" as variability over repeated visits to the same site during a single year (Kaufmann et al., 1999).

Although most metrics incorporated into multimetric indexes have high signal-to-noise ratios, i.e., small within-site variability compared to large between-site variability, a high signal-to-noise ratio alone does not guarantee that a candidate metric will be a meaningful indicator of biological condition. Metric values can be highly repeatable at individual sites but still be unrelated to human disturbance. Consider, for example, pool depth and embeddedness, two candidate metrics described by Kaufmann et al. (1999) in their assessment of habitat measures for MAIA. Pool depth is often considered an indicator of good fish habitat. It is expected to decline as erosion, dredging, and sedimentation fill pools, creating a homogeneous channel profile. Embeddedness was defined as an average of several substrate measures at the stream reach scale; it represented the proportion of the reach filled with sand and fine sediments.

Certainly statistical precision is a desirable property of a good metric, but statistical precision alone does not guarantee a predictable association with human disturbance. For the MAIA study, pool depth was very precise, with signal-to-noise ratio equal to 16. In contrast, embeddedness was more variable for repeat site visits with a signal-to-noise ratio of 1.9, which failed to meet the authors' suggested minimum value of 2. Embeddedness, however, showed a strong correlation with human disturbance. In contrast, pool depth was precise but not related to human disturbance. Embeddedness, though less statistically precise, was the better indicator of biological condition (Figure 5). Thus, statistical precision alone was not a good criterion for metric selection.

The most important signal for biomonitoring is a metric's response to human disturbance. The term "signal" in this case may, in fact, be misleading. The "signal" part of the signal-to-noise ratio is

actually just a measure of the observed range of values across all sites. In terms of biological assessment, the actual signal we are interested in detecting is the change in biological condition associated with human disturbance. In short, evaluation of the statistical properties of metrics should never substitute for actual metric testing against disturbance. These are two separate, though complementary, tests.

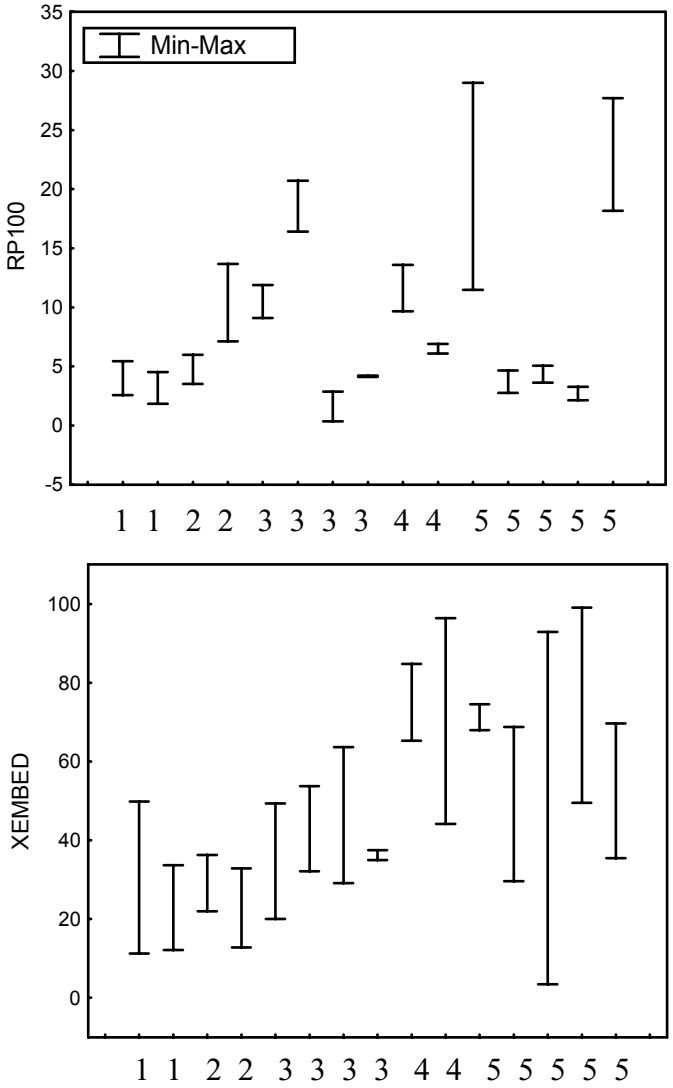

Figure 5. Ranges of values for mean residual pool depth (RP100, upper panel) and embeddedness (XEMBED, lower panel) for 15 sample sites sorted along the x-axis by disturbance class from least (1) to most (5) disturbed (Bryce et al., 1999). Vertical lines span the range of values recorded for two to six repeat visits to each site. Repeat visits to the same site yielded more similar values for RP100 than for embeddedness indicating greater precision (shorter vertical lines); however, embeddedness consistently increased with greater human disturbance while RP100 did not.

Watershed features were confounded with metric response to disturbance

A good monitoring tool must correlate with human disturbance and show little or no association with natural features. When human disturbance itself is associated with natural features, isolating the biological change associated with human disturbance can be tricky.

Watershed area was an example of this situation. Some fish taxa richness metrics tended to be correlated with watershed area because larger streams have more different types of habitat that support more species. To address this problem, metrics that correlated with watershed area were first regressed against watershed area using only reference sites. The residual values from this regression, that is, the metric values with the influence of watershed area removed statistically, were used to define a new version of the metric that was independent of watershed size. When the "corrected" metrics were significantly correlated with human disturbance, concerns about an underlying spurious correlation with watershed size were eliminated and the metrics were retained as good indicators of human influence. Watershed size was most important for fish, somewhat important for invertebrates, and did not influence diatom assemblages. The different sensitivity across assemblages to watershed area may reflect the relative range size of the organisms. Fish may travel throughout the watershed, invertebrates tend to stay within a reach or local stream, and diatoms may pass their lives on the same rock.

