EPA/620/R-99/005
May 2000

EVALUATION GUIDELINES FOR ECOLOGICAL INDICATORS

Edited by

Laura E. Jackson
Janis C. Kurtz
William S. Fisher

U.S. Environmental Protection Agency
Office of Research and Development
Research Triangle Park, NC 27711

♻ *Printed on Recycled Paper*

Notice

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency. It has been subjected to the Agency's review, and it has been approved for publication as EPA draft number NHEERL-RTP-MS-00-08. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The editors wish to thank the authors of Chapters Two, Three, and Four for their patience and dedication during numerous document revisions, and for their careful attention to review comments. Thanks also go to the members of the ORD Ecological Indicators Working Group, which was instrumental in framing this document and highlighting potential users. We are especially grateful to the 12 peer reviewers from inside and outside the U.S. Environmental Protection Agency for their insights on improving the final draft.

This report should be cited as follows:

Jackson, Laura E., Janis C. Kurtz, and William 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 p.

Abstract

This document presents fifteen technical guidelines to evaluate the suitability of an ecological indicator for a particular monitoring program. The guidelines are organized within four evaluation phases: conceptual relevance, feasibility of implementation, response variability, and interpretation and utility. The U.S. Environmental Protection Agency's Office of Research and Development has adopted these guidelines as an iterative process for internal and (EPA's) affiliated researchers during the course of indicator development, and as a consistent framework for indicator review. Chapter One describes the guidelines; Chapters Two, Three, and Four illustrate application of the guidelines to three indicators in various stages of development. The example indicators include a direct chemical measure, dissolved oxygen concentration, and two multi-metric biological indices, an index of estuarine benthic condition and one based on stream fish assemblages. The purpose of these illustrations is to demonstrate the evaluation process using real data and working with the limitations of research in progress. Furthermore, these chapters demonstrate that an evaluation may emphasize individual guidelines differently, depending on the type of indicator and the program design. The evaluation process identifies weaknesses that may require further indicator research and modification. This document represents a compilation and expansion of previous efforts, in particular, the initial guidance developed for EPA's Environmental Monitoring and Assessment Program (EMAP).

Keywords: ecological indicators, EMAP, environmental monitoring, ecological assessment, Environmental Monitoring and Assessment Program

Preface

This document describes a process for the technical evaluation of ecological indicators. It was developed by members of the U.S. Environmental Protection Agency's (EPA's) Office of Research and Development (ORD), to assist primarily the indicator research component of ORD's Environmental Monitoring and Assessment Program (EMAP). The *Evaluation Guidelines* are intended to direct ORD scientists during the course of indicator development, and provide a consistent framework for indicator review. The primary users will evaluate indicators for their suitability in ORD-affiliated ecological monitoring and assessment programs, including those involving other federal agencies. This document may also serve technical needs of users who are evaluating ecological indicators for other programs, including regional, state, and community-based initiatives.

The *Evaluation Guidelines* represent a compilation and expansion of previous ORD efforts, in particular, the initial guidance developed for EMAP. General criteria for indicator evaluation were identified for EMAP by Messer (1990) and incorporated into successive versions of the EMAP Indicator Development Strategy (Knapp 1991, Barber 1994). The early EMAP indicator evaluation criteria were included in program materials reviewed by EPA's Science Advisory Board (EPA 1991) and the National Research Council (NRC 1992, 1995). None of these reviews recommended changes to the evaluation criteria.

However, as one result of the National Research Council's review, EMAP incorporated additional temporal and spatial scales into its research mission. EMAP also expanded its indicator development component, through both internal and extramural research, to address additional indicator needs. Along with indicator development and testing, EMAP's indicator component is expanding the Indicator Development Strategy, and revising the general evaluation criteria in the form of technical guidelines presented here with more clarification, detail, and examples using ecological indicators currently under development.

The Ecological Indicators Working Group that compiled and detailed the *Evaluation Guidelines* consists of researchers from all of ORD's National Research Laboratories--Health and Environmental Effects, Exposure, and Risk Management--as well as ORD's National Center for Environmental Assessment. This group began in 1995 to chart a coordinated indicator research program. The working group has incorporated the *Evaluation Guidelines* into the ORD Indicator Research Strategy, which applies also to the extramural grants program, and is working with potential user groups in EPA Regions and Program Offices, states, and other federal agencies to explore the use of the *Evaluation Guidelines* for their indicator needs.

References

Barber, C.M., ed. 1994. Environmental Monitoring and Assessment Program: Indicator Development Strategy. EPA/620/R-94/022. U.S. Environmental Protection Agency, Office of Research and Development: Research Triangle Park, NC.

EPA Science Advisory Board. 1991. Evaluation of the Ecological Indicators Report for EMAP; A Report of the Ecological Monitoring Subcommittee of the Ecological Processes and Effects Committee. EPA/SAB/EPEC/91-01. U.S. Environmental Protection Agency, Science Advisory Board: Washington, DC.

Knapp, C.M., ed. 1991. Indicator Development Strategy for the Environmental Monitoring and Assessment Program. EPA/600/3-91/023. U.S. Environmental Protection Agency, Office of Research and Development: Corvallis, OR.

Messer, J.J. 1990. EMAP indicator concepts. *In:* Environmental Monitoring and Assessment Program: Ecological Indicators. EPA/600/3-90/060. Hunsaker, C.T. and D.E. Carpenter, eds. United States Environmental Protection Agency, Office of Research and Development: Research Triangle Park, NC, pp. 2-1 - 2-26.

National Research Council. 1992. Review of EPA's Environmental Monitoring and Assessment Program: Interim Report. National Academy Press: Washington, DC.

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

Contents

Introduction

Worldwide concern about environmental threats and sustainable development has led to increased efforts to monitor and assess status and trends in environmental condition. Environmental monitoring initially focused on obvious, discrete sources of stress such as chemical emissions. It soon became evident that remote and combined stressors, while difficult to measure, also significantly alter environmental condition. Consequently, monitoring efforts began to examine ecological receptors, since they expressed the effects of multiple and sometimes unknown stressors and their status was recognized as a societal concern. To characterize the condition of ecological receptors, national, state, and community-based environmental programs increasingly explored the use of ecological indicators.

An indicator is a sign or signal that relays a complex message, potentially from numerous sources, in a simplified and useful manner. An ecological indicator is defined here as a measure, an index of measures, or a model that characterizes an ecosystem or one of its critical components. An indicator may reflect biological, chemical or physical attributes of ecological condition. The primary uses of an indicator are to characterize current status and to track or predict significant change. With a foundation of diagnostic research, an ecological indicator may also be used to identify major ecosystem stress.

There are several paradigms currently available for selecting an indicator to estimate ecological condition. They derive from expert opinion, assessment science, ecological epidemiology, national and international agreements, and a variety of other sources (see Noon 1998, Anonymous 1995, Cairns *et al.* 1993, Hunsaker and Carpenter 1990, and Rapport *et al.* 1985). The chosen paradigm can significantly affect the indicator that is selected and is ultimately implemented in a monitoring program. One strategy is to work through several paradigms, giving priority to those indicators that emerge repeatedly during this exercise.

Under EPA's Framework for Ecological Risk Assessment (EPA 1992), indicators must provide information relevant to specific assessment questions, which are developed to focus monitoring data on environmental management issues. The process of identifying environmental values, developing assessment questions, and identifying potentially responsive indicators is presented elsewhere (Posner 1973, Bardwell 1991, Cowling 1992, Barber 1994, Thornton *et al.* 1994). Nonetheless, the importance of appropriate assessment questions cannot be overstated; an indicator may provide accurate information that is ultimately useless for making management decisions. In addition, development of assessment questions can be controversial because of competing interests for environmental resources. However important, it is not within the purview of this document to focus on the development and utility of assessment questions. Rather, it is intended to guide the technical evaluation of indicators within the presumed context of a pre-established assessment question or known management application.

Numerous sources have developed criteria to evaluate environmental indicators. This document assembles those factors most relevant to ORD-affiliated ecological monitoring and assessment programs into 15 guidelines and, using three ecological indicators as examples, illustrates the types of information that should be considered under each guideline. This format is intended to facilitate consistent and technically-defensible indicator research and review. Consistency is critical to developing a dynamic and iterative base of knowledge on the strengths and weaknesses of individual indicators; it allows comparisons among indicators and documents progress in indicator development.

Building on Previous Efforts

The *Evaluation Guidelines* document is not the first effort of its kind, nor are indicator needs and evaluation processes unique to EPA. As long as managers have accepted responsibility for environmental programs, they have required measures of performance (Reams *et al.* 1992). In an international effort to promote consistency in the collection and interpretation of environmental information, the Organization for Economic Cooperation and Development (OECD) developed a conceptual framework, known as the Pressure-State-Response (PSR) framework, for categorizing environmental indicators (OECD 1993). The PSR framework encompasses indicators of human activities (pressure), environmental condition (state), and resulting societal actions (response).

The PSR framework is used in OECD member countries including the Netherlands (Adriaanse 1993) and the U.S., such as in the Department of Commerce's National Oceanic and Atmospheric Administration (NOAA 1990) and the Department of Interior's Task Force on Resources and Environmental Indicators. Within EPA, the Office of Water adopted the PSR framework to select indicators for measuring progress towards clean water and safe drinking water (EPA 1996a). EPA's Office of Policy, Planning and Evaluation (OPPE) used the PSR framework to support the State Environmental Goals and Indicators Project of the Data Quality Action Team (EPA 1996b), and as a foundation for expanding the Environmental Indicators Team of the Environmental Statistics and Information Division. The Interagency Task Force on Monitoring Water Quality (ITFM 1995) refers to the PSR framework, as does the International Joint Commission in the Great Lakes Water Quality Agreement (IJC 1996).

OPPE expanded the PSR framework to include indicators of the interactions among pressures, states and responses (EPA 1995). These types of measures add an "effects" category to the PSR framework (now PSR/E). OPPE incorporated EMAP's indicator evaluation criteria (Barber 1994) into the PSR/E framework's discussion of those indicators that reflect the combined impacts of multiple stressors on ecological condition.

Measuring management success is now required by the U.S. Government Performance and Results Act (GPRA) of 1993, whereby agencies must develop program performance reports based on indicators and goals. In cooperation with EPA, the Florida Center for Public Management used the GPRA and the PSR framework to develop indicator evaluation criteria for EPA Regions and states. The Florida Center defined a hierarchy of six indicator types, ranging from measures of administrative actions such as the number of permits issued, to measures of ecological or human health, such as density of sensitive species. These criteria have been adopoted by EPA Region IV (EPA 1996c), and by state and local management groups. Generally, the focus for guiding environmental policy and decision-making is shifting from measures of program and administrative performance to measures of environmental condition.

ORD recognizes the need for consistency in indicator evaluation, and has adopted many of the tenets of the PSR/E framework. ORD indicator research focuses primarily on ecological condition (state), and the associations between condition and stressors (OPPE's "effects" category). As such, ORD develops and implements science-based, rather than administrative policy performance indicators. ORD researchers and clients have determined the need for detailed technical guidelines to ensure the reliability of ecological indicators for their intended applications. The *Evaluation Guidelines* expand on the information presented in existing frameworks by describing the statistical and implementation requirements for effective ecological indicator performance. This document does not address policy indicators or indicators of administrative action, which are emphasized in the PSR approach.

Four Phases of Evaluation

Chapter One presents 15 guidelines for indicator evaluation in four phases (originally suggested by Barber 1994): conceptual foundation, feasibility of implementation, response variability, and interpretation and utility. These phases describe an idealized progression for indicator development that flows from fundamental concepts to methodology, to examination of data from pilot or monitoring studies, and lastly to consideration of how the indicator serves the program objectives. The guidelines are presented in this sequence also because movement from one phase into the next can represent a large commitment of resources (*e.g.*, conceptual fallacies may be resolved less expensively than issues raised during method development or a large pilot study). However, in practice, application of the guidelines may be iterative and not necessarily sequential. For example, as new information is generated from a pilot study, it may be necessary to revisit conceptual or methodological issues. Or, if an established indicator is being modified for a new use, the first step in an evaluation may concern the indicator's feasibility of implementation rather than its well-established conceptual foundation.

Each phase in an evaluation process will highlight strengths or weaknesses of an indicator in its current stage of development. Weaknesses may be overcome through further indicator research and modification. Alternatively, weaknesses might be overlooked if an indicator has strengths that are particularly important to program objectives. The protocol in ORD is to demonstrate that an indicator performs satisfactorily in all phases before recommending its use. However, the *Evaluation Guidelines* may be customized to suit the needs and constraints of many applications. Certain guidelines may be weighted more heavily or reviewed more frequently. The phased approach described here allows interim reviews as well as comprehensive evaluations. Finally, there are no restrictions on the types of information (journal articles, data sets, unpublished results, models, *etc.*) that can be used to support an indicator during evaluation, so long as they are technically and scientifically defensible.

References

Adriaanse, A. 1993. Environmental Policy Performance Indicators: A Study on the Development of Indicators for Environmental Policy in the Netherlands. Netherlands Ministry of Housing, Physical Planning and Environment.

Anonymous, 1995. Sustaining the World's Forests: The Santiago Agreement. *Journal of Forestry* 93: 18-21.

Barber, M.C., ed. 1994. Indicator Development Strategy. EPA/620/R-94/022. U.S. Environmental Protection Agency, Office of Research and Development: Research Triangle Park, NC.

Bardwell, L.V. 1991. Problem-framing: a perspective on environmental problem-solving. *Environmental Management* 15:603-612.

Cairns J. Jr., P.V. McCormick and B.R. Niederlehner. 1993. A proposed framework for developing indicators of ecosystem health. *Hydrobiologia* 263:1-44.

Cowling, E.B. 1992. The performance and legacy of NAPAP. *Ecological Applications* 2:111-116.

EPA. 1992. Framework for Ecological Risk Assessment. EPA/630/R-92/001. U.S. Environmental Protection Agency, Office of Research and Development: Washington, DC.

EPA. 1995. A Conceptual Framework to Support Development and Use of Environmental Information in Decision-Making. EPA 239-R-95-012. United States Environmental Protection Agency, Office of Policy Planning and Evaluation, April, 1995.

EPA. 1996a. Environmental Indicators of Water Quality in the United States. EPA 841-R-96-002, United States Environmental Protection Agency, Office of Water, Washington, D.C.

EPA. 1996b. Revised Draft: Process for Selecting Indicators and Supporting Data; Second Edition. United States Environmental Protection Agency, Office of Policy Planning and Evaluation, Data Quality Action Team, May 1996.

EPA. 1996c. Measuring Environmental Progress for U.S. EPA and the States of Region IV: Environmental Indicator System. United States Environmental Protection Agency, Region IV, July, 1996.

Hunsaker, C.T. and D.E. Carpenter, eds. 1990. Ecological Indicators for the Environmental Monitoring and Assessment Program. EPA 600/3-90/060. The U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.

IJC. 1996. Indicators to Evaluate Progress under the Great Lakes Water Quality Agreement. Indicators for Evaluation Task Force; International Joint Commission.

ITFM. 1995. Strategy for Improving Water Quality Monitoring in the United States: Final Report. Intergovernmental Task Force on Monitoring Water Quality. United States Geological Survey, Washington, D.C.

NOAA. 1990. NOAA Environmental Digest - Selected Indicators of the United States and the Global Environment. National Oceanographic and Atmospheric Administration.

Noon, B.R., T.A. Spies, and M.G. Raphael. 1998. Conceptual Basis for Designing an Effectiveness Monitoring Program. Chapter 2 *In:* The Strategy and Design of the Effectiveness Monitoring Program for the Northwest Forest Plan, General Technical Report PNW-GTR-437, Portland, OR: USDA Forest Service Pacific Northwest Research Station. pp. 21-48.

OECD. 1993. OECD Core Set of Indicators for Environmental Performance Reviews. Environmental Monograph No. 83. Organization for Economic Cooperation and Development.

Posner, M.I. 1973. Cognition: An Introduction. Glenview, IL: Scott Foresman Publication.

Rapport, D.J., H.A. Reigier, and T.C. Hutchinson. 1985. Ecosystem Behavior under Stress. *American Naturalist* 125: 617-640.

Reams, M.A., S.R. Coffee, A.R. Machen, and K.J. Poche. 1992. Use of Environmental Indicators in Evaluating Effectiveness of State Environmental Regulatory Programs. *In:* Ecological Indicators, vol 2, D.H. McKenzie, D.E.Hyatt and V.J. McDonald, Editors. Elsevier Science Publishers, pp. 1245-1273.

Thornton, K.W., G.E. Saul, and D.E. Hyatt. 1994. Environmental Monitoring and Assessment Program: Assessment Framework. EPA/620/R-94/016. U.S. Environmental Protection Agency, Office of Research and Development: Research Triangle Park, NC.

Chapter 1

Presentation of the Guidelines

Phase 1: Conceptual Relevance

The indicator must provide information that is relevant to societal concerns about ecological condition. The indicator should clearly pertain to one or more identified assessment questions. These, in turn, should be germane to a management decision and clearly relate to ecological components or processes deemed important in ecological condition. Often, the selection of a relevant indicator is obvious from the assessment question and from professional judgement. However, a conceptual model can be helpful to demonstrate and ensure an indicator's ecological relevance, particularly if the indicator measurement is a surrogate for measurement of the valued resource. This phase of indicator evaluation does not require field activities or data analysis. Later in the process, however, information may come to light that necessitates re-evaluation of the conceptual relevance, and possibly indicator modification or replacement. Likewise, new information may lead to a refinement of the assessment question.

Guideline 1: Relevance to the Assessment

Early in the evaluation process, it must be demonstrated in concept that the proposed indicator is responsive to an identified assessment question and will provide information useful to a management decision. For indicators requiring multiple measurements (indices or aggregates), the relevance of each measurement to the management objective should be identified. In addition, the indicator should be evaluated for its potential to contribute information as part of a suite of indicators designed to address multiple assessment questions. The ability of the proposed indicator to complement indicators at other scales and levels of biological organization should also be considered. Redundancy with existing indicators may be permissible, particularly if improved performance or some unique and critical information is anticipated from the proposed indicator.

Guideline 2: Relevance to Ecological Function

It must be demonstrated that the proposed indicator is conceptually linked to the ecological function of concern. A straightforward link may require only a brief explanation. If the link is indirect or if the indicator itself is particularly complex, ecological relevance should be clarified with a description, or conceptual model. A conceptual model is recommended, for example, if an indicator is comprised of multiple measurements or if it will contribute to a weighted index. In such cases, the relevance of each component to ecological function and to the index should be described. At a minimum, explanations and models should include the principal stressors that are presumed to impact the indicator, as well as the resulting ecological response. This information should be supported by available environmental, ecological and resource management literature.