Metrics from different assemblage types were eliminated for different reasons

The list of plausible metrics proposed for testing in Mid-Atlantic streams was much shorter for fish (58), a bit longer for invertebrates (120) and much longer for periphyton (240). The initial list was shortest for fish because the greatest amount of metric testing has occurred for fish; invertebrates place a close second and periphyton a distant third. For periphyton, most of the candidate metrics represented untested hypotheses. Metrics were selected and eliminated for different reasons across assemblages.

The majority of candidate fish metrics were eliminated because they failed to correlate with disturbance (30 metrics; Table 3). In contrast, a larger percentage of candidate invertebrate metrics were consistently associated with disturbance. For the invertebrate index, many metrics that were significantly correlated with multiple measures of disturbance and that demonstrated good statistical properties were excluded because their correlation with one another exceeded 0.7. Approximately 25 metrics were eliminated for this reason, leaving relatively few (7) in the final index (Klemm et al., 2003). Because index precision increases as a function of the number of metrics, some caution may be warranted in eliminating good biological signal on the basis of statistical correlation (Fore, unpublished data). Results are not complete for periphyton, but based on preliminary results for a subset of diatom metrics, there will likely be many metrics to choose from that are significantly correlated with disturbance (Fore, 2002b). Additional criteria related to the type of environmental processes measured by the metrics and metric redundancy will probably be used to select metrics for inclusion in the final multimetric index.

Table 3. Numbers of candidate metrics tested for MAIA's fish and invertebrate multimetric indexes and a summary of the reasons for which they were eliminated. Starting with the total number of candidate metrics, the number of candidates listed in each column were eliminated because their values spanned an insufficient range, their signal-to-noise ratios were low (indicating low precision), they were redundant with other metrics, they failed to correlate with human disturbance, or their correlation with watershed size could not be corrected. This winnowing process resulted in fewer than 10 metrics included in the final indexes.

	Fish	**Invertebrates**
Total number of candidate metrics	58	120
Insufficient range	13	20
Poor signal/noise	2	66
Redundant	3	25
Fail to correlate with disturbance	30	2
Persistent correlation with watershed area	1	0
Number of metrics in final index	9	7

VI. Development and Application of Multimetric Indexes

The Clean Water Act requires all states to define water quality standards for their surface waters (Ransel, 1995). The standards include designated uses, criteria to protect those uses, and a prohibition against degradation of existing uses. States must assess the condition of water bodies and determine whether they support or fail to support their designated uses. Water bodies that fail are then listed as impaired (NRC, 2001). Currently, most states have narrative biological criteria in place, but are mandated to develop numeric criteria for biological condition (Karr, 1991; USEPA, 2002).

Multimetric indexes were created to fill this role as numeric assessment tools and are used by most states for this purpose (USEPA, 2002). The relevance and defensibility of the decision to list a site as impaired depends on both the biological meaning of index values and the statistical precision of the index. Decline in biological condition is a continuous process, and drawing a line of impairment, or defining biological criteria, at any single point along its range will inevitably be somewhat arbitrary. The most common approach for drawing this line defines impairment in terms of deviation from values observed for reference sites (Hughes, 1995; USEPA, 2002). A second approach uses statistical power analysis to calculate the change in biological condition an index can detect for a given statistical model and defines impairment in terms of that measurable change (Fore, 2002a). Other authors have recommended that better definitions of beneficial use should drive this process and that definitions of impairment should be tied to societal values, such as the support of salmonid spawning (NRC, 2001).

Biological criteria depend on the definition of reference sites

Many states currently define biological impairment in terms of multimetric index values observed at reference, or minimally disturbed, sites. Sites of unknown condition are then sampled and judged against this standard. Defining a set of reference sites may be arbitrary, either in terms of which sites are included or which criteria are used to define reference condition. For the MAIA study, we have seen that hand-picked reference sites did not necessarily match objective criteria for reference condition. This lesson learned in the Mid-Atlantic is relevant to states as they develop criteria for reference condition and rules for defining impairment. Currently, states vary both in the way they characterize reference condition and the way they define deviation from reference condition. Furthermore, the line of impairment ranges from the index value that corresponds to the 5th to the 95th percentile of reference sites (USEPA, 2002).

For the MAIA fish index, three methods were used to define reference sites that represented increasingly stringent criteria for reference condition in that more conditions had to be satisfied (McCormick et al., 2001). The least restrictive criteria were based on chemical criteria and RBP habitat measures; the moderately restrictive criteria also included measures based on watershed land use; and the most restrictive criteria included all these as well as Bryce et al.'s condition class. The value of the fish index that represented the 25th percentile for reference sites was selected as the line of impairment for each method. For the three sets of reference sites, the values of the fish index were very similar and their average was used to define reference condition for fish assemblages.

Patterns of index variability were similar across assemblage types

Statistical precision is an important feature of any monitoring tool because it determines the ability of an indicator to detect change should it occur. A highly variable indicator must show a large

change in value before the change is statistically significant. Lack of sensitivity translates into an inability to sound an alarm that will protect resources from degradation. Statistical power analysis can be used to estimate the magnitude of change that an indicator can detect. Two commonly used statistical models for power analysis are the t-test and regression (Peterman, 1990; Carlisle and Clements, 1999; Fore et al., 2001). Given EMAP's focus on estimating trends through time, a regression model with index regressed against year is arguably the more relevant statistical model for power analysis (Stevens, 1994; Larsen et al., 2001; Hughes et al., 1998; Urquhart et al., 1998). Results from the two approaches indicate that the MAIA multimetric indexes had adequate precision to distinguish between two and five categories of biological condition and could detect between 1.5% and 2.5% change per year after five years of monitoring.

Both approaches for estimating statistical power to detect change use estimates of variance components derived from ANOVA. For power analysis based on a t-test to compare two sites with three replicates each, within-year variance is used to calculate the minimum detectable difference for index values at two sites (MDD; Zar, 1984). By dividing the range of the index by the MDD, one can calculate the number of categories of biological condition that the index can detect (Fore et al., 1994; Fore et al., 2001; Fore 2002a; Blocksom, 2003). The regression model uses the within-year variance as well, but also uses estimates of variance associated with site x year interaction and year-to-year variability to calculate statistical power (Larsen et al., 1995; Urquhart et al., 1998).

The relative magnitude of the variance components illustrates which temporal influences were relatively more important for each index. The percentage of total variance associated with repeat visits to the same site, that is, all sources of variance besides that associated with site differences were approximately similar across assemblages (13–20%). For fish and diatom assemblages, the multimetric index tended to have less variance associated with repeat visits than did its component metrics (Figure 6; McCormick and Peck, 2000; Fore, 2002b). USEPA guidelines for biological indicators recommend that the overall variability associated with repeat sampling within a single year, or the error component, not exceed 10% (McCormick and Peck, 2000). Only the fish index met this target with 6% of total variance associated with error; at 17%, the diatom index was furthest from the mark and the invertebrate index was close with 13% (Table 4).

Indexes for each assemblage differed in the way that the "nuisance" variance was allocated to each of the different sources, i.e., year-to-year differences, site x year interaction, and repeat sampling within year (measurement error). For the diatom index, most of its variance was associated with repeat visits to a site while year-to-year variance was nearly zero. In contrast, fish and invertebrate indexes did tend to change together across years (year-to-year variance). This difference may be related to longer life cycles for fish and invertebrates compared with diatoms.

Based on statistical power calculations for the t-test model, the fish index was most precise and could detect approximately 5.5 categories of biological condition, the invertebrate index could detect 4 categories, and the diatom index 2.4. The diatom index had the largest percentage of its total variance associated with error; therefore, it follows that it had the lowest statistical power to detect differences.

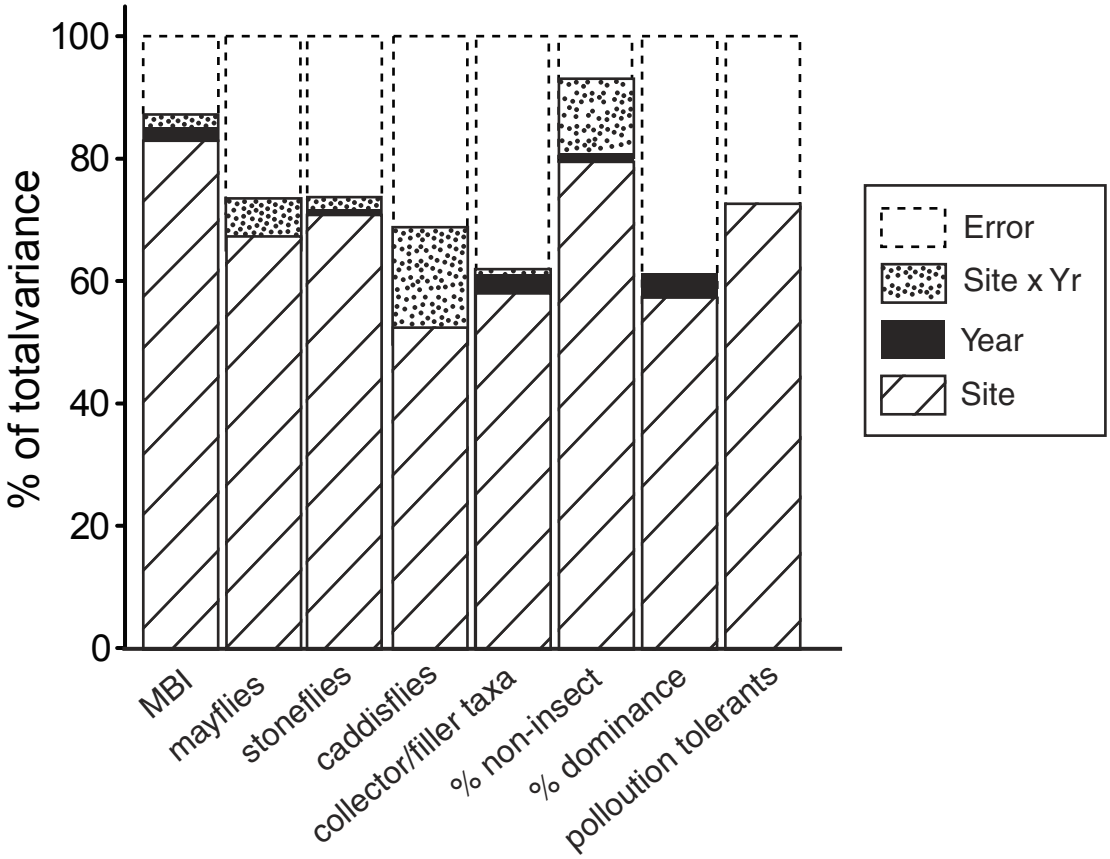

Figure 6. Variance components for the invertebrate index and its metrics. Variance associated with repeat visits to the same site during the same year ("error") and from year to year ("year") was lower for the index than for its metrics. Approximately 17% of the index's variability was associated with repeat visits (see Table 3).