Phase 2: Feasibility of Implementation

Adapting an indicator for use in a large or long-term monitoring program must be feasible and practical. Methods, logistics, cost, and other issues of implementation should be evaluated before routine data *collection begins. Sampling, processing and analytical methods should be documented for all measurements that comprise the indicator. The logistics and costs associated with training, travel, equipment and field and laboratory work should be evaluated and plans for information management and quality assurance developed.*

Note: Need For a Pilot Study

If an indicator demonstrates conceptual relevance to the environmental issue(s) of concern, tests of measurement practicality and reliability will be required before recommending the indicator for use. In all likelihood, existing literature will provide a basis for estimating the feasibility of implementation (Phase 2) and response variability (Phase 3). Nonetheless, both new and previously-developed indicators should undergo some degree of performance evaluation in the context of the program for which they are being proposed.

A pilot study is recommended in a subset of the region designated for monitoring. To the extent possible, pilot study sites should represent the range of elevations, biogeographic provinces, water temperatures, or other features of the monitoring region that are suspected or known to affect the indicator(s) under evaluation. Practical issues of data collection, such as time and equipment requirements, may be evaluated at any site. However, tests of response variability require a priori *knowledge of a site's baseline ecological condition.*

Pilot study sites should be selected to represent a gradient of ecological condition from best attainable to severely degraded. With this design, it is possible to document an indicator's behavior under the range of potential conditions that will be encountered during routine monitoring. Combining attributes of the planned survey design with an experimental design may best estimate the variance components. The pilot study will identify benchmarks of response for sensitive indicators so that routine monitoring sites can be classified on the condition gradient. The pilot study will also identify indicators that are insensitive to variations in ecological condition and therefore may not be recommended for use.

Clearly, determining the ecological condition of potential pilot study sites should be accomplished without the use of any of the indicators under evaluation. Preferably, sites should be located where intensive studies have already documented ecological status. Professional judgement may be required to select additional sites for more complete representation of the region or condition gradient.

Guideline 3: Data Collection Methods

Methods for collecting all indicator measurements should be described. Standard, well-documented methods are preferred. Novel methods should be defended with evidence of effective performance and, if applicable, with comparisons to standard methods. If multiple methods are necessary to accommodate diverse circumstances at different sites, the effects on data comparability across sites must be addressed. Expected sources of error should be evaluated.

Methods should be compatible with the monitoring design of the program for which the indicator is intended. Plot design and measurements should be appropriate for the spatial scale of analysis. Needs for specialized equipment and expertise should be identified.

Sampling activities for indicator measurements should not significantly disturb a site. Evidence should be provided to ensure that measurements made during a single visit do not affect the same measurement at subsequent visits or, in the case of integrated sampling regimes, simultaneous measurements at the site. Also, sampling should not create an adverse impact on protected species, species of special concern, or protected habitats.

Guideline 4: Logistics
The logistical requirements of an indicator can be costly and time-consuming. These requirements must be evaluated to ensure the practicality of indicator implementation, and to plan for personnel, equipment, training, and other needs. A logistics plan should be prepared that identifies requirements, as appropriate, for field personnel and vehicles, training, travel, sampling instruments, sample transport, analytical equipment, and laboratory facilities and personnel. The length of time required to collect, analyze and report the data should be estimated and compared with the needs of the program.

Guideline 5: Information Management
Management of information generated by an indicator, particularly in a long-term monitoring program, can become a substantial issue. Requirements should be identified for data processing, analysis, storage, and retrieval, and data documentation standards should be developed. Identified systems and standards must be compatible with those of the program for which the indicator is intended and should meet the interpretive needs of the program. Compatibility with other systems should also be considered, such as the internet, established federal standards, geographic information systems, and systems maintained by intended secondary data users.

Guideline 6: Quality Assurance
For accurate interpretation of indicator results, it is necessary to understand their degree of validity. A quality assurance plan should outline the steps in collection and computation of data, and should identify the data quality objectives for each step. It is important that means and methods to audit the quality of each step are incorporated into the monitoring design. Standards of quality assurance for an indicator must meet those of the targeted monitoring program.

Guideline 7: Monetary Costs
Cost is often the limiting factor in considering to implement an indicator. Estimates of all implementation costs should be evaluated. Cost evaluation should incorporate economy of scale, since cost per indicator or cost per sample may be considerably reduced when data are collected for multiple indicators at a given site. Costs of a pilot study or any other indicator development needs should be included if appropriate.

Phase 3: Response Variability

It is essential to understand the components of variability in indicator results to distinguish extraneous factors from a true environmental signal. Total variability includes both measurement error introduced during field and laboratory activities and natural variation, which includes influences of stressors. Natural variability can include temporal (within the field season and across years) and spatial (across sites) components. Depending on the context of the assessment question, some of these sources must be isolated and quantified in order to interpret indicator responses correctly. It may not be necessary or appropriate to address all components of natural variability. Ultimately, an indicator must exhibit significantly different responses at distinct points along a condition gradient. If an indicator is composed of multiple measurements, variability should be evaluated for each measurement as well as for the resulting indicator.

Guideline 8: Estimation of Measurement Error
The process of collecting, transporting, and analyzing ecological data generates errors that can obscure the discriminatory ability of an indicator. Variability introduced by human and instrument performance must be estimated and reported for all indicator measurements. Variability among field crews should also be estimated, if appropriate. If standard methods and equipment are employed, information on measurement error may be available in the literature. Regardless, this information should be derived or validated in dedicated testing or a pilot study.

Guideline 9: Temporal Variability - Within the Field Season
It is unlikely in a monitoring program that data can be collected simultaneously from a large number of sites. Instead, sampling may require several days, weeks, or months to complete, even though the data are ultimately to be consolidated into a single reporting period. Thus, within-field season variability should be estimated and evaluated. For some monitoring programs, indicators are applied only within a particular season, time of day, or other window of opportunity when their signals are determined to be strong, stable, and reliable, or when stressor influences are expected to be greatest. This optimal time frame, or index period, reduces temporal variability considered irrelevant to program objectives. The use of an index period should be defended and the variability within the index period should be estimated and evaluated.

Guideline 10: Temporal Variability - Across Years
Indicator responses may change over time, even when ecological condition remains relatively stable. Observed changes in this case may be attributable to weather, succession, population cycles or other natural inter-annual variations. Estimates of variability across years should be examined to ensure that the indicator reflects true trends in ecological condition for characteristics that are relevant to the assessment question. To determine inter-annual stability of an indicator, monitoring must proceed for several years at sites known to have remained in the same ecological condition.

Guideline 11: Spatial Variability
Indicator responses to various environmental conditions must be consistent across the monitoring region if that region is treated as a single reporting unit. Locations within the reporting unit that are known to be in similar ecological condition should exhibit similar indicator results. If spatial variability occurs due to regional differences in physiography or habitat, it may be necessary to normalize the indicator across the region, or to divide the reporting area into more homogeneous units.

Guideline 12: Discriminatory Ability
The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined. This analysis should incorporate all error components relevant to the program objectives, and separate extraneous variability to reveal the true environmental signal in the indicator data.

Phase 4: Interpretation and Utility

A useful ecological indicator must produce results that are clearly understood and accepted by scientists, policy makers, and the public. The statistical limitations of the indicator's performance should be documented. A range of values should be established that defines ecological condition as acceptable, marginal, and unacceptable in relation to indicator results. Finally, the presentation of indicator results should highlight their relevance for specific management decisions and public acceptability.

Guideline 13: Data Quality Objectives
The discriminatory ability of the indicator should be evaluated against program data quality objectives and constraints. It should be demonstrated how sample size, monitoring duration, and other variables affect the precision and confidence levels of reported results, and how these variables may be optimized to attain stated program goals. For example, a program may require that an indicator be able to detect a twenty percent change in some aspect of ecological condition over a ten-year period, with ninety-five percent confidence. With magnitude, duration, and confidence level constrained, sample size and extraneous variability must be optimized in order to meet the program's data quality objectives. Statistical power curves are recommended to explore the effects of different optimization strategies on indicator performance.

Guideline 14: Assessment Thresholds
To facilitate interpretation of indicator results by the user community, threshold values or ranges of values should be proposed that delineate acceptable from unacceptable ecological condition. Justification can be based on documented thresholds, regulatory criteria, historical records, experimental studies, or observed responses at reference sites along a condition gradient. Thresholds may also include safety margins or risk considerations. Regardless, the basis for threshold selection must be documented.

Guideline 15: Linkage to Management Action
Ultimately, an indicator is useful only if it can provide information to support a management decision or to quantify the success of past decisions. Policy makers and resource managers must be able to recognize the implications of indicator results for stewardship, regulation, or research. An indicator with practical application should display one or more of the following characteristics: responsiveness to a specific stressor, linkage to policy indicators, utility in cost-benefit assessments, limitations and boundaries of application, and public understanding and acceptance. Detailed consideration of an indicator's management utility may lead to a re-examination of its conceptual relevance and to a refinement of the original assessment question.

Application of the Guidelines

This document was developed both to guide indicator development and to facilitate indicator review. Researchers can use the guidelines informally to find weaknesses or gaps in indicators that may be corrected with further development. Indicator development will also benefit from formal peer reviews, accomplished through a panel or other appropriate means that bring experienced professionals together. It is important to include both technical experts and environmental managers in such a review, since the *Evaluation Guidelines* incorporate issues from both arenas. This document recommends that a review address information and data supporting the indicator in the context of the four phases described. The guidelines included in each phase are functionally related and allow the reviewers to focus on four fundamental questions:

Phase 1 - Conceptual Relevance: Is the indicator relevant to the assessment question (management concern) and to the ecological resource or function at risk?

Phase 2 - Feasibility of Implementation: Are the methods for sampling and measuring the environmental variables technically feasible, appropriate, and efficient for use in a monitoring program?

Phase 3 - Response Variability: Are human errors of measurement and natural variability over time and space sufficiently understood and documented?

Phase 4 - Interpretation and Utility: Will the indicator convey information on ecological condition that is meaningful to environmental decision-making?

Upon completion of a review, panel members should make written responses to each guideline. Documentation of the indicator presentation and the panel comments and recommendations will establish a knowledge base for further research and indicator comparisons. Information from ORD indicator reviews will be maintained with public access so that scientists outside of EPA who are applying for grant support can address the most critical weaknesses of an indicator or an indicator area.

It is important to recognize that the *Evaluation Guidelines* by themselves do not determine indicator applicability or effectiveness. Users must decide the acceptability of an indicator in relation to their specific needs and objectives. This document was developed to evaluate indicators for ORD-affiliated monitoring programs, but it should be useful for other programs as well. To increase its potential utility, this document avoids labeling individual guidelines as either essential or optional, and does not establish thresholds for acceptable or unacceptable performance. Some users may be willing to accept a weakness in an indicator if it provides vital information. Or, the cost may be too high for the information gained. These decisions should be made on a case-by-case basis and are not prescribed here.

Example Indicators

Ecological indicators vary in methodology, type (biological, chemical, physical), resource application (fresh water, forest, *etc.*), and system scale, among other ways. Because of the diversity and complexity of ecological indicators, three different indicator examples are provided in the following chapters to illustrate application of the guidelines. The examples include a direct measurement (dissolved oxygen concentration), an index (benthic condition) and a multimetric indicator (stream fish assemblages) of ecological condition. All three examples employ data from EMAP studies, but each varies in the type of information and extent of analysis provided for each guideline, as well as the approach and terminology used. The authors of these chapters present their best interpretations of the available information. Even though certain indicator strengths and weaknesses may emerge, the examples are *not* evaluations, which should be performed in a peer-review format. Rather, the presentations are intended to illustrate the types of information relevant to each guideline.

Chapter 2

Application of the Indicator Evaluation Guidelines to Dissolved Oxygen Concentration as an Indicator of the Spatial Extent of Hypoxia in Estuarine Waters

Charles J. Strobel, U.S. EPA, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI and James Heltshe, OAO Corporation, Narragansett, RI

This chapter provides an example of how ORD's indicator evaluation process can be applied to a simple ecological indicator - dissolved oxygen (DO) concentration in estuarine water.

The intent of these guidelines is to provide a process for evaluating the utility of an ecological indicator in answering a specific assessment question for a specific program. This is important to keep in mind because any given indicator may be ideal for one application but inappropriate for another. The dissolved oxygen indicator is being evaluated here in the context of a large-scale monitoring program such as EPA's Environmental Monitoring and Assessment Program (EMAP). Program managers developed a series of assessment questions early in the planning process to focus indicator selection and monitoring design. The assessment question being addressed in this example is *What percent of estuarine area is hypoxic/anoxic?* Note that this discussion is not intended to address the validity of the assessment question, whether or not other appropriate indicators are available, or the biological significance of hypoxia. It is intended only to evaluate the utility of dissolved oxygen measurements as an indicator of hypoxia.

This example of how the indicator evaluation guidelines can be applied is a very simple one, and one in which the proposed indicator, DO concentration, is nearly synonymous with the focus of the assessment question, hypoxia. Relatively simple statistical techniques were chosen for this analysis to illustrate the ease with which the guidelines can be applied. More complex indicators, as discussed in subsequent chapters, may require more sophisticated analytical techniques.

Phase 1: Conceptual Relevance

> ***Guideline 1: Relevance to the Assessment***
> *Early in the evaluation process, it must be demonstrated in concept that the proposed indicator is responsive to an identified assessment question and will provide information useful to a management decision. For indicators requiring multiple measurements (indices or aggregates), the relevance of each measurement to the management objective should be identified. In addition, the indicator should be evaluated for its potential to contribute information as part of a suite of indicators designed to address multiple assessment questions. The ability of the proposed indicator to complement indicators at other scales and levels of biological organization should also be considered. Redundancy with existing indicators may be permissible, particularly if improved performance or some unique and critical information is anticipated from the proposed indicator.*

In this example, the assessment question is: *What percent of estuarine area is hypoxic/anoxic?* Since hypoxia and anoxia are defined as low levels of oxygen and the absence of oxygen, respectively, the relevance of the proposed indicator to the assessment is obvious. It is important to note that, in this evaluation, we are examining the use of DO concentrations only to answer the specific assessment question, not to comment on the eutrophic state of an estuary. This is a much larger issue that requires additional indicators.

Guideline 2: Relevance to Ecological Function
It must be demonstrated that the proposed indicator is conceptually linked to the ecological function of concern. A straightforward link may require only a brief explanation. If the link is indirect or if the indicator itself is particularly complex, ecological relevance should be clarified with a description, or conceptual model. A conceptual model is recommended, for example, if an indicator is comprised of multiple measurements or if it will contribute to a weighted index. In such cases, the relevance of each component to ecological function and to the index should be described. At a minimum, explanations and models should include the principal stressors that are presumed to impact the indicator, as well as the resulting ecological response. This information should be supported by available environmental, ecological and resource management literature.

The presence of oxygen is critical to the proper functioning of most ecosystems. Oxygen is needed by aquatic organisms for respiration and by sediment microorganisms for oxidative processes. It also affects chemical processes, including the adsorption or release of pollutants in sediments. Low concentrations are often associated with areas of little mixing and high oxygen consumption (from bacterial decomposition).

Figure 2-1 presents a conceptual model of oxygen dynamics in an estuarine ecosystem, and how hypoxic conditions form. Oxygen enters the system from the atmosphere or via photosynthesis. Under certain conditions, stratification of the water column may occur, creating two layers. The upper layer contains less dense water (warmer, lower salinity). This segment is in direct contact with the atmosphere, and since it is generally well illuminated, contains living phytoplankton. As a result, the dissolved oxygen concentration is generally high. As plants in this upper layer die, they sink to the bottom where bacterial decomposition occurs. This process uses oxygen. Since there is generally little mixing of water between these two layers, oxygen is not rapidly replenished.

This may lead to hypoxic or anoxic conditions near the bottom. This problem is intensified by nutrient enrichment commonly caused by anthropogenic activities. High nutrient levels often result in high concentrations of phytoplankton and algae. They eventually die and add to the mass of decomposing organic matter in the bottom layer, hence aggravating the problem of hypoxia.

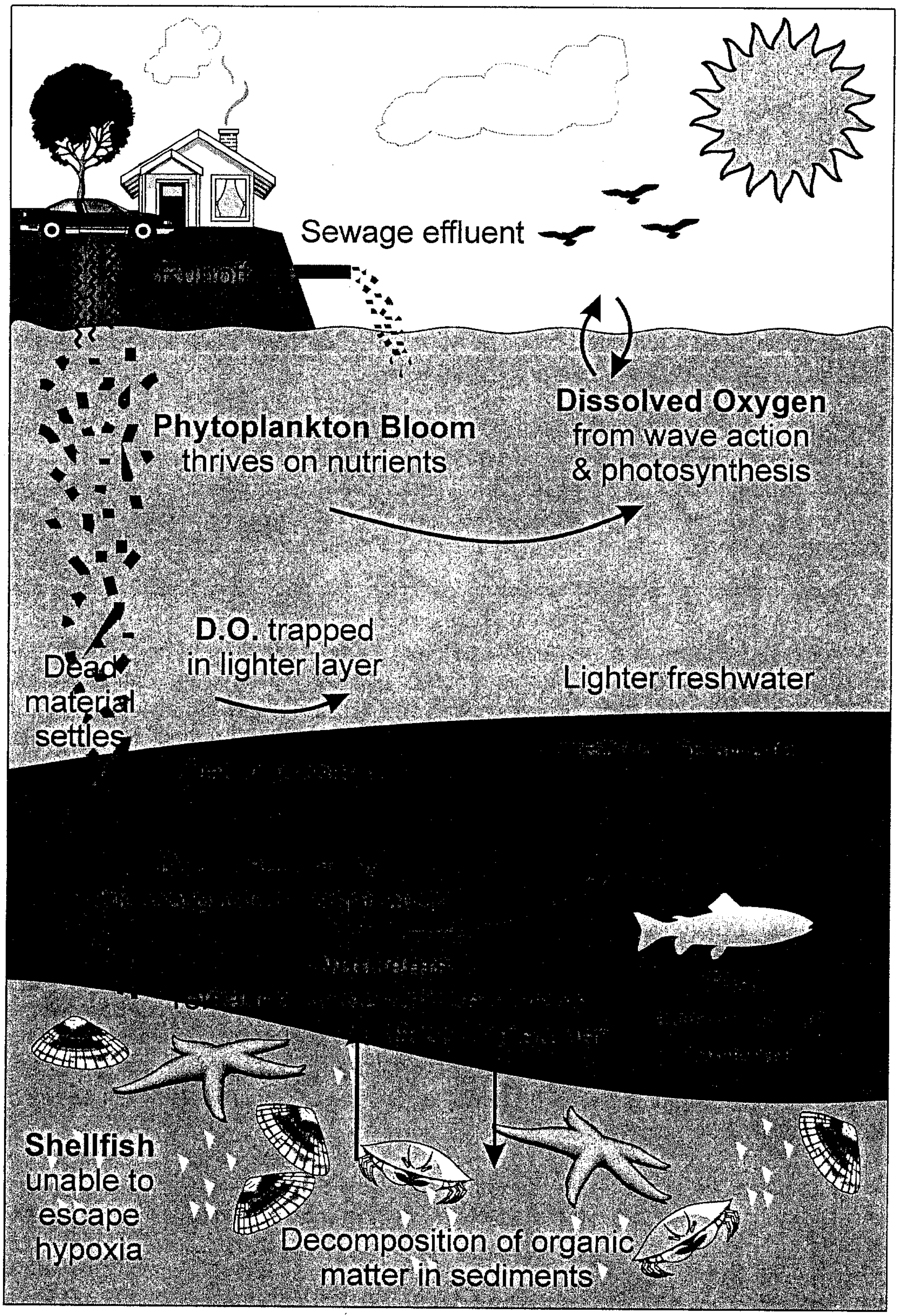


Figure 2-1. Conceptual model showing the ecological relevance of dissolved oxygen concentration in estuarine water.

Phase 2. Feasibility of Implementation

Guideline 3: Data Collection Methods

Methods for collecting all indicator measurements should be described. Standard, well-documented methods are preferred. Novel methods should be defended with evidence of effective performance and, if applicable, with comparisons to standard methods. If multiple methods are necessary to accommodate diverse circumstances at different sites, the effects on data comparability across sites must be addressed. Expected sources of error should be evaluated.

Methods should be compatible with the monitoring design of the program for which the indicator is intended. Plot design and measurements should be appropriate for the spatial scale of analysis. Needs for specialized equipment and expertise should be identified.

Sampling activities for indicator measurements should not significantly disturb a site. Evidence should be provided to ensure that measurements made during a single visit do not affect the same measurement at subsequent visits or, in the case of integrated sampling regimes, simultaneous measurements at the site. Also, sampling should not create an adverse impact on protected species, species of special concern, or protected habitats.

Once it is determined that the proposed indicator is relevant to the assessment being conducted, the next phase of evaluation consists of determining if the indicator can be implemented within the context of the program. Are well-documented data collection and analysis methods currently available? Do the logistics and costs associated with this indicator fit into the overall program plan? In some cases a pilot study may be needed to adequately address these questions. As described below, the answer to all these questions is yes for dissolved oxygen. Once again, this applies only to using DO to address the extent of hypoxia/anoxia for a regional monitoring program.

A variety of well-documented methods are currently available for the collection of dissolved oxygen data in estuarine waters. Electronic instruments are most commonly used. These include simple dissolved oxygen meters as well as more sophisticated CTDs (instruments designed to measure conductivity, temperature, and depth) equipped with DO probes. A less expensive, although more labor intensive method, is a Winkler titration. This "wet chemistry" technique requires the collection and fixation of a water sample from the field, and the subsequent titration of the sample with a thiosulphate solution either in the field or back in the laboratory. Because this method is labor intensive, it is probably not appropriate for large monitoring programs and will not be considered further. The remainder of this discussion will focus on the collection of DO data using electronic instrumentation.

Other variations in methodology include differences in sampling period, duration, and location. The first consideration is the time of year. Hypoxia is most severe during the summer months when water temperatures are high and the biota are most active. This is therefore the most appropriate time to monitor DO, and it is the field season for the program in which we are considering using this indicator. The next consideration is whether to collect data at a single point in time or to deploy an instrument to collect data over an extended period. Making this determination requires *a priori* knowledge of the DO dynamics of the area being studied. This issue will be discussed further in Guideline 9. For the purpose of this evaluation guideline, we will focus on single point-in-time measurements.

The third aspect to be considered is where in the water column to make the measurements. Because hypoxia is generally most severe near the bottom, a bottom measurement is critical. For this program, we will be considering a vertical profile using a CTD. This provides us with information on the DO concentration not only at the bottom, but throughout the water column. The additional information can be used to determine the depth of the pycnocline (a sharp, vertical density gradient in the water column), and potentially the volume of hypoxic water. Using a CTD instead of a DO meter provides ancillary information on the water column (salinity, temperature, and depth of the measurements). This information is needed to characterize the water column at the station, so using a CTD eliminates the need for multiple measurement with different instruments.

The proposed methodology consists of lowering a CTD through the water column to obtain a vertical profile. The instrument is connected to a surface display. Descent is halted at one meter intervals and the CTD held at that depth until the DO reading stabilizes. This process is continued until the unit is one meter above the bottom, which defines the depth of the bottom measurement.

Guideline 4: Logistics
The logistical requirements of an indicator can be costly and time-consuming. These requirements must be evaluated to ensure the practicality of indicator implementation, and to plan for personnel, equipment, training, and other needs. A logistics plan should be prepared that identifies requirements, as appropriate, for field personnel and vehicles, training, travel, sampling instruments, sample transport, analytical equipment, and laboratory facilities and personnel. The length of time required to collect, analyze and report the data should be estimated and compared with the needs of the program.

The collection of dissolved oxygen data in the manner described under Guideline 3 requires little additional planning over and above that required to mount a field effort involving sampling from boats. Collecting DO data adds approximately 15 to 30 minutes at each station, depending on water depth and any problems that may be encountered. The required gear is easily obtainable from a number of vendors (see Guideline 7 for estimated costs), and is compact, requiring little storage space on the boat. Each field crew should be provided with at least two CTD units, a primary unit and a backup unit. Operation of the equipment is fairly simple, but at least one day of training and practice is recommended before personnel are allowed to collect actual data.

Dissolved oxygen probes require frequent maintenance, including changing membranes. This should be conducted at least weekly, depending on the intensity of usage. This process needs to be worked into logistics as the membrane must be allowed to "relax" for at least 12 hours after installation before the unit can be recalibrated. In addition, the dissolved oxygen probe must be air-calibrated at least once per day. This process takes about 30 minutes and can be easily conducted prior to sampling while the boat is being readied for the day.

No laboratory analysis of samples is required for this indicator; however, the data collected by field crews should be examined by qualified personnel.

In summary, with the proper instrumentation and training, field personnel can collect data supporting this indicator with only minimal effort.

Guideline 5: Information Management
Management of information generated by an indicator, particularly in a long-term monitoring program, can become a substantial issue. Requirements should be identified for data processing, analysis, storage, and retrieval, and data documentation standards should be developed. Identified systems and standards must be compatible with those of the program for which the indicator is intended and should meet the interpretive needs of the program. Compatibility with other systems should also be considered, such as the internet, established federal standards, geographic information systems, and systems maintained by intended secondary data users.

This indicator should present no significant problems from the perspective of information management. Based on the proposed methodology, data are collected at one-meter intervals. The values are written on hard-copy datasheets and concurrently logged electronically in a surface unit attached to the CTD. (Note that this process will vary with the method used. Other options include not using a deck unit and logging data in the CTD itself for later uploading to a computer; or simply typing values from the hard-copy datasheet directly into a computer spreadsheet). After sampling has been completed, data from the deck unit can be uploaded to a computer and processed in a spreadsheet package. Processing would most likely consist of plotting out dissolved oxygen with depth to view the profile. Data should be uploaded to a computer daily. The user needs to pay particular attention to the memory size of the CTD or deck unit. Many instruments may contain sufficient memory for only a few casts. To avoid data loss it is important that the data be uploaded before the unit's memory is exhausted. The use of hard-copy datasheets provides a back-up in case of the loss of electronic data.

Guideline 6: Quality Assurance
For accurate interpretation of indicator results, it is necessary to understand their degree of validity. A quality assurance plan should outline the steps in collection and computation of data, and should identify the data quality objectives for each step. It is important that means and methods to audit the quality of each step are incorporated into the monitoring design. Standards of quality assurance for an indicator must meet those of the targeted monitoring program.

The importance of a well-designed quality assurance plan to any monitoring program cannot be overstated. One important aspect of any proposed ecological indicator is the ability to validate the results. Several methods are available to assure the quality of dissolved oxygen data collected in this example. The simplest method is to obtain a concurrent measurement with a second instrument, preferably a different type than is used for the primary measurement *(e.g.*, using a DO meter rather than a CTD). This is most easily performed at the surface, and can be accomplished by hanging both the CTD and the meter's probe over the side of the boat and allowing them to come to equilibrium. The DO measurements can then be compared and, if they agree within set specifications (*e.g.*, 0.5 mg/L), the CTD is assumed to be functioning properly. The DO meter should be air-calibrated immediately prior to use at each station. One could argue against the use of an electronic instrument to check another electronic instrument, but it is unlikely that both would be out of calibration in the same direction, to the same magnitude. An alternative method is to collect a water sample for Winkler titration; however, this would not provide immediate feedback. One would not know that the data were questionable until the sample is returned to the laboratory and it is too late to repeat the CTD cast. Although Winkler titrations can be performed in the field, the rocking of the boat can lead to erroneous titration.

Additional QA of the instrumentation can be conducted periodically in the laboratory under more controlled conditions. This might include daily tests in air-saturated water in the laboratory, with Winkler titrations verifying the results. Much of this depends upon the logistics of the program, for example, whether the program is run in proximity to a laboratory or remotely.

Three potential sources of error could invalidate results for this indicator: 1) improper calibration of the CTD, 2) malfunction of the CTD, and 3) the operator not allowing sufficient time for the instrument to equilibrate before each reading is taken. Taking a concurrent surface measurement should identify problems 1 and 2. The third source of error is more difficult to control, but can be minimized with proper training. If data are not uploaded directly from the CTD or surface unit into a computer, another source of error, transcription error, is also possible. However, this can be easily determined through careful review of the data.

Guideline 7: Monetary Costs
Cost is often the limiting factor in considering to implement an indicator. Estimates of all implementation costs should be evaluated. Cost evaluation should incorporate economy of scale, since cost per indicator or cost per sample may be considerably reduced when data are collected for multiple indicators at a given site. Costs of a pilot study or any other indicator development needs should be included if appropriate.

Cost is not a major factor in the implementation of this indicator. The sampling platform (boat) and personnel costs are spread across all indicators. As stated earlier, this indicator adds approximately 30 minutes to each station; however, one person can be collecting DO data while other crew members are collecting other types of data or samples.

The biggest expense is the equipment itself. Currently the most commonly used type of CTD costs approximately $6,000 each, the deck unit $3,000 and a DO meter approximately $1,500. A properly outfitted crew would need two of each, which totals $21,000. Assuming this equipment lasts for four years at 150 stations per year, the average equipment cost per station would be only $35. Expendable supplies (DO membranes and electrolyte) should be budgeted at approximately $200 per year, depending upon the size of the program.

Phase 3: Response Variability

Once it is determined that an indicator is relevant and can be implemented within the context of a specific monitoring program, the next phase consists of evaluating the expected variability in the response of that indicator. In this phase of the evaluation, it is very important to keep in mind the specific assessment question and the program design. For this example, the program is a large-scale monitoring program and the assessment question is focused on the spatial extent of hypoxia. This is very different from evaluating the hypoxic state at a specific station, as will be shown below in our evaluation of variability.

The data used in this evaluation come from two related sources. The majority of the data were collected as part of EMAP's effort in the estuaries of the Virginian Province (Cape Cod, MA to Cape Henry, VA) from 1990 to 1993. The distribution of sampling locations is shown in Figure 2-2. This effort is described in Holland (1990), Weisberg *et al.* (1993), and Strobel *et al.* (1995). Additional data from EPA's Mid-Atlantic Integrated Assessment (MAIA) program, collected in 1997, were also used. These data were collected in the small estuaries associated with Chesapeake Bay.

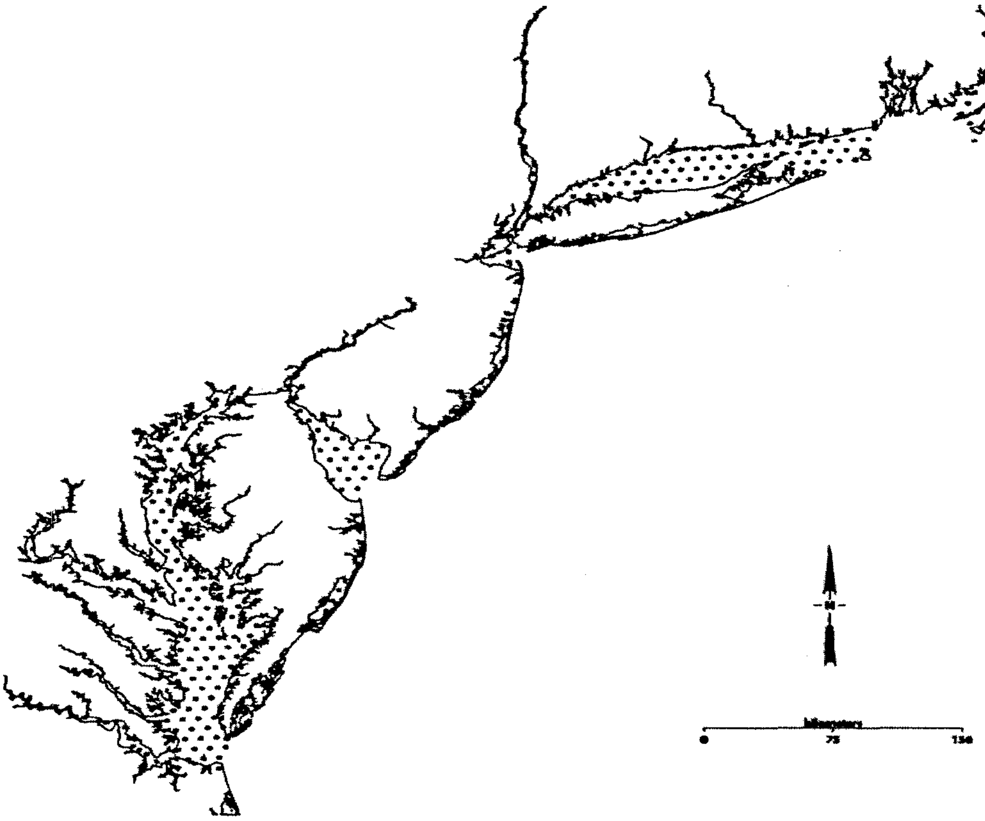


Figure 2-2. Each dot identifies an EMAP-Virginian Province station location in estuaries, 1990-1993.

Guideline 8: Estimation of Measurement Error
The process of collecting, transporting, and analyzing ecological data generates errors that can obscure the discriminatory ability of an indicator. Variability introduced by human and instrument performance must be estimated and reported for all indicator measurements. Variability among field crews should also be estimated, if appropriate. If standard methods and equipment are employed, information on measurement error may be available in the literature. Regardless, this information should be derived or validated in dedicated testing or a pilot study.

Using the QA information collected by EMAP over the period from 1991 to 1993 (a different method was employed in 1990, so those data were excluded from this analysis), we can estimate the error associated with this measurement. Figure 2-3 is a frequency distribution for 784 stations of the absolute difference between the DO measurements collected by the CTD and the DO meter used as a cross check (Δ DO). The data included in this figure were collected over three years by nine different field crews. Therefore, the figure illustrates the total measurement error--that associated with instrumentation as well as with operation of the instruments. Of the 784 stations, the measurement quality objective of $\leq$ 0.5 mg/L was met at over 90 percent. No bias was detected, meaning the CTD values were not consistently higher or lower than those from the DO meter.

It is of course possible to analyze instrumentation and operation errors separately. Such analyses would be necessary if total error exceeded a program's measurement quality objectives, in order to isolate and attempt to minimize the source of error. In fact, EMAP-Estuaries field crews conducted side-by-side testing during training to minimize between-crew differences. Good comparability between crews was achieved. However, because this was considered a training exercise, these data were not saved. Such side-by-side testing could be incorporated into any future analyses of the dissolved oxygen indicator. This would need to be conducted in the laboratory rather than in the field to eliminate the inherent temporal and spatial varability at any given site.

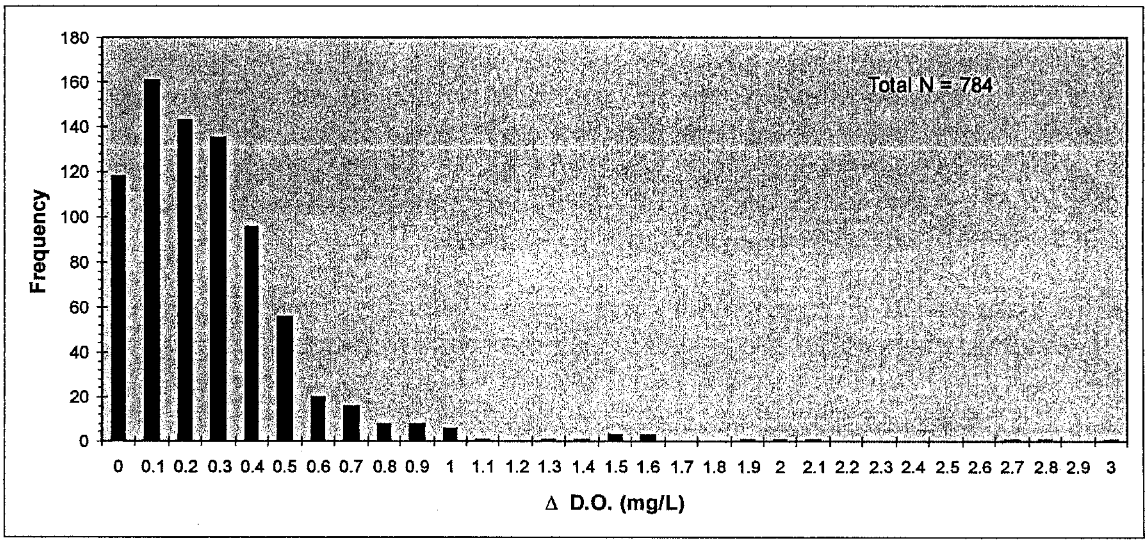


Figure 2-3. Frequency distribution of EMAP dissolved oxygen quality assurance data. Δ DO represents the absolute difference between the CTD measurement and that from a second instrument. Over 90% of the stations met the measurement quality objective (Δ DO $\leq$ 0.5 mg/L) .

Other potential sources of measurement error include inadequate thermal equilibration of the instrumentation prior to conducting a cast, and allowing insufficient time for the DO probe to repond to changes in DO concentration across an oxycline. Both can be addressed by proper training and evaluated by examining the full vertical profile for several parameters (*i.e.*, temperature and DO).