Table 4. Components of variance expressed as a percentage of the total variance for diatom, invertebrate, and fish multimetric indexes. Variance associated with site differences, year-to-year differences, site x year interaction, and repeat visits within years are shown for each index.

	Diatom	Invertebrate	Fish
Site	80.4	83.3	86.8
Year	0	2.1	1.5
Site x year	2.7	1.6	5.6
Error (repeat visits)	16.9	12.9	6.2

For the trend (regression) model, the statistical power of the fish index was highest and the invertebrate index lowest (Figure 7). The differences were small after the first few years of monitoring, however, and the indexes had very similar statistical power over the long term. USEPA's performance objectives for trend detection specify that an indicator should be capable of detecting a 2% change per year after approximately five years of sampling 30–50 sites assuming a type I error rate of 0.1 and a type II error rate of 0.2 (McCormick and Peck, 2000). On the basis of five years of monitoring at 40 sites, the diatom index could detect a 1.5% change per year, the fish index 2.1%, and the invertebrate index 2.5%.

Thus, the two statistical models ranked the three indexes differently in terms of their statistical power; these differences are entirely an artifact of the statistical model used to calculate power. Because the year-to-year component of variability is relatively more important for detecting trends through time, higher annual variability of the invertebrate index caused it to have the lowest power for the trend (regression) model (Larsen et al., 1995; Urquhart et al., 1998). These results illustrate that when comparing statistical power, identical statistical models must be used because results depend entirely on the statistical model selected and its underlying assumptions (Blocksom, 2003).

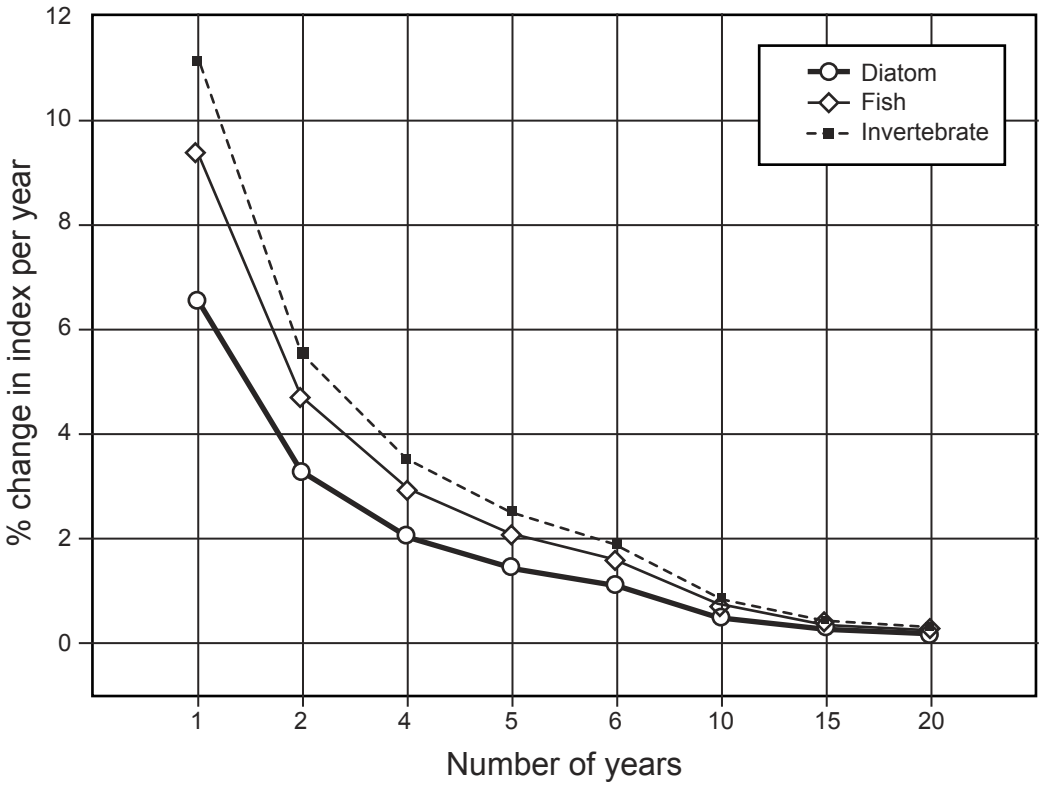

Figure 7. The percentage change per year that would represent a statistically significant change in multimetric index values decreases as the number of years of sampling increases. A smaller percentage change indicates a more precise index and greater statistical power to detect change. Thus, the invertebrate index was the *least* precise although all three indexes were very similar. Based on repeat visits to the same 40 sites each year.

Invertebrate and diatom index values were comparable for pool and riffle samples

Stream resources are naturally varied and complex but index values used for assessment must be independent of natural features and mean the same thing whether the assessment is from a low or high elevation site or a wide or narrow reach. In this way, an index must be applicable to all types of streams in a region.

The tendency when developing multimetric indexes is to select similar sites for metric testing, and eliminate data from small or unique ecoregions, unusual sites, or different habitat types. The purpose is to identify a set of homogeneous sites with similar underlying physical and geographic features so that metric response will be stronger in the absence of other confounding factors.

In the Mid-Atlantic region, a homogeneous set of sites could not be easily defined because there were so many choices for criteria including watershed size, ecoregion, and lowlands vs. uplands. A key lesson learned from the MAIA pilot was that including as much data as possible in the metric testing process was more efficient and simpler than testing multiple sets of sites. In addition, unique sites were not 'orphaned' from the assessment process.