Guideline 9: Temporal Variability - Within the Field Season
It is unlikely in a monitoring program that data can be collected simultaneously from a large number of sites. Instead, sampling may require several days, weeks, or months to complete, even though the data are ultimately to be consolidated into a single reporting period. Thus, within-field season variability should be estimated and evaluated. For some monitoring programs, indicators are applied only within a particular season, time of day, or other window of opportunity when their signals are determined to be strong, stable, and reliable, or when stressor influences are expected to be greatest. This optimal time frame, or index period, reduces temporal variability considered irrelevant to program objectives. The use of an index period should be defended and the variability within the index period should be estimated and evaluated.

The dissolved oxygen concentration of estuarine water is highly dependent on a variety of factors, including photosynthesis (which is affected by nutrient levels), temperature, salinity, tidal currents, stratification, winds, and water depth. These factors make DO concentrations highly variable over relatively short time periods. There is also a strong seasonal component, with lowest dissolved oxygen concentrations experienced during the summer months of late July through September. In the EMAP program, estuarine monitoring was conducted during the summer when the biotic community is most active. Since we are interested in DO because of its effects on aquatic biota, and since summer is the season when organisms are most active and dissolved oxygen concentrations are generally the lowest, it is also the most appropriate season for evaluating the extent of hypoxic conditions. In 1990, EMAP conducted sampling in the Virginian Province estuaries to determine the most appropriate index period within the summer season. A subset of stations were sampled in each of three sampling intervals; 20 June to 18 July, 19 July to 31 August, and 1 September to 22 September. The results of analysis of the data collected at these stations showed the DO concentrations to be most consistent in Intervals 2 and 3, suggesting that July 19-September 22 is the most appropriate definition of the index period for the study area. Similar reconnaissance would need to be performed in other parts of the country where this indicator may be employed.

Even within the index period, DO concentrations at a given station vary hourly, daily and weekly. The high degree of temporal variability in DO at one station over the period from July 28 through August 26 is shown in Figure 2-4.

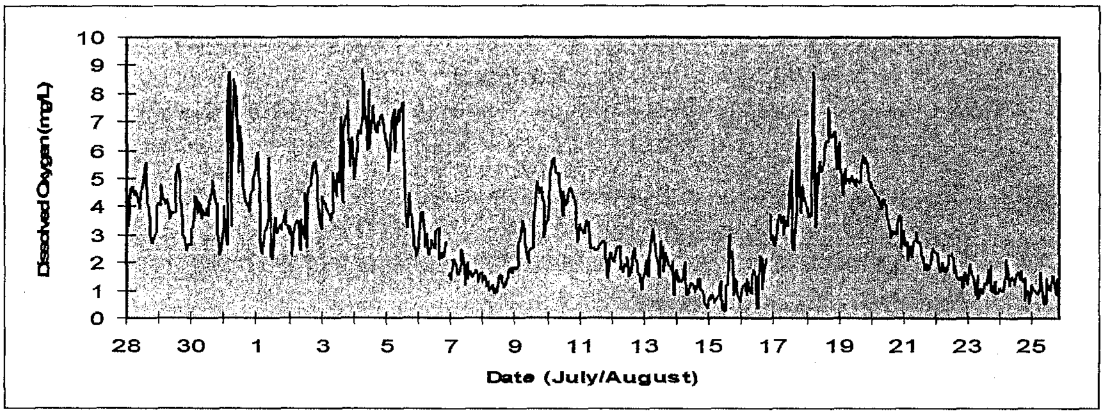


Figure 2-4. Continuous plot of bottom dissolved oxygen concentration at EMAP station 088 in Chesapeake Bay, 1990.

Figure 2-5 illustrates a 24-hour record of bottom dissolved oxygen from the same station. Although concentrations vary throughout the day, most mid-Atlantic estuaries generally do not exhibit a strong diurnal signal; most of the daily variability is associated with other factors such as tides (Weisberg *et al.* 1993). This is not the case in other regions, such as the Gulf of Mexico, where EMAP showed a strong diurnal signal (Summers *et al.* 1993). Such regional differences in temporal variability illustrate the need to tailor implementation of the indicator to the specific study area.

Short-term variability, as illustrated in Figures 2-4 and 2-5, makes this indicator, using single point-in-time measurements, inappropriate for characterizing a specific station. However, single stations are not the focus of the program for which this indicator is being evaluated in this example. The purpose of EMAP is to evaluate ecological condition across a broad geographic expanse, not at individual stations. The percent area hypoxic throughout the index period is more stable on a regional scale.

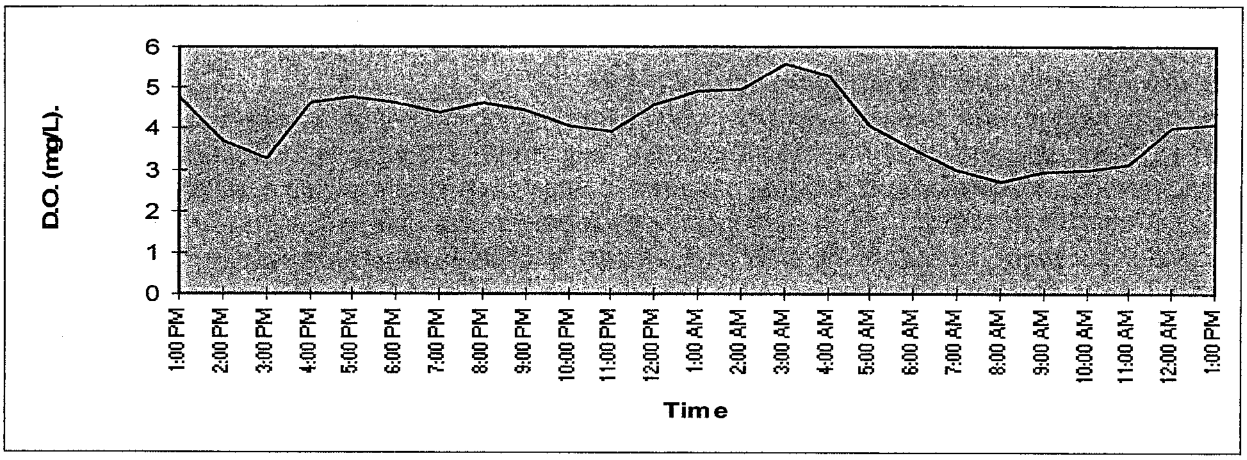


Figure 2-5. A 24-hour segment of the DO plot from Figure 2-4 showing a lack of strong diurnal signal.

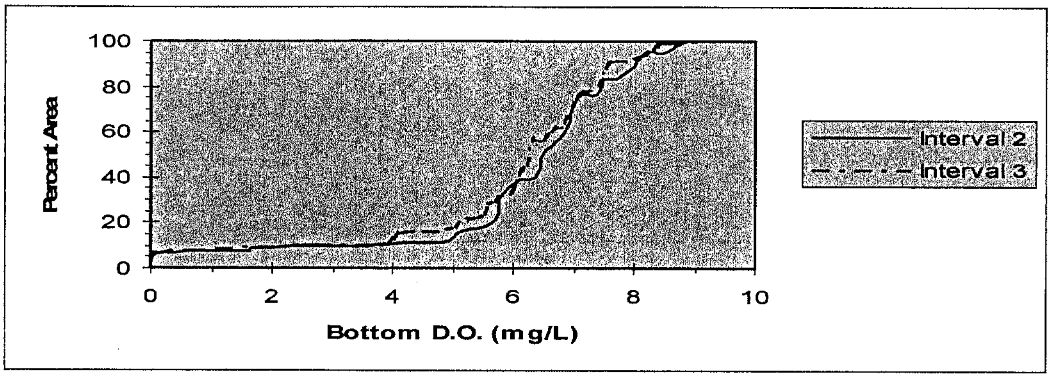


Figure 2-6. Comparison of cumulative distribution functions for EMAP-Virginian Province Intervals 2 and 3.

By plotting the cumulative distribution of bottom DO concentrations as a function of their weighted frequency (based on the spatial area represented by each station), we can estimate the percent area across the region with a DO concentration below any given value. Figure 2-6 shows cumulative distribution functions (CDFs) for point-in-time bottom DO measurements collected in 1990 during intervals 2 and 3 (*i.e.*, the index period). To determine the percent of estuarine area with a dissolved oxygen concentration less than 5 mg/L, one would look for the point on the y axis where the curve intersects the value of 5 on the x axis (*i.e.*, 15 to 20% in Figure 2-6). Confidence intervals can also be constructed around these CDFs, but were eliminated here for the sake of clarity. This figure shows that the percent area classified as hypoxic (*i.e.*, DO <5 or <2 mg/L) was approximately the same in the first half of the index period as it was in the second half. This stability makes this indicator appropriate for monitoring programs documenting the spatial extent of environmental condition on large (*i.e.*, regional) scales.

Guideline 10: Temporal Variability - Across Years
Indicator responses may change over time, even when ecological condition remains relatively stable. Observed changes in this case may be attributable to weather, succession, population cycles or other natural inter-annual variations. Estimates of variability across years should be examined to ensure that the indicator reflects true trends in ecological condition for characteristics that are relevant to the assessment question. To determine inter-annual stability of an indicator, monitoring must proceed for several years at sites known to have remained in the same ecological condition.

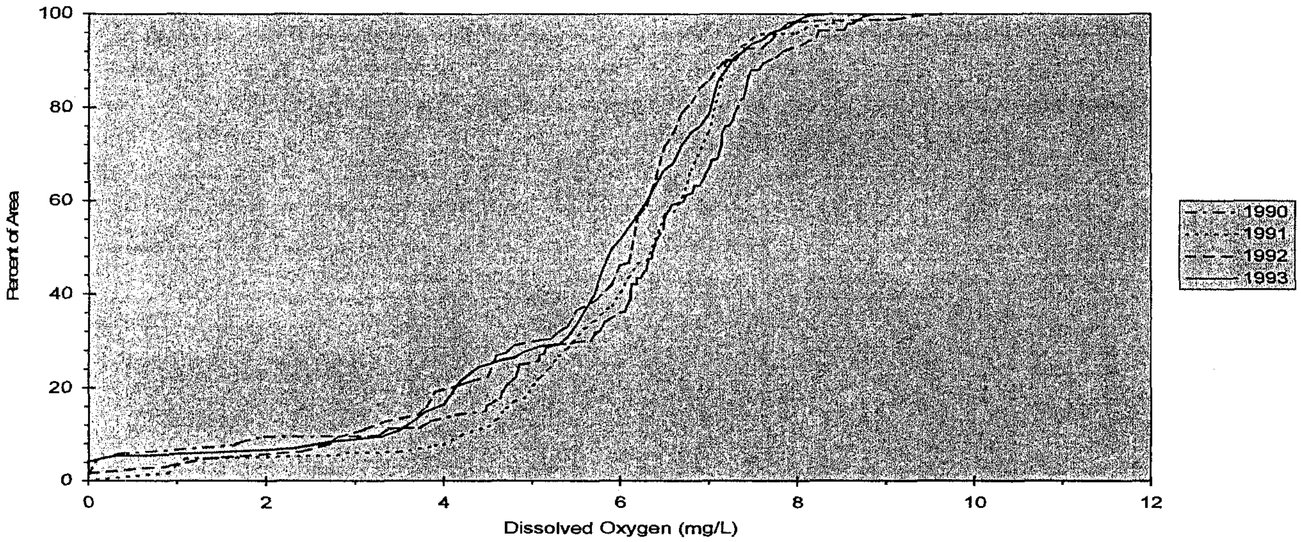


Figure 2-7. Annual cumulative distribution functions of bottom dissolved oxygen concentration for the Virginian Province, 1990-1993.

As discussed above, point-in-time DO measurements can be highly variable at any given station. This applies to across-year comparisons as well as within-season comparisons. However, when using this information to address the spatial extent of hypoxia across a broad region, this indicator is reasonably stable. Figure 2-7 shows the similarity of individual CDFs of bottom dissolved oxygen concentration in the estuaries of the Virginian Province for 1990 through 1993. Figure 2-8 shows the percent area below 5 and 2 mg/L (defined by EMAP as criteria for hypoxic and very hypoxic conditions, respectively) for those same years. Note that the percent area considered hypoxic by these criteria do not differ significantly from year to year, despite differences in climatic conditions (temperature and rainfall: Figure 2-9).

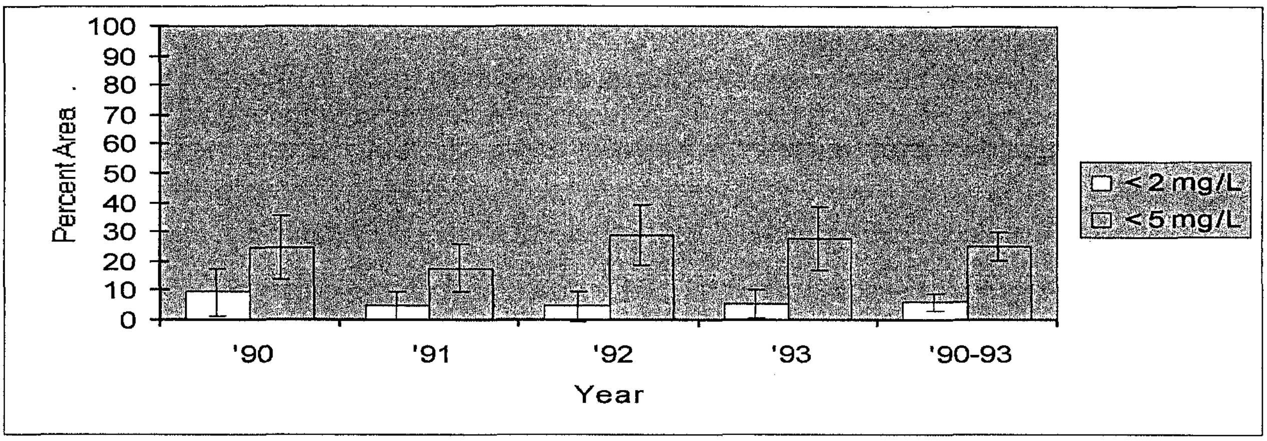


Figure 2-8. Annual estimates of percent area hypoxic in the Virginian Province based on EMAP dissolved oxygen measurements and criteria of < 2 mg/L and > 5 mg/L. '90-93 represents the four-year mean. Error bars represent 95% confidence intervals.

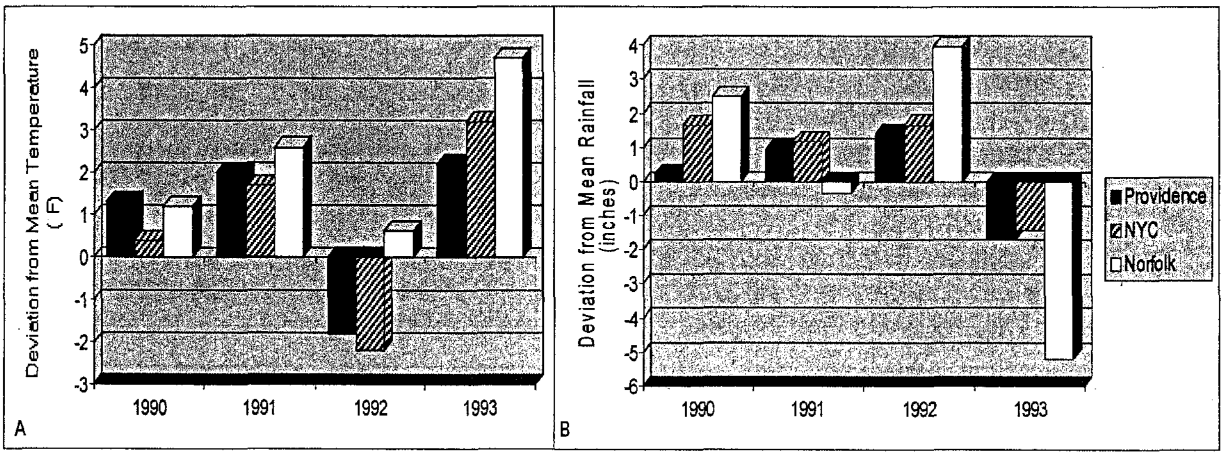


Figure 2-9. Climatic conditions in the Virginian Province, 1990-1993. (A) deviation from mean air temperature, and (B) deviation from mean rainfall.

Guideline 11: Spatial Variability
Indicator responses to various environmental conditions must be consistent across the monitoring region if that region is treated as a single reporting unit. Locations within the reporting unit that are known to be in similar ecological condition should exhibit similar indicator results. If spatial variability occurs due to regional differences in physiography or habitat, it may be necessary to normalize the indicator across the region, or to divide the reporting area into more homogeneous units.

Guideline 12: Discriminatory Ability
The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined. This analysis should incorporate all error components relevant to the program objectives, and separate extraneous variability to reveal the true environmental signal in the indicator data.

Since we are evaluating the use of dissolved oxygen concentration as an indicator of hypoxia, which is defined as a DO concentration below a certain value, there is no spatial variability associated with this indicator. Simply stated, using a criterion of 5 mg/L to define hypoxia, a DO concentration of 4 mg/L will indicate hypoxic conditions regardless of where the sample is collected. Note that this does NOT mean that adverse biological effects will always occur if the DO concentration falls below 5 mg/L. Nor does it mean that a given level of nutrient enrichment will result in the same degree of hypoxia in all areas. Both of these components are spatially variable and are affected by a number of environmental factors. However, they do not affect the *relationship* between dissolved oxygen concentration (the indicator) and hypoxia as defined (the asssessment issue).

Because of the large number of variables known to affect the dissolved oxygen concentration in sea water, most of which are not routinely measured, the utility of variability component analyses is limited. However, this indicator is really a direct measurement of the focus of the assessment question; therefore, discriminatory ability is inherently high.

Since the program's objective is to estimate the percent of estuarine area with hypoxic/anoxic condition on a broad geographic scale rather than to compare individual sites, an alternative way to look at this indicator's discriminatory ability is to plot out the CDF along with its confidence intervals. Figure 2-10 illustrates such a plot for the EMAP Virginian Province data collected from 1990 to 1993. (See Strobel *et al.* [1995] for a discussion of how the confidence intervals were developed). The tight 95% confidence intervals suggest that this indicator, as applied, has a high degree of discriminatory ability – a relatively small shift in the curve

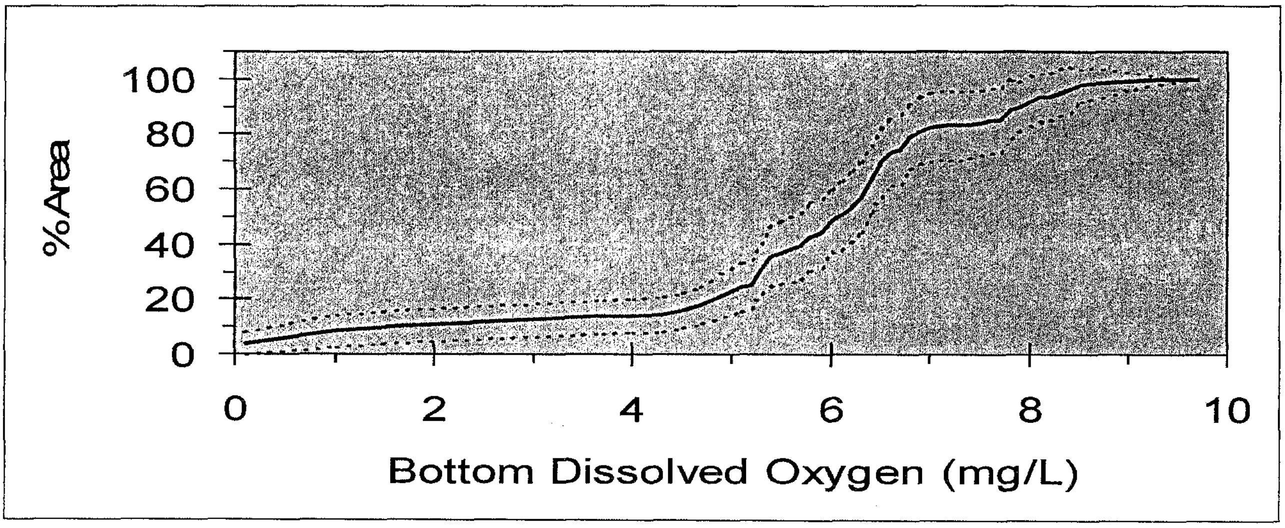


Figure 2-10. Cumulative distribution function of bottom dissolved oxygen concentration for the EMAP Virginian Province, 1990-1993. Error bars represent 95% confidence intervals.

can be determined to be significant. These confidence intervals are a function of both the variability in the data and the sampling design. Alternative approaches might be needed to evaluate the utility of this indicator for programs with significantly different designs.

Comparisons between curves can be made for those generated in two different regions (*i.e.*, status comparison) or from the same region at two different times (*i.e.*, trends comparison). Although this analysis does not separate out variability due to extraneous factors, it does provide insight into the utility of the indicator to discriminate condition using the design of the EMAP-Virginian Province program.

Phase 4: Interpretation and Utility

Once it is determined that the indicator is relevant, applicable, and responsive, the final phase of evaluation is to determine if the results can be clearly understood and useful.

Guideline 13: Data Quality Objectives
The discriminatory ability of the indicator should be evaluated against program data quality objectives and constraints. It should be demonstrated how sample size, monitoring duration, and other variables affect the precision and confidence levels of reported results, and how these variables may be optimized to attain stated program goals. For example, a program may require that an indicator be able to detect a twenty percent change in some aspect of ecological condition over a ten-year period, with ninety-five percent confidence. With magnitude, duration, and confidence level constrained, sample size and extraneous variability must be optimized in order to meet the program's data quality objectives. Statistical power curves are recommended to explore the effects of different optimization strategies on indicator performance.

The Data Quality Objective for trends in EMAP-Estuaries was to be able to detect a two percent change per year over 12 years with 90% confidence. This indicator meets that requirement as shown in Figure 2-11. This figure shows several power curves for annual changes ranging from one to three percent. Note that these curves are based on data from more than 400 stations sampled over a period of four years. The ability to detect trends will differ using different sampling designs. If fewer stations were to be sampled, a new set of power curves could be generated to show the ability to detect trends with that number of stations.

Guideline 14: Assessment Thresholds
To facilitate interpretation of indicator results by the user community, threshold values or ranges of values should be proposed that delineate acceptable from unacceptable ecological condition. Justification can be based on documented thresholds, regulatory criteria, historical records, experimental studies, or observed responses at reference sites along a condition gradient. Thresholds may also include safety margins or risk considerations. Regardless, the basis for threshold selection must be documented.

Although there is debate regarding their validity, assessment thresholds already exist for dissolved oxygen. Several states have adopted 5 mg/L as a criterion for 24-hour continuous concentrations and 2 mg/L as a point-in-time minimum concentration for supporting a healthy ecosystem. This is supported by EPA research (U.S. EPA 1998) which shows long-term effects at 4.6 mg/L and acute effects at 2.1 mg/L. If these thresholds change in the future, data collected on this indicator can easily be re-analyzed to produce new assessments of the hypoxic area.

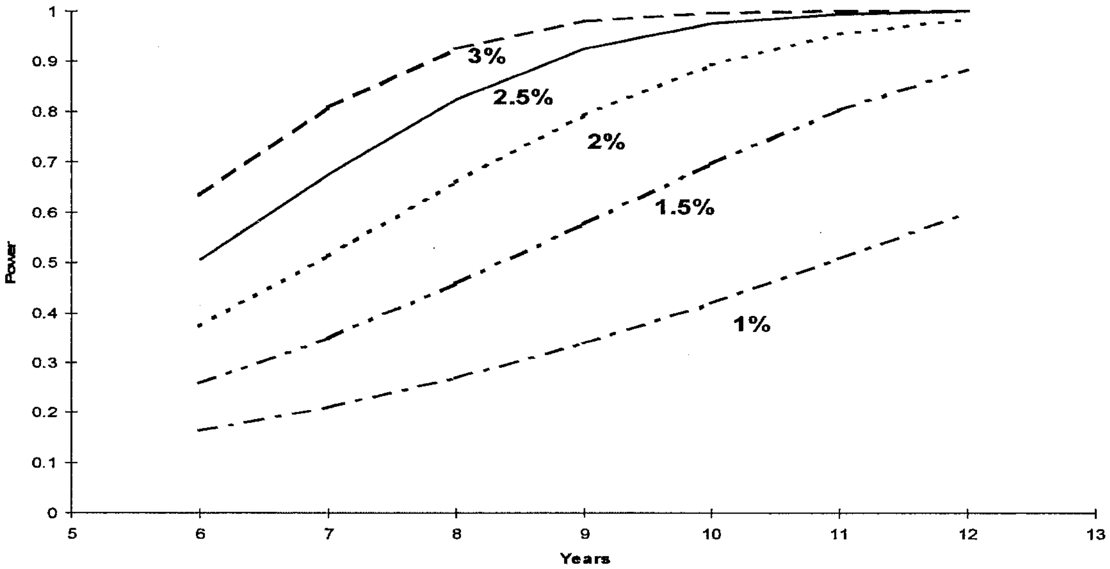


Figure 2-11. Power curves for detecting annual changes of 1, 1.5, 2, 2.5 and 3% in the percent of area exhibiting hypoxia based on EMAP Virginian Province data, 1990-1993.

Guideline 15: Linkage to Management Action

Ultimately, an indicator is useful only if it can provide information to support a management decision or to quantify the success of past decisions. Policy makers and resource managers must be able to recognize the implications of indicator results for stewardship, regulation, or research. An indicator with practical application should display one or more of the following characteristics: responsiveness to a specific stressor, linkage to policy indicators, utility in cost-benefit assessments, limitations and boundaries of application, and public understanding and acceptance. Detailed consideration of an indicator's management utility may lead to a re-examination of its conceptual relevance and to a refinement of the original assessment question.

Currently, hypoxia and eutrophication are important issues, particularly in the mid-Atlantic states. Millions of dollars are being spent on sewage treatment upgrades and controls for non-point sources. If these actions are successful, they will result in a decrease in the percent area with low dissolved oxygen, making this indicator important for measuring the efficacy of these efforts.

Mitigation of hypoxia is a complicated issue. Sewage treatment plants, non-point source pollution, and a variety of natural sources introduce nutrients to our estuaries. Increased nutrients can lead to hypoxic conditions, but the effects of hypoxia are not always easy to predict. For example, increased turbidity may inhibit phytoplankton growth, which, through a series of complicated interactions, may decrease a system's susceptibility to reductions in DO. Management interest is not necessarily to reduce nutrient levels, but to protect the biota of our estuaries from hypoxic conditions, which is exactly what this indicator is measuring.

Summary

The results of this evaluation show that point-in-time bottom dissolved oxygen measurement can be an appropriate indicator for determining the spatial extent of hypoxia in a regional monitoring program. The indicator is conceptually relevant to both the assessment question and ecological function. It is easily implemented at reasonable cost with well-defined methods. Probably the greatest concern in the implementation of this indicator is the temporal and spatial variability of DO concentrations. This variability limits the utility of point-in-time measurements in describing the conditions at a given station. However, when the indicator is applied across a large region to generate an estimate of the overall percent area hypoxic, this evaluation indicates reasonable stability of the indicator. This scale-dependent conclusion clearly illustrates the need to evaluate an indicator in the context of a specific monitoring program, as an indicator that may be ideal for one type of program may be inappropriate for another. In this case, the indicator itself could be applied to monitoring programs designed to characterize conditions at individual stations if alternative methods were employed (*e.g.,* continuous monitoring). Lastly, dissolved oxygen data are easily interpretable relative to the assessment question on the extent of hypoxia, and are of high value to environmental managers.

Acknowledgements

This work was supported in part by EPA contract number 68-W5-0065 to OAO Corporation. This chapter is NHEERL-Atlantic Ecology Division contribution # NHEERL-NAR-2068.

References

Holland, A.F. (ed.). 1990. Near Coastal Program Plan for 1990: Estuaries. EPA/600/4- 900/033. U.S. Environmental Protection Agency, Office of Research and Development, Narragansett, RI.

Strobel, C.J., H.W. Buffum, S.J. Benyi, E.A. Petrocelli, D.R. Reifsteck, and D.J. Keith. 1995. Statistical Summary: EMAP-Estuaries Virginian Province - 1990-1993. EPA/620/R-94/026.

Summers, J.K., J.M. Macauley, V.D. Engle, G.T. Brooks, P.T. Heitmuller, and M.T. Adams. 1993. Louisianian Province Demonstration Project Report: EMAP-Estuaries-1991. U.S. Environmental Protection Agency, Office of Research and Development, Gulf Breeze, FL.

U.S. EPA, 1992. What is the Long Island Sound Study? Fact sheet # 15, U.S. Environmental Protection Agency, Long Island Sound Study, Stamford, CT.

U.S. EPA. 1998. (Draft) Ambient Water Quality Criteria - Dissolved Oxygen. U.S. Environmental Protection Agency, Office of Water, Washington, DC. (In Review).

Weisberg, S.B., J.B. Frithsen, A.F. Holland, J.F. Paul, K.J. Scott, J.K. Summers, H.T. Wilson, R.M.Valente, D.G. Heimbuch, J. Gerritsen, S.C. Schimmel, and R.W. Latimer. 1993. EMAP-Estuaries, Virginian Province 1990 Demonstration Project Report. EPA/620/R- 93/006. U.S. Environmental Protection Agency, Office of Research and Development, Narragansett, RI.

Chapter Three

Application of the Indicator Evaluation Guidelines to an Index of Benthic Condition for Gulf of Mexico Estuaries

Virginia D. Engle, U.S. EPA, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL

This section provides an example of how the *Evaluation Guidelines for Ecological Indicators* can be applied to a multimetric ecological indicator - a benthic index for estuarine waters.

The intent of the *Evaluation Guidelines* is to provide a process for evaluating the utility of an ecological indicator in answering a specific assessment question for a specific program. This is important to keep in mind because any given indicator may be ideal for one application but inappropriate for another. The benthic index is evaluated here in the context of a large-scale monitoring program, specifically EPA's Environmental Monitoring and Assessment Program - Estuaries (EMAP-E). Program managers developed a series of assessment questions early in the planning process and focused the monitoring design accordingly.

One of the primary goals of EMAP-E was to develop and monitor indicators of pollution exposure and habitat condition in order to determine the magnitude and geographical distribution of resources that are adversely affected by pollution and other environmental stresses (Messer *et al.* 1991). In its first year of implementation in the estuaries of the Gulf of Mexico, EMAP-E collected data to develop a preliminary assessment of the association between benthic communities, sediment contamination and hypoxia. A benthic index of estuarine integrity was developed that incorporated measures of community composition and diversity, and discriminated between areas of undegraded *vs.* degraded environmental conditions. In this way, a benthic index would reflect the collective response of the benthic community to pollution exposure or adverse habitat conditions.

Information gained from monitoring benthic macroinvertebrate communities has been widely used to measure the status of and trends in the ecological condition of estuaries. Benthic macroinvertebrates are good indicators of estuarine condition because they are relatively sedentary within the sediment-water interface and deeper sediments (Dauer *et al.* 1987). Both short-term disturbances such as hypoxia and long-term disturbances such as accumulation of sediment contaminants affect the population and community dynamics of benthic macroinvertebrates (Rosenberg 1977, Harper *et al.* 1981, Rygg 1986). Many of the effects of such disturbances on the benthos have been documented and include changes in indicators such as benthic diversity, long-lived to short-lived species, biomass, abundance of opportunistic or pollution-tolerant organisms, and the trophic or functional structure of the community (Pearson and Rosenberg 1978, Santos and Simon 1980, Gaston 1985, Warwick 1986, Gaston and Nasci 1988, Gaston and Young 1992).

The search for an index that both integrates parameters of macrobenthic community structure and distinguishes between polluted and unpolluted areas has been a recent focus of marine and estuarine benthic monitoring programs (Warwick 1986, Chapman 1989, McManus and Pauly 1990). An ideal indicator of the response of benthic organisms to perturbations in the environment would not only quantify their present condition in ecosystems but also would integrate the effects of anthropogenic and natural stressors on the organisms over time (Boesch and Rosenberg 1981, Messer *et al.* 1991).

Recently, researchers have successfully developed multimetric indices that combine the various effects of natural and anthropogenic disturbances on benthic communities. Although initially developed for freshwater systems (Lenat 1988, Lang *et al.* 1989, Plafkin *et al.* 1989, Kerans and Karr 1994, Lang and Reymond 1995), variations of the benthic index of biotic integrity (B-IBI) concept have been successfully applied to estuaries (Engle *et al.* 1994, Ranasinghe *et al.* 1994, Weisberg *et al.* 1997, Engle and Summers 1999, Van Dolah *et al.* 1999). There are some basic differences between the approach we have used and the traditional IBI approach. The parameters that comprise our benthic index were chosen empirically as the parameters that provided the best statistical discrimination between sites with known degraded or undegraded conditions (where degraded is defined as having undesirable or unacceptable ecological condition). The weighting factors applied to these parameters were also determined empirically based on the contribution of each parameter to the fit of the model. The parameters included in a traditional IBI approach were chosen by the researchers based on evaluations of cumulative ecological dose response curves. The rank scoring of each parameter (*e.g.*, as a 1, 3, or 5) was based on a subjective weighting of the distribution of values from known sites. The parameters in the IBI are equally weighted in the calculation of the overall rank score. Both approaches to developing multimetric indices have advantages and criticisms; however, the ultimate goal is the same - to combine complex community information into a meaningful index of condition.

Multimetric benthic indices can help environmental managers who require a standardized means of tracking the ecological condition of estuaries. However, environmental managers and policy makers also desire an easy, manageable method of identifying the extent of potentially degraded areas and a means of associating biotic responses with environmental stressors (Summers *et al.* 1995). In order for an indicator to be appropriate for the assessment of estuarine health, it should incorporate geographic variation and should recognize the inherent multivariate nature of estuarine systems (Karr 1993, Wilson and Jeffrey 1994). While the statistical methods used to develop indicators may often be complex, it is the end product, an index of condition, that is of interest to resource managers. By applying a mathematical formula to multivariate benthic data, resource managers can calculate a single, scaled index that can then be used to evaluate the benthic condition of estuaries in their region. Although indices have been accused of oversimplifying or overgeneralizing biological processes, they play an important role in resource management (*i.e.*, to provide criteria with which to characterize a resource as impaired or healthy) (Rakocinski *et al.* 1997). While ecological indicators were developed to serve as tools for the preliminary assessment of ecological condition, they are not intended to replace a complete analysis of the benthic biological dynamics nor were they intended to stand alone. They also should be used in conjunction with other synoptic data on sediment toxicity and pollutant concentrations to provide a weight-of-evidence basis for judging the incidence of anthropogenically induced disturbances (Hyland *et al.* 1998).

Phase 1: Conceptual Relevance

Guideline 1: Relevance To The Assessment

Early in the evaluation process, it must be demonstrated in concept that the proposed indicator is responsive to an identified assessment question and will provide information useful to a management decision. For indicators requiring multiple measurements (indices or aggregates), the relevance of each measurement to the management objective should be identified. In addition, the indicator should be evaluated for its potential to contribute information as part of a suite of indicators designed to address multiple assessment questions. The ability of the proposed indicator to complement indicators at other scales and levels of biological organization should also be considered. Redundancy with existing indicators may be permissible, particularly if improved performance or some unique and critical information is anticipated from the proposed indicator.

The Environmental Monitoring and Assessment Program (EMAP) focused on providing much needed information about the condition of the Nation's ecological resources. EMAP was designed to answer the following questions (Summers *et al.* 1995):

1. What is the status, extent, and geographical distribution of our ecological resources?
2. What proportions of these resources are declining or improving? Where? At what rate?
3. What factors are likely to be contributing to declining conditions?
4. Are pollution control, reduction, mitigation, and prevention programs achieving overall improvement in ecological condition?

To accomplish these management objectives, EMAP sought to develop a suite of indicators that would represent the response of biota to environmental perturbations. These indicators were categorized as response, exposure, habitat, or stressor. Our indicator, the benthic index, was classified as a response indicator because it represents the response of the estuarine benthic community to environmental stressors (*e.g.*, sediment contaminants and hypoxia). A good response indicator should demonstrate the ability to associate responses with well-defined exposures. EMAP-Estuaries (EMAP-E) sought to apply these management objectives to the development of indicators to represent the condition of estuaries.

The specific assessment question addressed by the benthic index emerged from a hierarchy of assessment questions that were relevant to EMAP-E management goals. The broad assessment question for EMAP-E is: *What is the condition of estuaries?* Our project was geographically limited to estuaries in the Gulf of Mexico; therefore, our regional assessment question became: *What percent of estuarine area in the Gulf of Mexico is in good (or degraded) ecological condition?* Because biological integrity is one component of ecological condition, the next logical assessment question was: *What percent of estuarine area in the Gulf of Mexico exhibited acceptable (or unacceptable) biological integrity?* The condition of benthic biota is one measure of biological integrity. This tenet led to the specific assessment question addressed by the benthic index: *What percent of estuarine area has degraded benthic communities?* As a response indicator for estuaries, the benthic index was intended to contribute information to the broad assessment question above and to be used in conjunction with a suite of indicators to evaluate the overall condition of estuaries.