Riffles and pools within a reach represent different habitats to invertebrates and diatoms and sampling in these areas typically yield different taxa. The MAIA sampling protocol kept the samples separate when both habitats occurred at a sample site. For invertebrates, pool and riffles were different: some metrics calculated for pools tended to indicate poorer conditions. Fortunately, similar metrics were correlated with disturbance in each habitat. To compensate for these natural differences, metrics were simply adjusted when they were converted to unit-less scores. In this way, the final index was comparable for samples from a pool or a riffle (Klemm et al., 2003). For diatoms, similar adjustments were not necessary for pool and riffle habitats because index values were similar from each habitat and contributed no measurable source of variability to the index (Fore, 2002b).

Assemblages differed in their sensitivity to disturbance types

When multiple assemblages are monitored, potentially conflicting assessments must be reconciled. What if a stream reach falls below the "impaired" threshold according to its fish index value but not according to its invertebrate index value? Several studies have found that biological indexes based on different assemblages generally agree in terms of assessment condition, that is, indexes for different assemblages are highly correlated (Lammert and Allan, 1999; O'Connor et al., 2000). Nonetheless, assemblages may disagree at specific sites and the conflict must be resolved to determine whether the site is impaired. Some of the disagreement may be associated with differential sensitivities of assemblages to specific types of disturbance (Bryce and Hughes, 2002; Norton et al., 2002).

The lesson learned from Mid-Atlantic streams was that any of the three assemblages could be used to monitor stream condition because multimetric indexes for all three assemblages could reliably distinguish degraded sites from sites with little or no human influence. However, different assemblages showed different sensitivities associated with their natural history and assessments based on different assemblages provide more information about the type and source of disturbance.

Multimetric indexes for all three assemblages could distinguish reference sites from sites disturbed by nutrient enrichment, mixed impacts of development and agriculture, and acid mine deposition (see Figure 4 on page 17). Overall, the indexes agreed, with lower values for mine drainage than for other types of disturbance. The indexes disagreed somewhat on the relative magnitude of degradation associated with nutrient and acid deposition, in that diatoms indicated greater degradation associated with nutrients, probably because several diatom metrics were related to different aspects of nutrient enrichment (e.g., organic vs. inorganic sources; Table 5). In contrast, acid deposition sites looked more like reference sites from the diatom point of view. Because acid deposition sites were typically nutrient poor and steeply sloped, they lacked diatom taxa that were associated with alkalinity, nutrients, or sediment.

The different assemblages tended to respond differently to specific stressors. Invertebrate and diatom indexes were more highly correlated with specific stressors measured at the reach scale, such as nitrogen or riparian condition than the fish index (see Table 1 on page 18). In contrast, the fish index was less significantly correlated with chemical measures and measures at the reach scale, possibly because fish are not limited to the smaller spatial scale of a reach, as are invertebrates and diatoms.

Table 5. Biological metrics included in the fish, invertebrate, and diatom indexes. Percent sign (%) denotes the percentage of individuals belonging to a given group out of the total number of sampled individuals.

Metric type	Diatom	Invertebrate	Fish
Taxonomic composition		No. Ephemeroptera taxa No. Plecoptera taxa No. Trichoptera taxa	No. cyprinid taxa % cottid
Tolerance & intolerance	% intolerant % very tolerant % salt tolerant % tolerant of low oxygen	Tolerance index	No. sensitive taxa % tolerant
Assemblage structure		% non-insects % dominance	No. benthic taxa % exotics
Autecological guild	% eutrophic % N heterotrophs % polysaprobic % alkaliphilic		
Trophic guild		No. collector/filterer taxa	% piscivore/invertivore % macro-omnivore
Morphometric guild	% very motile		
Reproductive guild			% gravel-spawning taxa

VII. CONCLUSIONS

The goal of the MAIA study was to demonstrate how a probabilistic sampling design could be used to develop monitoring protocols and assess surface waters at a regional scale. The purpose of this document was to distill the critical lessons from the Mid-Atlantic experience and present approaches and conclusions that will be relevant to states as they move forward with their own monitoring programs.

Although EMAP's objectives are to support states in their assessment of surface waters by developing survey design methods, biomonitoring tools, and biocriteria, these tools will not necessarily be adopted by states. The MAIA pilot was a demonstration; and states in the region are not obligated to adopt EMAP methods. Currently, monitoring and assessment under the Clean Water Act is each state's responsibility, although the USEPA may intervene if a state's program or reporting is inadequate. Whether states in the region of this pilot will adopt or adapt EMAP protocols is unclear. Nonetheless, the initial MAIA pilot study has metamorphosed into a larger collaborative effort that coordinates data and information across multiple agencies responsible for water resources in this region.

Whether or not the specific methods and protocols developed by EMAP are appropriate for state monitoring programs, the issues related to sampling design, data management, and the development of assessment tools are relevant to any emerging monitoring program. For sampling design, the primary issue relates to sample site selection. Resolution of this issue depends on a clear understanding of how the data will be used before they are collected (Ward et al., 1986). The way data are entered into the computer seems remote from data analysis and interpretation, but when data management issues are left unresolved, they can derail or corrupt any analysis.

For the MAIA study, objective criteria were used to establish links between human disturbance and biological response. To ensure that biological monitoring tools were both meaningful and reliable, they were evaluated for their correlation with independent measures of human disturbance and for their variability through time. Multimetric indexes developed for fish, invertebrates and diatoms showed a strong and consistent correlation with multiple aspects of human disturbance measured at a variety of spatial scales. For all biological indexes, the more integrative the measure of human disturbance, the higher the correlation. All three indexes were capable of detecting small changes in biological condition within a few years of sampling (2.5% or less per year). Both their strong correlation with disturbance and their statistical precision support multimetric indexes in their role as biological monitoring tools for protecting water resources within the legal framework of the Clean Water Act.