Macroinvertebrates provide an ideal measure of the response of the benthic community to environmental perturbations for many reasons (*e.g.*, see Boesch and Rosenberg 1981, Reish 1986). Benthos are primarily sedentary and, thus, have limited escape mechanisms to avoid disturbances (Bilyard 1987). Benthic invertebrates are relatively easy to monitor and tend to reflect the cumulative impacts of environmental perturbations, thereby providing good indications of the changes in an ecosystem over time. They have been used extensively as indicators of the impacts of both pollution and natural fluctuations in the estuarine environment (Gaston *et al.* 1985, Bilyard 1987, Holland *et al.* 1987, Boesch and Rabalais 1991). Benthic assemblages are often comprised of a variety of species (across multiple phyla) that represent a range of biotic responses to potential pollutant impacts.

The concept behind development of our benthic index begins with the assumption that adverse environmental conditions (*e.g.*, hypoxia and sediment contamination) affect benthic communities in predictable ways. The basic tenets of Pearson and Rosenberg (1978) for organic pollution provide a good example of the biological principles that operate in benthic communities. Pollution induces a decrease in diversity in favor of (sometimes high) abundances of relatively few species labeled as pollution-tolerant or opportunist. In pristine areas or areas unaffected by pollution, benthic communities exhibit higher diversity and stable populations of species labeled as pollution-sensitive or equilibrium. In general, although pollution-tolerant species may thrive in relatively undegraded areas, the converse is almost never true - pollution-sensitive species do not normally exist in polluted areas. Groups composed of higher-order levels of taxonomy are more often used in this context than a single indicator species. The fact that indicator species are not found in certain areas may be due to factors other than environmental condition (*i.e.*, inability to disperse, seasonal absence, or biotic competition; Sheehan 1984).

Through a mathematical process of determining which components of the benthic community best discriminate between degraded and undegraded sites, the benthic index is composed of the following parameters: diversity, proportional abundance of capitellids, bivalves, and amphipods, and the abundance of tubificids. Each parameter is directly or indirectly related to the condition of the benthic community. Each component of the benthic index can be and has been used individually as an indicator of benthic community condition in various monitoring programs. For our purposes, however, none of the components retains an individual relevance to the management objective. The strength of an indicator like the benthic index lies in the checks and balances associated with combining these components.

The benthic index does, indeed, indicate if a site has a degraded benthic community and this index is used by EMAP-E to compute the proportion of estuarine area with this subnominal condition. The benthic index was intended to be part of an overall assessment of ecological condition of estuaries that incorporated indicators of biological integrity, sediment and water quality, and aesthetic values. As a component of biological integrity, the benthic index provides insight into one aspect of the biotic community; complementary indicators could be developed for fish, zooplankton, and phytoplankton if sufficient data were available. One step in the validation of the benthic index showed that the benthic index had a greater success rate in classifying degraded or undegraded sites than any of its components did individually. Although the component parameters that make up the benthic index are useful for specific assessments, by combining them, the benthic index provides a more comprehensive assessment of benthic condition without being redundant.

Guideline 2: Relevance to Ecological Function

It must be demonstrated that the proposed indicator is conceptually linked to the ecological function of concern. A straightforward link may require only a brief explanation. If the link is indirect or if the indicator itself is particularly complex, ecological relevance should be clarified with a description, or conceptual model. A conceptual model is recommended, for example, if an indicator is comprised of multiple measurements or if it will contribute to a weighted index. In such cases, the relevance of each component to ecological function and to the index should be described. At a minimum, explanations and models should include the principal stressors that are presumed to impact the indicator, as well as the resulting ecological response. This information should be supported by available environmental, ecological and resource management literature.

Benthos are vital to ecosystem structure and function as a food resource for demersal fish and as intermediate links between higher and lower trophic levels. They provide a significant transfer of carbon in the energy dynamics of an estuary, and act as agents of bioturbation and nutrient regeneration (Flint *et al.* 1982). Benthic organisms often provide the first step in the bioaccumulation of pollutants in estuarine food chains, especially heavy metals. An index of environmental condition based on benthos, therefore, would provide useful information for management decisions based on long-term trend analysis, spatial patterns of enrichment or contamination, or the recognition of "hot spots" exhibited by total defaunation. An ideal indicator that incorporates the characteristics of benthic community structure would be sensitive to contaminant and dissolved oxygen stress and serve as a good integrator of estuarine sediment quality (Scott 1990).

Adverse environmental conditions that may have human influences and may affect the benthic community can be grouped into five general categories (Karr 1991, 1993; Fig. 3-1):

1. Water & Sediment Quality - hypoxia, salinity, temperature, contaminants
2. Habitat Structure - substrate type, water depth, complexity of physical habitat
3. Flow Regime - water volume and season flow distributions
4. Energy Source - characteristics of organic material entering waterbody
5. Biotic Interactions - competition, predation, disease, parasitism.

For the purposes of this assessment, we sought to evaluate the effects of water and sediment quality (specifically, contaminants and hypoxia) on the benthic community. While the other factors are equally important in determining benthic community structure, they were not included in this assessment. Figure 3-1 illustrates the primary pathways by which contaminants enter estuaries. Contaminants enter the estuary primarily via land-based non-point sources (*e.g.*, runoff from agricultural or livestock operations or urban runoff) and point sources (*e.g.*, industrial effluent or municipal wastewater). Contaminant stress is evident if the sediments are toxic to test organisms or if the levels of certain chemicals are high when compared to established guidelines. The benthic community responds to contaminant stress by an overall reduction in abundance and number of species, an increase in the proportion of pollution-tolerant or opportunistic species, or both.

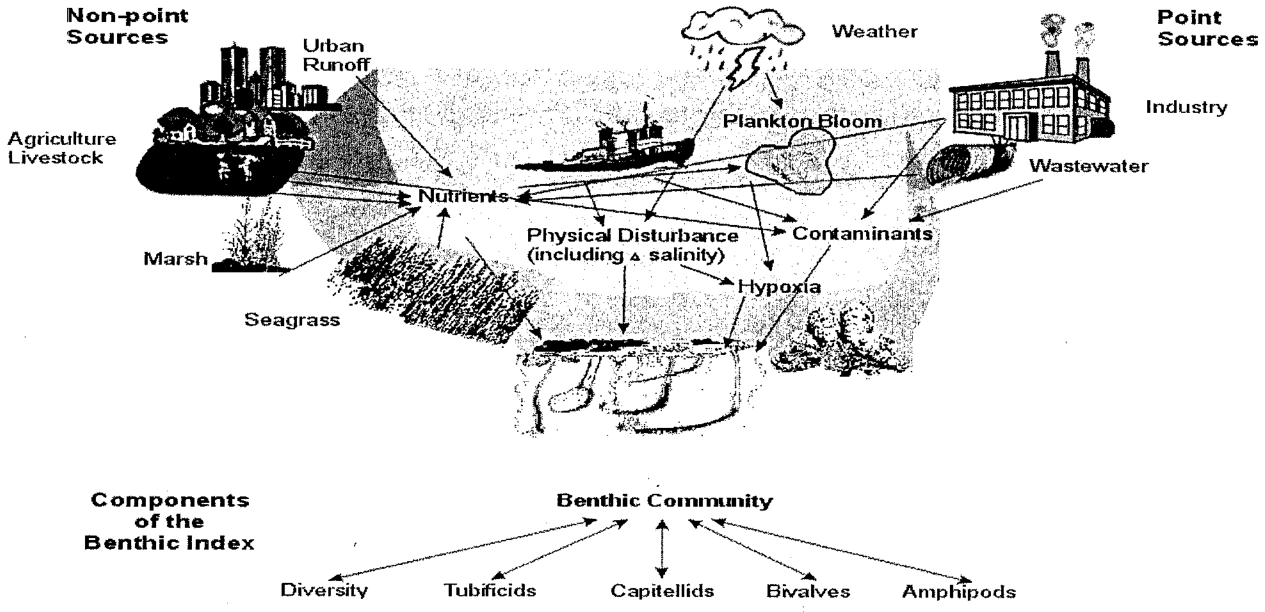


Figure 3-1. Conceptual diagram of a typical estuary showing the environmental stressors that may contribute to altering benthic macroinvertebrate community structure and the components of the benthic index.

Nutrients enter an estuarine system via the same point and non-point source pathways as contaminants but may also come from atmospheric deposition. Excess nutrients that lead to eutrophication can cause shifts in species composition and abundance. Eutrophication can deplete the oxygen in the bottom waters or climatic conditions may drive hypoxia because they influence the stratification of the water column. Oxygen depletion causes acute stress to the benthic community resulting in die-offs or chronic stress that may lead to shifts in species composition. Habitat disturbance includes physical scouring of the bottom as a result of storms, trawling, or dredging, as well as salinity and temperature changes brought about by climatic events. Because any of these stressors may induce alterations of the benthic community, monitoring community changes reflects the environmental conditions to which the benthos are exposed.

The benthic index represents the response of the benthic community to stressors like contaminants and hypoxia (Fig. 3-2). As a multimetric indicator, the benthic index is composed of the community measures that best discriminate between stressed and unstressed areas. Although these components (diversity, proportional abundance of capitellids, bivalves, and amphipods, and the abundance of tubificids) were chosen empirically by statistical analyses, they have biological relevance to the function of the benthic community. Diversity is directly related to the relative stability of a community. Sites that have been affected by contamination or hypoxia exhibit lower diversity than sites that have not been so adversely affected. Capitellid worms are often regarded as opportunists because of their high reproductive rate, small body size, and short life span. They are often found in great abundance in organically enriched areas and are usually the first to colonize disturbed sediments. Tubificid worms have life histories similar to capitellids but may extend the range of habitats in which capitellids are commonly found. Bivalves are usually indicative of stable environments because most bivalves are large-bodied, have slower reproductive rates, and longer life spans than worms. As filter feeders, most bivalves are the first to show signs of stress when water quality becomes unsuitable. Amphipods have long been used as test organisms in toxicity tests because of their demonstrated sensitivity to contaminants.

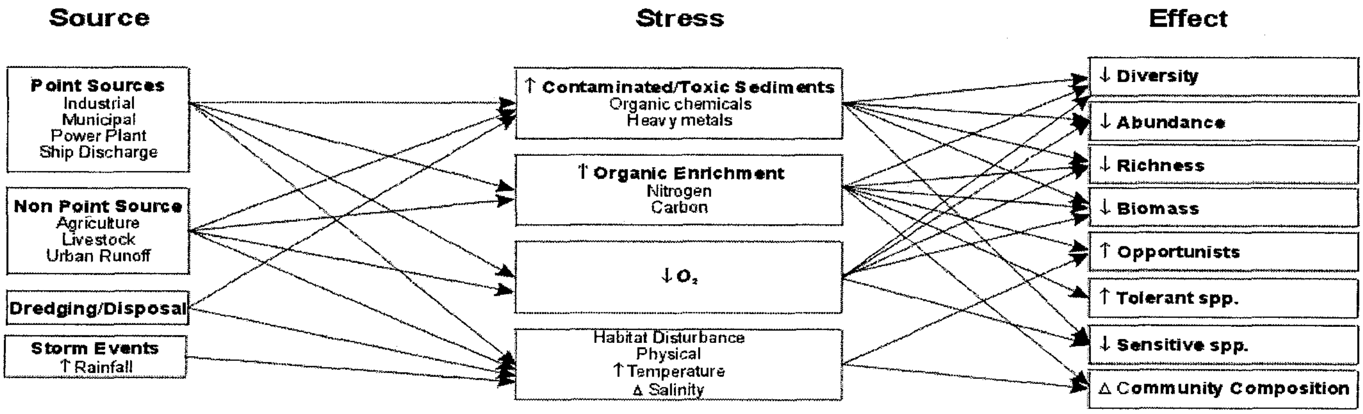


Figure 3-2. Conceptual model of the indicator, showing linkages between sources of stress, types of stress, and effect on the benthic community.

Phase 2: Feasibility of Implementation

Given that aspects of the benthic macroinvertebrate community can be used as indicators of degraded areas, how does a benthic index represent degraded benthic communities? The development of the benthic index is based on a simple question: if you have a set of known degraded and undegraded areas, what components of the benthic community best discriminate between the two? The degraded and undegraded sites were chosen from the EMAP-E database according to strict criteria involving environmental parameters that are known to affect the benthic community (*e.g.*, hypoxia, sediment contaminants, and sediment toxicity). This was done without prior knowledge of the benthic communities at those sites. A suite of parameters that represented aspects of benthic community structure and function were compiled for these sites (Engle *et al.* 1994, Engle and Summers 1999). The statistical techniques of discriminant analysis were applied in order to choose a subset of parameters that best discriminated between the degraded and undegraded sites. Appropriate coefficients (or weights) were computed by canonical discriminant analysis. The benthic index is a linear function of the weighted parameters that represent the subset that best discriminates between degraded and undegraded sites. The benthic index was scaled to range from 0 to 10 and threshold values were determined empirically from the data. The final result is an index that, when calculated to be below the threshold value for a given site, indicates that the benthic community at that site is similar to benthic communities found at known degraded sites. The summary of the development of EMAP-E's benthic index given here is brief; more detail on the development, validation, and application of the benthic index may be found elsewhere (Engle *et al.* 1994, Engle and Summers 1998, Engle and Summers 1999).

Guideline 3: Data Collection Methods

Methods for collecting all indicator measurements should be described. Standard, well-documented methods are preferred. Novel methods should be defended with evidence of effective performance and, if applicable, with comparisons to standard methods. If multiple methods are necessary to accommodate diverse circumstances at different sites, the effects on data comparability across sites must be addressed. Expected sources of error should be evaluated.

Methods should be compatible with the monitoring design of the program for which the indicator is intended. Plot design and measurements should be appropriate for the spatial scale of analysis. Needs for specialized equipment and expertise should be identified.

Sampling activities for indicator measurements should not significantly disturb a site. Evidence should be provided to ensure that measurements made during a single visit do not affect the same measurement at subsequent visits or, in the case of integrated sampling regimes, simultaneous measurements at the site. Also, sampling should not create an adverse impact on protected species, species of special concern, or protected habitats.

Field and laboratory methods for collecting and processing benthic macroinvertebrate samples are thoroughly documented in Heitmuller and Valente (1991), Macauley (1991), and U.S. EPA (1995).

Field Sampling Methods

A 1/25 m^2, stainless steel, Young-modified Van Veen Grab sampler was used to collect 3 replicate sediment grabs for enumeration of benthic invertebrate community composition. This grab sampled an area of 413 cm^2 with a maximum depth of penetration in the sediment of 10 cm. Each acceptable benthic grab sample was rinsed into a plastic dishpan for transport to the sieving station for immediate, aboard processing. The sediment from an individual grab was sieved through a 500 µm sieve to wash away sediments and leave organisms, detritus, sand and shell particles larger than 500 µm. The contents on the sieve were rinsed with site water, into 500-ml wide-mouth polypropylene jar(s). The contents of each jar were preserved by the addition of 100 ml of formalin:seawater (50:50) containing Rose Bengal vital stain to yield a final formalin concentration of 10% by volume.

Expected sources of error in the field sampling methods were reduced by mandatory training of all field personnel in proper collection methods including determination of acceptable grabs, sieving techniques, and preservation of the sample. However, slight measurement error could occur if the volume of sediment in a grab was not consistent among grabs or if there was not sufficient water and formalin in the sample jars to fill the jar (in order to prevent agitation of fixed organisms). Human error could occur also in the transfer of samples from grab to sieve to sampling jar. These sources of measurement error were minimized by thorough training of all field personnel and random quality control audits to ensure that proper sampling techniques were being employed.

The field sampling methods employed by EMAP-E are not expected to cause significant disturbance to a site. The grab samples an area of 413 cm^2 with a maximum penetration depth of 10 cm; this results in a maximum sample volume of 4 liters. After 3 grabs were taken, whether successful or not, the anchor line was let out to move the boat 5 m downstream to ensure that the exact 413 cm^2 location was not sampled repetitively. Because EMAP-E samples each station at a single point in time, the adverse effects of disturbance at any site are minimal. There are no species of estuarine benthic macroinvertebrates currently listed as endangered, threatened, or protected in the Gulf of Mexico. Although most sampling occurred in open-bay bottom, occasionally a site was located in a protected submerged aquatic vegetation (SAV) habitat, in which case, special care was taken not to unduly disturb SAV.

Laboratory Processing Methods

The samples were shipped directly from the field to the laboratory where they were again washed through 500 µm mesh sieves. Benthic fauna were sorted from the sediments, identified to lowest practical taxa, and enumerated. Only benthic macroinvertebrates were identified. Meiofauna and taxonomic groups having only planktonic forms were excluded from the identification process. At least one qualified benthic taxonomist was required in order to provide authority on identification of taxa.

Expected sources of error in laboratory processing included errors in handling of samples, inefficient sorting and inaccurate identifications. These sources of error were minimized by rigorous quality control measures that included random resorts and recounts and by having qualified and trained personnel performing the sorting and identification of taxa.

Data Manipulation Methods

The benthic macrofaunal count data were sent to a data manager in dBase® format. The data manager translated this data into SAS® format. The data was then checked for transcription errors, and inconsistencies in taxonomic coding as well as new taxonomic codes. Detailed methods for the development and application of the benthic index are found elsewhere (Engle *et al.* 1994, Engle and Summers 1998, Engle and Summers 1999). The benthic parameters that make up the components of the benthic index were calculated for each sampling station (Table 3-1). The benthic index was calculated by combining the components in a linear fashion as illustrated in Table 3-2.

Table 3-1. Methods used to calculate components of the benthic index from raw benthic data.

Component	Calculation Method
Proportion of Expected Shannon-Wiener Diversity Index	Shannon-Wiener Diversity Index: $H' = -\sum p_i \log_2 p_i$ where p_i = proportion of total abundance represented by species *i*. Proportion of Expected Shannon-Wiener Diversity Index = $H' \equiv [2.618426 - 0.044795X + 0.007278X^2 - 0.000119X^3]$ where X = bottom salinity (ppt)
Mean Abundance of Tubificids	Calculate sum of total abundance of members of Family:Tubificidae and divide by number of grabs (3)
Percent Capitellids	Calculate sum of abundance of members of Family:Capitellidae in each grab and divide by the total abundance of all fauna in that grab. Calculate average proportion in 3 grabs and multiply by 100.
Percent Bivalves	Calculate sum of abundance of members of Class:Bivalvia in each grab and divide by the total abundance of all fauna in that grab. Calculate average proportion in 3 grabs and multiply by 100.
Percent Amphipods	Calculate sum of abundance of members of Order:Amphipoda in each grab and divide by the total abundance of all fauna in that grab. Calculate average proportion in 3 grabs and multiply by 100.

Table 3-2. Methods used to calculate the benthic index from the components listed in Table 1.

- Step 1: Log transform abundances and arc-sine transform proportions.

- Step 2: Standardize variables to mean = 0 and standard deviation = 1 by applying the following formula to all of the data:

$$x_i' = \frac{S * (x_i - 0) + M}{s_x}$$

where

x_i' = new standardized value
S = desired standard deviation (1)
M = desired mean (0)
x_i = observation's original value
0 = variable's mean
s_x = variable's standard deviation

- Step 3: Calculate the discriminant score of the benthic index as follows using the standardized variables from Step 2.

Discriminant Score = (1.5710 x Proportion of expected diversity) +
(-1.0335 x Mean Abundance of tubificids) +
(-0.5607 x Percent capitellids) +
(-0.4470 x Percent bivalves) +
(0.5023 x Percent amphipods).

- Step 4: Calculate the benthic index by normalizing the discriminant scores from Step 3 using the following formula which includes the minimum (-3.21) and range (7.50) of discriminant scores from the original test data used to develop the benthic index:
Benthic Index = ((Discriminant Score - (-3.21)) /7.50) * 10

Guideline 4: Logistics
The logistical requirements of an indicator can be costly and time-consuming. These requirements must be evaluated to ensure the practicality of indicator implementation, and to plan for personnel, equipment, training, and other needs. A logistics plan should be prepared that identifies requirements, as appropriate, for field personnel and vehicles, training, travel, sampling instruments, sample transport, analytical equipment, and laboratory facilities and personnel. The length of time required to collect, analyze and report the data should be estimated and compared with the needs of the program.

Monitoring programs that utilize sediment sampling for the collection of benthic macroinvertebrates vary in their logistic requirements depending on the spatial extent and temporal duration of the monitoring design. All field operations conducted by EMAP-E were planned and implemented according to an approved logistics plan that was prepared following guidelines established for EMAP (Baker and Merritt 1990). Elements of the logistics plan address major areas of project implementation, including project management, site access and scheduling, safety and waste disposal, procurement and inventory control, training and data collection, and the assessment of the operation upon completion. EMAP-E in the Lousianian Province was tasked with sampling ~150 stations that ranged hundreds of miles from Texas to Florida. The time frame was ~2 months during the summer. Because the success of EMAP-E depended on standardized sampling and processing, all boat crews underwent rigorous training that covered boat operation, collection methods, and proper QA procedures. A field operations manual was prepared each year to give detailed instructions to the field teams on safety, operation of equipment, handling of samples, and quality assurance (Macauley 1991). A logistics plan was prepared each year that gave a day-to-day account of locations to be sampled, personnel assignments, and suggested hotels, boat ramps, and other necessary resources. In addition, a quality assurance project plan was prepared to identify quality control guidelines for collection, handling, and shipping of field samples as well as laboratory analytical methods (Heitmuller and Valente 1991). Table 3-3 lists the logistical requirements to carry out sampling of benthos by the EMAP-E field teams; however, the magnitude of equipment required by EMAP-E would not necessarily be appropriate for small-scale, localized monitoring programs.

EMAP-E set a goal of producing a statistical summary approximately 9 months after sampling concluded. Collection and field processing of benthic samples was completed aboard the boat within a short time frame (*i.e.*, 1-2 hours). In the laboratory, processing of the samples was more tedious and, on average, it took 5-10 man-hours to process a sample, from the initial bench sieving to transfer of raw data from handwritten sheets to an electronic file. The original development of the index from receiving the electronic data to publishing a manuscript took on the order of several years. The application of the index, now that it has been finalized, is relatively straightforward; the length of time from receipt of the data to calculating the index is on the order of a few days.

Guideline 5: Information Management
Management of information generated by an indicator, particularly in a long-term monitoring program, can become a substantial issue. Requirements should be identified for data processing, analysis, storage, and retrieval, and data documentation standards should be developed. Identified systems and standards must be compatible with those of the program for which the indicator is intended and should meet the interpretive needs of the program. Compatibility with other systems should also be considered, such as the internet, established federal standards, geographic information systems, and systems maintained by intended secondary data users.

Table 3-3. Logistical requirements for sampling and processing of benthic macroinvertebrates by EMAP-E in the Louisianian Province.

Field Personnel	3 Teams consisting of 2 crews with 5 members each (1 crew chief, 2 boat crew members, and 2 shore crew members)
Vehicles	1-ton, 4WD, dual rear wheel pickup truck with heavy duty, dual axle trailer (with brakes, winch, spare tire and rollers) 25-foot SeaArk work boat equipped with 7.5 L gas engine fitted with Bravo II outdrive, an "A" frame boom assembly and hydraulic winch. On-board electronics included Loran C, 2 VHF radios, radar, compass, and depth finder. Mobile laboratory - 15-foot truck equipped with VHF radio, GRiD laptop computer, shelves, and work bench. Full-size panel van for transporting crew members.
Training	A 2-week training course is mandatory for all crew members. Crew members must show proficiency in towing and launching the boat, using navigation equipment, locating stations, entering and retrieving data from computers, using all sampling gear, first aid, and safety.
Travel	The two crews comprising a team worked alternating 6-day schedules. Extensive travel from the Field Operations Center to the staging area was required of all crews (as much as 1000 miles by road, trailering the boat). Site reconnaissance was performed prior to initiation of field activities in order to determine locations of boat ramps and hotels and to identify any stations unsuitable for sampling.
Sampling Gear	Stainless steel Young-modified Van Veen grab sampler (self-leveling with a hinged top)
Data Transport	Samples were transferred from 0.5 mm sieve to wide mouth Nalgene bottles and preserved with 10% buffered formalin containing Rose Bengal stain. Bottles are labeled with bar code, packed individually in ziploc plastic bags and placed into shipping container. Samples were shipped via Federal Express to the benthic sample processing laboratory.
Laboratory Facilities	State-of-the-art benthic processing laboratory equipped with compound and dissecting microscopes, magnifying lights, computers for data entry, complete specimen voucher collection.
Laboratory personnel	3 benthic taxonomists, 3-5 student sorters, 1 Ph.D. level benthic ecologist /supervisor.

Information management was thoroughly addressed in EMAP-E and one of its major goals was to disseminate the data to the public in a timely manner. This goal requires that the data be collected with adequate sample tracking methods and that it is stored in a standardized format that is made easily accessible to the public. All samples that were collected by EMAP-E were tracked via a bar code system from the field to the laboratory. At the laboratory, benthic taxa identifications and counts were handwritten onto data sheets that were then transcribed into an electronic dBase® file. This file was sent via e-mail to the EMAP-E data manager who translated the data into SAS® format. All data manipulations and calculations of the benthic index were accomplished using SAS® on an Intel® PC. EMAP-E data is currently stored in three venues: 1) as SAS® data sets on a Windows® NT server at the Louisianian Province office, 2) as Oracle 7™ DBMS files on a Windows® NT server at the Louisianian Province office, and 3) as ASCII text downloadable files on the EMAP web page (http://www.epa.gov/emap) which is housed on a server at the centralized EMAP Information Management (EMAP-IM) office. EMAP-IM produced an information management plan as an evolving document to outline EMAP's strategy for maintaining, archiving, and distributing all of the data collected by EMAP (Hale *et al.* 1998). In addition to the data sets, metadata files are cataloged that describe methods, contacts, quality assurance, and other information pertinent to the data.

The complex information management requirements of EMAP were necessary to ensure that the large amounts of data from this national monitoring program were consistently documented, standardized, and made available to end-users in a timely and efficient manner. These might not be necessary for smaller monitoring programs. At a minimum, the benthic taxa identifications and counts could be entered electronically into a spreadsheet or database format and all calculations for the benthic index could be accomplished there. In this case the hardware and software required would be a high-end PC with a package such as Excel, Lotus, or dBase installed.

Guideline 6: Quality Assurance
For accurate interpretation of indicator results, it is necessary to understand their degree of validity. A quality assurance plan should outline the steps in collection and computation of data, and should identify the data quality objectives for each step. It is important that means and methods to audit the quality of each step are incorporated into the monitoring design. Standards of quality assurance for an indicator must meet those of the targeted monitoring program.

EMAP-E emphasized the collection of data via standardized methods. This ensured the comparability of data collected by different field teams across large geographic areas. Rigorous quality control (QC) was necessary to achieve this goal. All field crew personnel were trained prior to the sampling season in, among other things, the proper techniques for grab sampling, sieving, and preservation of benthic samples. Both the field crews and the laboratory personnel were provided with manuals that outlined the correct techniques for handling the samples (see Macauley 1991 and U.S. EPA 1995). Random QC audits were performed by the Quality Assurance Officer both in the field and at the laboratory. Table 3-4 lists the various aspects of quality control for field and laboratory operations as well as information management for EMAP-E Louisianian Province benthic data (Heitmuller and Valente 1991).

Guideline 7: Monetary Costs
Cost is often the limiting factor in considering to implement an indicator. Estimates of all implementation costs should be evaluated. Cost evaluation should incorporate economy of scale, since cost per indicator or cost per sample may be considerably reduced when data are collected for multiple indicators at a given site. Costs of a pilot study or any other indicator development needs should be included if appropriate.

It is difficult to separate the cost of implementing this particular indicator from the overall cost of implementing EMAP. The cost of personnel, vehicles, and travel are spread across all indicators measured by the program. Even the cost of the equipment is spread across several indicators (*i.e.*, benthos, sediment characterization, sediment toxicity, and sediment chemistry) as the same gear is used to collect all of these samples. The Young-modified Van Veen grab cost $1250 to purchase initially. The sieve boxes, forceps, Nalgene bottles, labels and other miscellaneous equipment for one team cost $350 for one year. The entire process of collecting, sieving, and preserving benthic samples took an average of 1 hour per station. Given that each team sampled an average of 50 stations per year and assuming that the grab lasts for four years, the average equipment cost per station would be only $13.

Table 3-4. Quality Control procedures for field sampling, laboratory processing, and data analysis of EMAP-E Louisianian Province benthic data.

Field Operations

Sample Collection

- Sediment should not extrude from the upper face of sampler.
- Overlying water should be present.
- Sediment surface should be relatively flat.
- Entire surface of sample should be included in sampler.
- Grab must have penetrated sediment to a minimum depth of 7 cm.
- If these QC guidelines were not met, the sample was rejected.

Sample Processing

- Gentle rinsing of sample through sieve - no forceful jets of water.
- Preservation of sample with 10% buffered formalin containing Rose Bengal stain.

Laboratory Operations

Sample Storage

- Samples should be stored between 5°C and 30°C to avoid freezing or evaporation.
- After sorting, organisms should be stored in vials with 70% ethanol.
- Minimize exposure of samples to direct sunlight.

Sorting (i.e., separating organisms from sediment and debris)

- 10% of all samples are resorted independently by a second, experienced sorter.
- Re-sorts are randomly chosen (1 out of 10) on a regular basis
- Sorting efficiency is calculated as:

$$\frac{\text{\# of organisms originally sorted}}{\text{\# of organisms originally sorted + additional \# found in re-sort}} \times 100$$

- Actions for unsatisfactory sorting efficiencies:
  90-95% - Technician should be retrained.
  < 90% - All samples in batch must be resorted and any organisms found in re-sort will be added to the original data sheet.

Species Identification and Enumeration

- 10% of all samples are checked for accuracy by a senior taxonomist.
- Re-identification samples are randomly chosen (1 out of 10) on a regular basis
- Accuracy will be calculated as:

$$\frac{\text{Total \# of organisms in QC recount -Total \# of errors}}{\text{Total \# of organisms in QC recount}} \times 100$$

where errors include:
Counting error (e.g., counting 11 of a given species instead of 10)
Identification error (e.g., misidentifying species X as species Y)
Unrecorded taxa errors (e.g., not identifying species X when it is present)

- Actions for unsatisfactory level of taxonomic accuracy:

  90-95% - Original technician advised, species identifications reviewed, and any changes to species identifications recorded on original data sheet.
  < 90% - Same as for 90-95% but numerical counts should also be corrected on original data sheet.

- Maintain voucher specimen collection and permanent undegraded collection.

The bulk of the expense to implement a benthic monitoring program comes from the laboratory costs. The benthic laboratory for this study charged $300 per sample for sorting and taxonomic identification. Three replicate samples were collected at each station bringing the laboratory cost to $900 per station. This brings the total cost of implementing the benthic index to $913 per station, excluding travel, lodging and personnel costs. These costs were well within the budget allocated to this program.

Phase 3: Response Variability

Once the indicator has been determined to be pertinent to the assessment question, ecologically relevant, and feasible to implement, then the next phase is to evaluate the expected variability in the response of the indicator. Variability can arise from many sources (*i.e.,* human error in field or laboratory processing of samples, temporal variability, and spatial variability). EMAP-E has incorporated, as part of its design, mechanisms to address these sources of variability in any indicator. It is important to evaluate the variability in an indicator in the context of the specific assessment question. The evaluation of response variability is very different for an indicator that is designed to measure the current condition at a particular site versus an indicator like EMAP-E's benthic index. This benthic index was designed to measure the proportion of estuarine area in the Louisianian Province that had degraded benthic communities. In this example, the assessment question was aimed at estimating the spatial extent of degraded benthic communities across a large geographical area (~25,000 km^2 of estuarine area in the Louisianian Province).

The data used in this evaluation came entirely from EMAP-E's efforts in the Louisianian Province estuaries (northern Gulf of Mexico, from Anclote Key, Florida to Rio Grande, Texas) from 1991 to 1994. Figure 3-3 shows the distribution of sampling sites for a single year, 1991. The sample design, methods, results, and statistical evaluations are documented in Summers *et al.* (1991), Summers *et al.* (1992), Summers *et al.* (1993), Macauley *et al.* (1994), and Macauley *et al.* (1996).

Guideline 8: Estimation of Measurement Error
The process of collecting, transporting, and analyzing ecological data generates errors that can obscure the discriminatory ability of an indicator. Variability introduced by human and instrument performance must be estimated and reported for all indicator measurements. Variability among field crews should also be estimated, if appropriate. If standard methods and equipment are employed, information on measurement error may be available in the literature. Regardless, this information should be derived or validated in dedicated testing or a pilot study.

While the parsing of overall variance into specific components (*e.g.*, measurement error) is essential to the estimation of trends, our program was initially more concerned with the estimation of status. We did not evaluate measurement error specifically; to do so would require a redesign of our program. We did, however, determine the most likely sources of measurement error and sought to minimize this error through rigorous training and quality control. Measurement errors can be introduced into the benthic data from three primary sources: collection of the sample, handling and preservation of the sample, and activities in the laboratory. In the field, variability in the sample would be associated with the volume of the grab, incorporation of water in the sample, and human error associated with field sieving and preservation of the sample. This variability

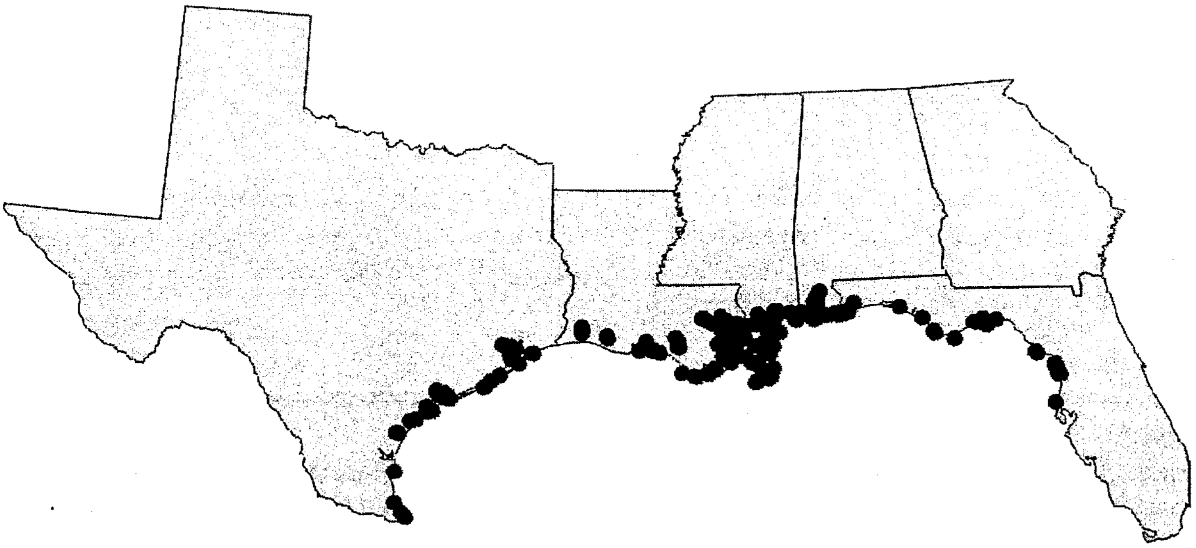


Figure 3-3. Map of EMAP-E sampling sites for 1991 in the Louisianian province.

Table 3-5. Results of the quality control measures employed by the benthic laboratory in the sorting, identification, and enumeration of benthic macroinvertebrate samples from EMAP-E Louisianian Province, 1991 to 1994.

QC Measure	1991	1992	1993	1994
Sorting QC				
Total # jars sorted	558	nd	527	544
Total # jars QC'd	64	nd	53	56
# jars passed	57	nd	49	55
Sorting success rate	89%	nd	92%	98%
Corrective action	yes	nd	yes	no
Taxonomy and Enumeration QC				
Total # vials ID'd	~800	nd	738	895
Total # vials QC'd	83	nd	74	92
# QC vials with >10% error	3	nd	0	3
Taxonomy success rate	96%	nd	100%	97%
Corrective action	no	nd	no	no

nd = no data available

was minimized by rigorous training of the field crew prior to initiation of the sampling season and QC audits of the field crew that were performed throughout the sampling season. Although the magnitude of variability in the indicator that is associated with these sources of measurement error were not quantified, field crews consistently scored >95% efficiency in all field activities. We believe that measurement errors were minimal due to the standardized methods employed and the thorough training and QC requirements imposed on the personnel.

As detailed in *Guideline 6: Quality Assurance*, quality control measures were employed in the laboratory to minimize the variability in the benthic data resulting from potential human processing error. In addition to the QC requirements of 10% resorts and reidentifications, QC audits were performed on the laboratory as well. The results of the QC for the benthic laboratory are listed in Table 3-5.

Guideline 9: Temporal Variability - Within the Field Season
It is unlikely in a monitoring program that data can be collected simultaneously from a large number of sites. Instead, sampling may require several days, weeks, or months to complete, even though the data are ultimately to be consolidated into a single reporting period. Thus, within-field season variability should be estimated and evaluated. For some monitoring programs, indicators are applied only within a particular season, time of day, or other window of opportunity when their signals are determined to be strong, stable, and reliable, or when stressor influences are expected to be greatest. This optimal time frame, or index period, reduces temporal variability considered irrelevant to program objectives. The use of an index period should be defended and the variability within the index period should be estimated and evaluated.