The size of the MAIA project created some of its own new challenges; hundreds of people have been involved in creating the sampling design and collecting, managing and analyzing the data. As state programs grow to meet the requirements of the Clean Water Act, the scale of this project appears less daunting (Ransel, 1995; NRC 2001). In fact, many states already assess hundreds of sites each year. Consequently, the process and outcomes of the MAIA pilot serve to highlight persistent issues or sticking points that states will continue to grapple with as they develop adequate biological monitoring programs (Yoder and Rankin, 1998).

REFERENCES

Barbour M.T., Gerritsen J., G.E. Griffith, R. Frydenborg, E. McCarron, J.S. White, and M.L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15: 185-211.

Barbour M.T., Gerritsen J., Snyder B.D. and Stribling J.B. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: Periphyton, benthic macroinvertebrates, and fish. Second edition. EPA 841-B-99-002. US Environmental Protection Agency, Office of Water, Washington, D.C.

Beyers, D.W. 1998. Causal inference in environmental impact studies. Journal of the North American Benthological Society 17:367-373.

Blocksom, K.A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. Environmental Management, *in press*.

Bryce, S.A., D.P. Larsen, R.M. Hughes, and P.R. Kaufmann. 1999. Assessing relative risks to aquatic ecosystems: A mid-Appalachian case study, Journal of the American Water Resources Association, 35(1), 23-36.

Bryce, S.A. and R.M. Hughes. 2002. Variable Assemblage Responses to Multiple Disturbance Gradients: Oregon and Appalachian, USA, Case Studies. Biological Response Signatures: Indicator Patterns Using Aquatic Communities (Ed T.P. Simon), pp. 539-560. CRC Press LLC, Boca Raton, FL.

Carlisle D.M. and Clements W.H. 1999. Sensitivity and variability of metrics used in biological assessments of running waters. Environmental Toxicology and Chemistry, 18, 285-291.

Davis, W.S. and Scott. 2000. Mid-Atlantic Highlands Streams Assessment: Technical Support Document. EPA/903/B-00/004. USEPA Mid-Atlantic Integrated Assessment, Ft. Meade, Maryland.

Fore L.S., Karr J.R. and Conquest L.L. 1994. Statistical properties of an index of biotic integrity used to evaluate water resources. Canadian Journal of Fisheries and Aquatic Sciences, 51, 212-231.

Fore L.S., Paulsen K., K. O'Laughlin. 2001. Assessing the performance of volunteers in monitoring streams. Freshwater Biology, 46, 109-123

Fore L.S. 2002a. Biological assessment of mining disturbance on stream invertebrates in mineralized areas of Colorado. Biological Response Signatures: Indicator Patterns Using Aquatic Communities (Ed T.P. Simon), pp. 445-480. CRC Press LLC, Boca Raton, FL.

Fore L.S. 2002b. Response of diatom assemblages to human disturbance: development and testing of a multimetric index for the Mid-Atlantic Region (USA). Biological Response Signatures: Indicator Patterns Using Aquatic Communities (Ed T.P. Simon), pp. 347-370. CRC Press LLC, Boca Raton, FL.

Fore, L.S. and C. Grafe. 2002. Using diatoms to assess the biological condition of large rivers in Idaho (USA). Freshwater Biology 47:2015-2037..

Gerritsen, J.G., J. Green, and R. Preston. 1994. Establishment of regional reference conditions for stream biological assessment and watershed management. Proceedings of Watersheds '93: A National Conference on Watershed Management, Arlington, Va., March 21-24, 1993, U.S. Environmental Protection Agency, EPA-804-R-94-002, p. 797-801.

Hale, S.S., L.H. Bahner and J.F. Paul. 2000. Finding common ground in managing data used for regional environmental assessments. Environmental Monitoring and Assessment 63:143-157.

Hale, S.S. and H.W. Buffum. 2000. Designing environmental databases for statistical analyses. Environmental Monitoring and Assessment 64:55-68.

Hale, S.S., M.M. Hughes, J.F. Paul, R.S. McAskill, S.A. Rego, D.R. Bender, N.J. Dodge, R.L. Richter and J.L. Copeland. 1998. Managing scientific data: The EMAP approach. Environmental Monitoring and Assessment 51: 429-440.

Hale, S.S., J. Rosen, D. Scott, J. Paul, and M. Hughes. 1999. EMAP Information Management Plan: 1998-2001. EPA/620/R-99/001a. National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency.

Heffner, R.A., M.J. Butler IV, and C.K. Reilly. 1996. Pseudoreplication revisited. Ecology 77:2558-2562.

Herlihy, A.T., P.R. Kaufmann, and M.E. Mitch. 1990. Regional estimates of acidic mine drainage impact on streams in the Mid-Atlantic and Southeastern United States. Water, Air and Soil Pollution 50:91-107.

Herlihy, A.T., P.R. Kaufmann, M.R. Church, P.J. Wigington, Jr., J.R. Webb, and M.J. Sale. 1993. The effects of acidic deposition on streams in the Appalachian Mountain and Piedmont region of the Mid-Atlantic United States. Water Resources Research 29:2687-2703.

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., J.L. Stoddard, and C.B. Johnson. 1998. The relationship between stream chemistry and watershed land cover data in the Mid-Atlantic region, U.S. Water, Air and Soil Pollution, 105, 377-386.

Hill, B.H., A.T. Herlihy, P.R. Kaufmann, R.J. Stevenson, F.H. McCormick, and C. Burch Johnson. 2000. Use of periphyton assemblage data as an index of biotic integrity, 19(1), 50-67.

Hill, B.H., R.J. Stevenson, Y. Pan, A.T. Herlihy, P.R. Kaufmann, C.B. Johnson. 2001. Comparison of correlations between environmental characteristics and stream diatom assemblages characterized at genus and species levels. Journal of the North American Benthological Society 20:299-310.

Hudson, L.I. and J.J.H. Cibrowski. 1996. Teratogenic and genotoxic responses of larval *Chironomus salinarius* group (Diptera: Chironomidae) to contaminated sediment. Environmental Toxicology and Chemistry 15: 1375-1381.

Hughes, R.M. 1995. Defining acceptable biological status by comparing with reference conditions. In Davis, W.S. and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, FL:31-47.

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., S.G. Paulsen, and J.L. Stoddard. 2000. EMAP-Surface Waters: A national, multiassemblage, probability survey of ecological integrity, Hydrobiologia, 423, 429-443.

Hurlbert S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, 54, 187-211.

Jackson, L.E., J.C. Kurtz, and W.S. Fisher (Eds.). 2000. Evaluation Guidelines for Ecological Indicators. EPA/620/R-99/005. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC. 107 pp.

Karr J.R. 1991. Biological integrity: a long-neglected aspect of water resource management. Ecological Applications, 1, 66-84.

Karr J.R. and Chu E.W. 1999. Restoring Life in Running Waters: Better Biological Monitoring. Island Press, Washington, DC.

Karr J.R., Fausch K.D., Angermeier P.L., Yant P.R. and Schlosser I.J. 1986. Assessment of biological integrity in running water: a method and its rationale. Illinois Natural History Survey Special Publication Number 5, Champaign, IL.

Karr J.R., Allan J.D. and Benke A.C. 2000. River conservation in the United States and Canada. Global Perspectives on River Conservation: Science, Policy, Practice (Eds P.J. Boon, B.R. Davies and G.E. Petts), pp. 3-39. J. Wiley, Chichester, UK.

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, 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, 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(5): 656-669.

Klemm, D.J., K.A. Blocksom, W.T. Thoeny, F.A. Fulk, A.T. Herlihy, P.R. Kaufmann, S.M. Cormier. 2002. Using macroinvertebrates as indicators of ecological conditions for streams in the Mid-Atlantic Highlands region. Environmental Monitoring and Assessment 78:169-212.

Lammert, M. and J.D. Allan. 1999. Assessing biotic integrity of streams: Effects of scale in measuring the influence of land use/cover and habitat structure on fish and macroinvertebrates, Environmental Management, 23(2), 257-270.

Larsen, D.P., N.S. Urquhart, and D.L. Kugler. 1995. Regional scale trend monitoring of indicators of trophic condition of lakes. Water Resources Bulletin 31: 117-140.

Larsen, D. P., T. M. Kincaid, S. E. Jacobs, and N. S. Urquhart. 2001. Designs for evaluating local and regional scale trends. BioScience 12:1069-1078.

Lemly, A.D. 2000. Using bacterial growth on insects to assess nutrient impacts in streams. Environmental Monitoring and Assessment 63:431-446.

Loftis, J.C. C.H. Taylor, A.D. Newell, and P.L. Chapman. 1991. Multivariate trend testing of lake water quality. Water Resources Bulletin 27:461-473.

McCormick, F.H., R.M. Hughes, P.R. Kaufmann, D.V. Peck, J.L. Stoddard, A.T. Herlihy. 2001. Development of an index of biotic integrity for the Mid-Atlantic Highlands region. Transactions of the American Fisheries Society 130:857-87.

McCormick, F.H. and D.V. Peck. 2000. Application of the indicator evaluation guidelines to a multimetric indicator of ecological condition based on stream fish assemblages. Chapter 4 in Evaluation Guidelines for Ecological Indicators (L.E. Jackson, J.C. Kurtz, and W.S. Fisher, eds.), EPA/620/R-99/005. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.

Mebane, C.A. 2002. Effects of metals on freshwater macroinvertebrates: a review and case study of the correspondence of multimetric index, toxicity testing, and copper concentrations in sediment and water. Biological Response Signatures: Indicator Patterns Using Aquatic Communities (Ed T.P. Simon), pp. 287-312. CRC Press LLC, Boca Raton, FL.

Miltner, R.J. and E.T. Rankin. 1998. Primary nutrients and the biotic integrity of rivers and streams. Freshwater Biology 40:145-158.

National Research Council (NRC). 1995. Review of EPA's Environmental Monitoring and Assessment Program: Overall Evaluation. National Academy Press, Washington, DC.

National Research Council (NRC). 2001. Assessing the TMDL Approach to Water Quality Management. National Academy Press, Washington, DC. 109 pp.

Norton, S.B., S.M. Cormier, M. Smith, R. Christian Jones. 2000. Can biological assessments discriminate among types of stress? A case study from the Eastern Corn Belt Plains ecoregion. Environmental Toxicology and Chemistry 19:1113-1119.

O'Connell, T.J., L.E. Jackson, and R.P. Brooks. 1998. A bird community index of biotic integrity for the mid-Atlantic highlands, Environmental Monitoring and Assessment, 51(1-2), 145-156.

O'Connor, R.J., T.E. Walls, and R.M. Hughes. 2000. Using multiple taxonomic groups to index the ecological condition of lakes. Environmental Monitoring and Assessment 61:207-228.

Olden, J.D. and D. A. Jackson. 2000. Torturing data for the sake of generality: how valid are our regression models? Ecoscience 7:501-510.

Olsen, A.R., J. Sedransk, D. Edwards, Gotway, C.A., Liggett, W., Rathbun, S., Reckhow, K.H. and Young, L.J. 1999. Statistical issues for monitoring ecological and natural resources in the United States. Environmental Monitoring and Assessment 54: 1-45

Overton, W.S. and Stehman, S.V. 1996. Desirable design characteristics for long-term monitoring of ecological variables. Environmental and Ecological Statistics. 3: 349-361.

Pan, Y.R.J. Stevenson, B.H. Hill, A.T. Herlihy, and G.B. Collins. 1996. Using diatoms as indicators of ecological conditions in lotic systems: a regional assessment. Journal of the North American Benthological Society 15: 481-495.

Pan Y., Stevenson R.J., Hill B.H. and Herlihy A.T. 2000. Ecoregions and benthic diatom assemblages in Mid-Atlantic Highlands streams, USA. Journal of the North American Benthological Society, 19, 518-540.

Pan, Y.R.J. Stevenson, B.H. Hill, P.R. Kaufmann, and A.T. Herlihy. 1999. Spatial patterns and ecological determinants of benthic algal assemblages in Mid-Atlantic streams, USA. Journal of Phycology 35: 460-468.

Paulsen, S.G., R.M. Hughes, and D.P. Larsen. 1998. Critical elements in describing and understanding our nation's aquatic resources. Journal of the American Water Resources Association 34:995-1005.

Peterman R.M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences, 47, 2-15.

Peterson, S.A, N.S. Urquhart, and E.B. Welch. 1999. Sample representativeness: a must for reliable regional lake condition estimates. Environmental Science and Technology. 33: 1559-1565.

Ransel K.P. 1995. The sleeping giant awakes: PUD No. 1 of Jefferson County v. Washington Department of Ecology. Environmental Law, 25, 255-283.

Richardson, J.S. and P.M. Kiffney. 2000. Response of a stream macroinvertebrate community from a pristine, southern BC stream to metals in experimental mesocosms. Environmental Toxicology and Chemistry19:736-743.

Stevens, D.L. Jr. 1994. Implementation of a national environmental monitoring program. Journal of Environmental Management 42:1-29.

Stevenson R.J. and Bahls L. 1999. Chapter six: periphyton protocols. Rapid bioassessment protocols for use in streams and wadeable rivers: Periphyton, benthic macroinvertebrates, and fish, 2nd edn. (Eds M.T. Barbour, J. Gerritsen, B.D. Snyder and J.B. Stribling) EPA 841-B-99-002. US Environmental Protection Agency, Office of Water, Washington, D.C.

Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. 1986.Environmental impact assessment: "Pseudoreplication" in time? Ecology 67:929-940.

Stewart-Oaten, A., J.R. Bence, and C.W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology 73: 1396-1404.

Stoddard, J.L., C.T. Driscoll, J.S. Kahl, and J.H. Kellogg. 1998. Can site-specific trends be extrapolated to a region? An acidification example for the Northeast. Ecological Applications 8: 288-299.

Suter, G.W. II. 2001. Applicability of indicator monitoring to ecological risk assessment. Ecological Indicators 1:1010-112.

Urquhart, N.S., S.G. Paulsen, and D.P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8:246-257.

U.S. Environmental Protection Agency (EPA). 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, Office of Environmental Information and Office of Water, Washington, DC.

U.S. Environmental Protection Agency (EPA). 2003. Aquatic Resource Monitoring Web Site. www.epa.gov/nheerl/arm/index.htm

Waite, I.R., A.T. Herlihy, D.P. Larsen, D.J. Klemm. 2000. Comparing strengths of geographic and nongeographic classifications of stream benthic macroinvertebrates in the Mid-Atlantic Highlands, USA. Journal of the North American Benthological Society 19:429-441.

Wallace, J.B., J.W. Grubaugh and M.R. Whiles. 1996. Biotic indices and stream ecosystem processes: results from an experimental study. Ecological Applications 6: 140-151.

Wang, L., J., Lyons, P. Kanehl. 1998. Development and evaluation of a habitat rating system for low-gradient Wisconsin streams. North American Journal of Fisheries Management 19:775-785.

Ward, R.C., Loftis, J.C., and G.B. McBride. 1986. The "data rich but information poor" syndrome in water quality monitoring. Environmental Management 10:291-297.

White, J. and G. Merritt. 1998. Evaluation of R-EMAP techniques for the measurement of ecological integrity of streams in Washington state's coast range ecoregion. Environmental Monitoring and Assessment 51:345-355.

Yoder, C.O. and E.T. Rankin. 1998. The role of biological indicators in a state water quality management process. Environmental Monitoring and Assessment 51: 61-88.

Yoder, C.O. and J.E. DeShon. 2002. Using biological response signatures within a framework of multiple indicators to assess and diagnose causes and sources of impairments to aquatic assemblages in selected Ohio rivers and streams. Biological Response Signatures: Indicator Patterns Using Aquatic Communities (Ed T.P. Simon), pp. 23-82. CRC Press LLC, Boca Raton, FL.

Zar J.H. 1984. Biostatistical Analysis, 2nd ed. Prentice-Hall, Inc., Englewood Cliffs, NJ.

United States
Environmental Protection
Agency

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE ☐; detach, or copy the cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35

Office of Environmental Information
and Office of Water
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA/903/R-03/003
March 2003