EMAP-E chose to implement sampling during the summer (July to September) because this was the period during which the stressors of concern (*i.e.*, contaminants and DO) would most severely impact the biota. During the summer index period, benthic organisms are most active, temperatures are highest, hypoxia occurs more frequently, and predation is at its peak. In estuaries especially, the added pressures attributed to human populations (*e.g.*, recreational boating, fishing, increased water usage and municipal effluent, nonpoint source runoff of nutrients from agricultural lands) are at their highest during the summer. It is during this index period that we are most likely to detect impacts of stressors on benthic communities.

The benthic index was computed for all sites sampled by EMAP-E in the Louisianian Province from 1991 to 1994. Because the index was developed from a subset of sites sampled in 1991 and 1992, validation of the benthic index was accomplished by using an independent set of data from two subsequent years, 1993 and 1994, as well as data from special study sites representing between-year and within-year replicates. Validation of the benthic index consisted of three steps: assessment of the correct classification by the index of an independent set of degraded and undegraded sites, comparison of the cumulative distribution function of the index among four years, and correct classification of replicate sites by the index.

Within each year (excluding 1991), 13 estuaries were visited more than once in order to evaluate spatial and temporal replication. These sites were used to validate the consistency of classification of the benthic index. The same classification by the benthic index should be given to a single site on replicate visits within a sampling season. The distribution of benthic index values between the first and second visits to a site within the same sampling period was compared (Fig. 3-4). The shaded areas indicate the marginal zone between the threshold values of 3.0 for degraded sites and 5.0 for undegraded sites. Ideally, all of the points should fall in quadrants 2 and 4 where sites were classified as degraded on both visits or as undegraded on both visits. However, although several points fall within the lightly shaded area, indicating that the site classification changed from degraded or undegraded to marginal (or *vice versa*) from the first visit to the second visit, no sites fell within quadrants 1 and 3, indicating that no sites were inversely classified as degraded or undegraded. Correlation between the benthic index from the temporal replicates was significant ($p < 0.05$; $r = 0.83$). The Kappa statistic (κ) was used to test the degree of agreement in classification by the benthic index between site visits. The null hypothesis of no agreement (H_o: $\kappa = 0$) was rejected at $\alpha = 0.5$. The Kappa statistic ($\kappa = .644$) indicated moderate agreement in the classification between visits. This validation of the benthic index was determined to be successful in showing minimal within-year temporal variability.

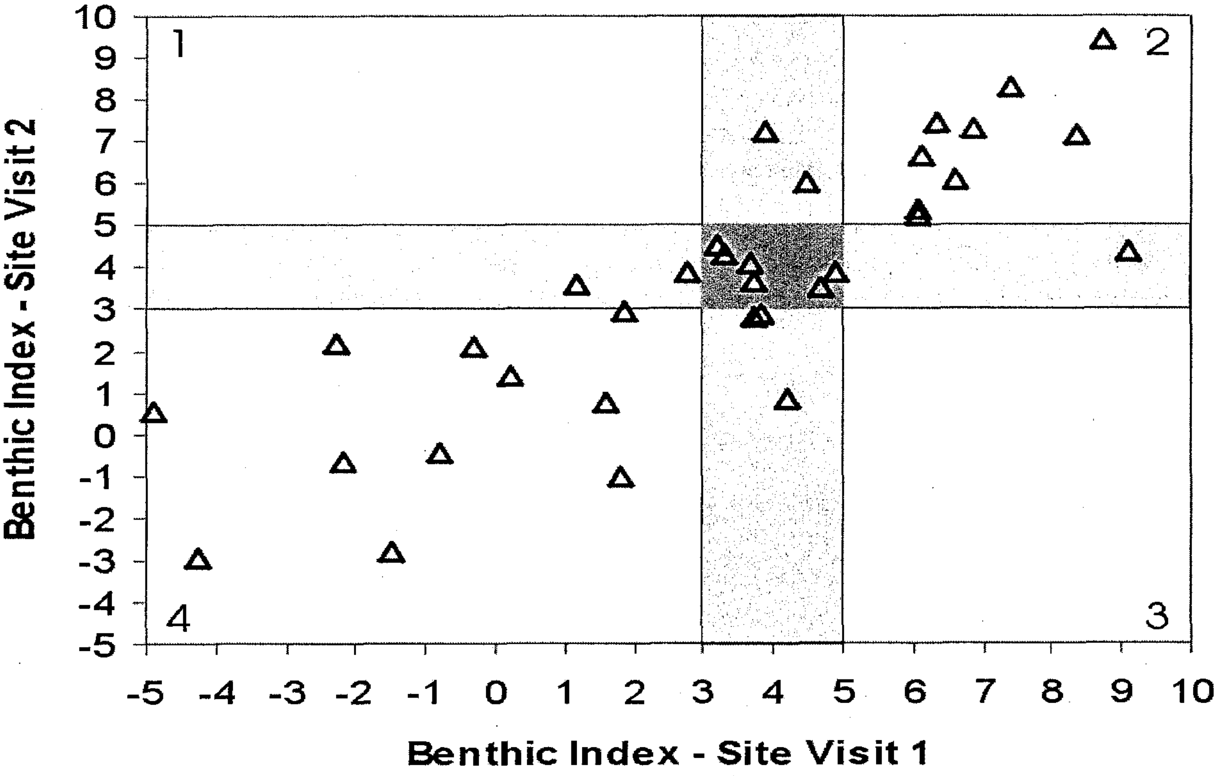


Figure 3-4. Comparison of benthic index values from replicate visits within a sampling season to a site. Quadrant 2 indicates sites classified as undegraded for both visits; quadrant 4 indicates sites classified as degraded for both visits; quadrants 1 and 3 indicate sites that were classified differently for both visits. Sites that fall within the gray shaded area (except for those sites in the center of the cross) changed classification from degraded or undegraded to marginal (or vice versa) from visit 1 to visit 2.

Guideline 10: Temporal Variability - Across Years
Indicator responses may change over time, even when ecological condition remains relatively stable. Observed changes in this case may be attributable to weather, succession, population cycles or other natural inter-annual variations. Estimates of variability across years should be examined to ensure that the indicator reflects true trends in ecological condition for characteristics that are relevant to the assessment question. To determine inter-annual stability of an indicator, monitoring must proceed for several years at sites known to have remained in the same ecological condition.

A subset of 13 sites from 1991 was sampled every year to provide an estimate of between-year variation. An analysis of variance was performed on the data from the between-year temporal replicates to test for a year effect on the benthic index and a pairwise T-test was used to detect significant differences in benthic index values paired by station between any two years. For both tests, the null hypothesis of "no significant difference" was not rejected (p > 0.05). The benthic index values among years at the thirteen stations are compared categorically in Figure 3-5. We defined a change in classification as a change from degraded to undegraded (or *vice versa*). Benthic index values between 3 and 5 represented moderate or marginal conditions. A change in classification from degraded to moderate or undegraded to moderate was not viewed as a misclassification by the index. Although the range of values for any given station is sometimes large, the classification of a station does not change from degraded to undegraded (or *vice versa*) except in the case of stations 1 and 10. The classification of a station does change from degraded or undegraded to moderate for many stations, however. This validation exercise showed that the benthic index maintained relative inter-annual stability in addition to minimal within-year variability.

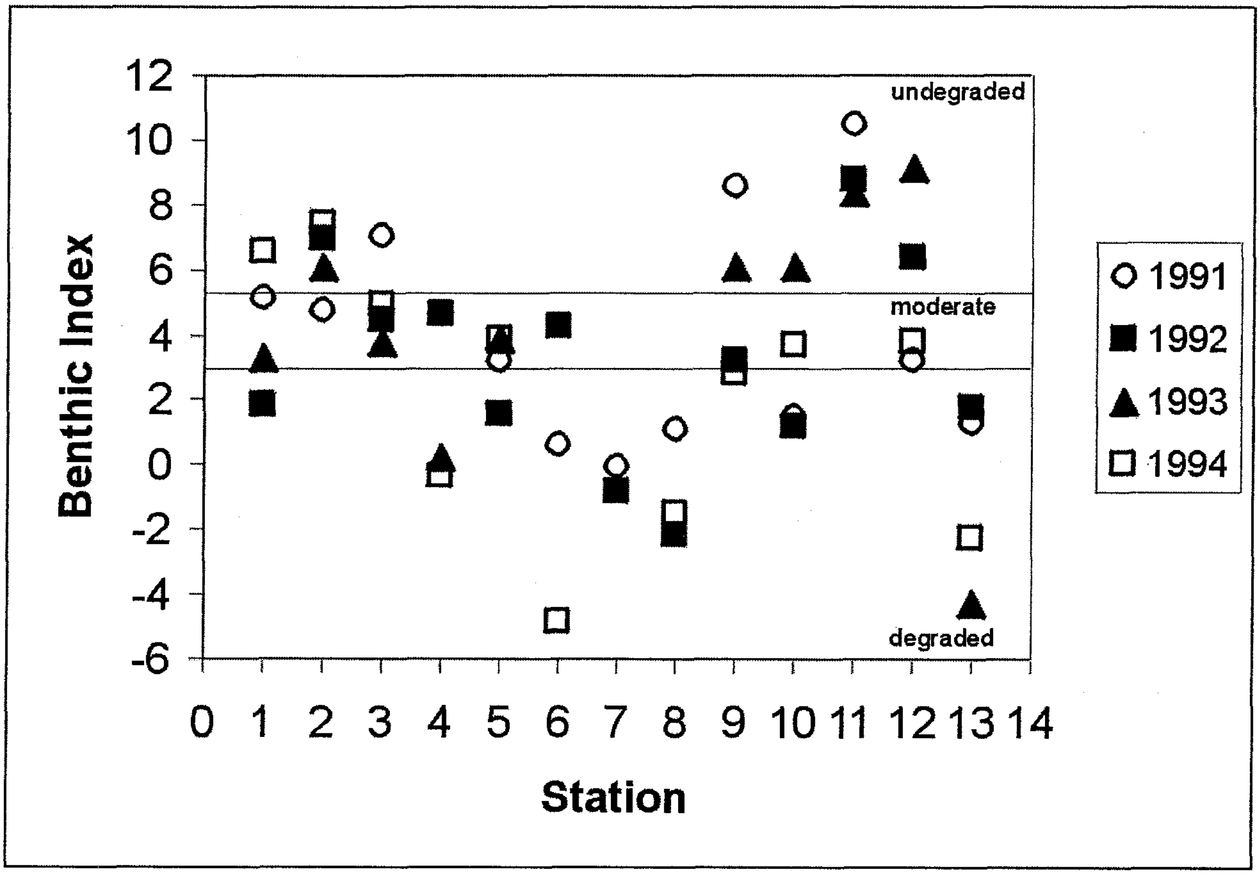


Figure 3-5. Comparison of benthic index values at stations that were sampled each year. Horizontal lines at benthic index values of 3 and 5 indicate the boundaries of degraded, moderate and undegraded benthic conditions.

Guideline 11: Spatial Variability
Indicator responses to various environmental conditions must be consistent across the monitoring region if that region is treated as a single reporting unit. Locations within the reporting unit that are known to be in similar ecological condition should exhibit similar indicator results. If spatial variability occurs due to regional differences in physiography or habitat, it may be necessary to normalize the indicator across the region, or to divide the reporting area into more homogeneous units.

Monitoring programs that cover large geographic areas must contend with the inherent spatial variability in the data. In the Louisianian Province, which spans estuaries from Texas to Florida, EMAP-E encountered the full range of expected benthic habitat types (from sand to mud, tidal freshwater to marine, and impacted to pristine) and identified more than 1000 different benthic invertebrate species. The 46 test sites that were used to develop the benthic index were chosen, therefore, not only to represent extreme reference and degraded environmental conditions, but also to cover the expected range of salinity, sediment types, and biogeographical divisions found in the estuaries of the northern Gulf of Mexico (Engle *et al.* 1994, Engle and Summers 1999). The sites ranged in salinity regimes from tidal-freshwater (0 ppt) to marine (>35 ppt) and most sites were located in muddy (>80% silt-clay) sediment (Table 3-6). The location of the majority of sites in Louisiana is simply an artifact of the EMAP probability-based sample design (Louisiana has proportionately more estuarine area than the other four gulf states). In this way, we sought to minimize the spatial variability in the benthic index.

Table 3- 6. Distribution of the number of degraded and undegraded test sites among categories of salinity, sediment types, and states.

Salinity	Number of Degraded Sites	Number of Undegraded Sites
Fresh (0 ppt)	8	2
Brackish (>0-5ppt)	2	3
Mesohaline (>5-18 ppt)	4	5
Polyhaline (>18-35 ppt)	7	11
Marine (>35 ppt)	1	3
Sediment Type		
Mud (>80% Silt-Clay)	17	10
Mud/Sand (20-80% Silt-Clay)	5	10
Sand (<20% Silt-Clay)	0	4
State		
Florida	4	2
Alabama	3	3
Mississippi	2	2
Louisiana	12	10
Texas	1	7

Guideline 12: Discriminatory Ability
The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined. This analysis should incorporate all error components relevant to the program objectives, and separate extraneous variability to reveal the true environmental signal in the indicator data.

Sites from 1993 and 1994 were classified as degraded or undegraded based on our *a priori* criteria for dissolved oxygen, sediment chemistry, and sediment toxicity that were used to choose test sites in the development of the index. Of the 310 sites sampled in 1993 and 1994, only 195 could be classified as either degraded or undegraded and these were used in the first validation step. In a Monte Carlo exercise, we randomly chose 50 subsets of the 195 sites where each subset consisted of 50 degraded and 50 undegraded sites. Correct classification occurred when the benthic index was either $\leq$ 3 at degraded sites or $\geq$ 5 at undegraded sites. Misclassification occurred when the benthic index was $\leq$ 3 at undegraded sites (false negative) or $\geq$ 5 at degraded sites (false positive). Using the 50 trials, we determined the percent of sites that were correctly classified as degraded and undegraded by the benthic index. The benthic index correctly classified 66-82% of degraded sites ($\bar{x}$ = 74%; SE = 0.5) and 70-84% of undegraded sites ($\bar{x}$ = 77%; SE = 0.4). The high degree of variability in the benthic communities in the Gulf of Mexico region influenced the classification success. Although we attempted to minimize this variability during the development phase, we may have sacrificed a level of precision in favor of a generalized index that is applicable across a wide geographic area with an inherently large spatial variation. We also investigated the kappa coefficient (κ) to measure the degree of agreement (Stokes *et al.* 1995) between classification of a site by the benthic index versus classification by sediment contaminants, toxicity, and dissolved oxygen. The average kappa coefficient for the 50 trials was 0.509 where $\kappa \geq 0.4$ indicates moderate agreement and the null hypothesis that there was no agreement (H_0: $\kappa = 0$) was rejected at the $\alpha = 0.05$ level of significance.

An important consideration for the benthic index was that it not be significantly correlated with any natural habitat factors like salinity or sediment type. This was addressed during the development of the index by adjusting any benthic parameters that were correlated with salinity, or sediment type. One of the components of the benthic index, proportion of expected diversity, represents a salinity-adjusted variable because diversity is highly correlated with salinity in estuarine waters. Figures 3-6 and 3-7 show the relationship between the benthic index and salinity and percent silt-clay content of sediments. The benthic index was still significantly correlated with salinity and percent silt-clay but the R^2 for both of these correlations was <15%. We determined that these relationships were insignificant from an ecological perspective with statistical significance primarily driven by the large number of samples (n = 338).

Anthropogenic impacts may be correlated with salinity and silt-clay as well; therefore, residual correlations between the benthic index and salinity or silt-clay may not indicate a lack of discriminatory power in the index. This is important because the benthic index was designed to be an indicator of environmental condition that is representative of the degree of sediment contamination and hypoxia experienced at a site, regardless of the inherent salinity and sediment characteristics.

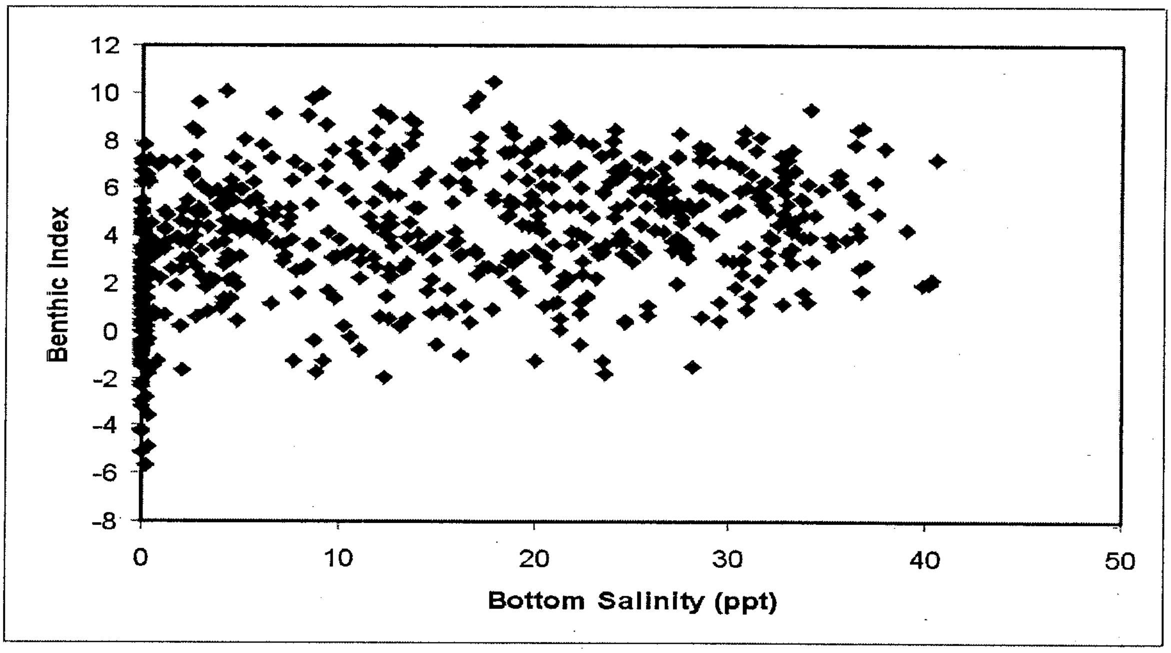


Figure 3-6. Benthic index versus bottom salinity for all sites from EMAP-E Louisianian Province, 1991 to 1994.

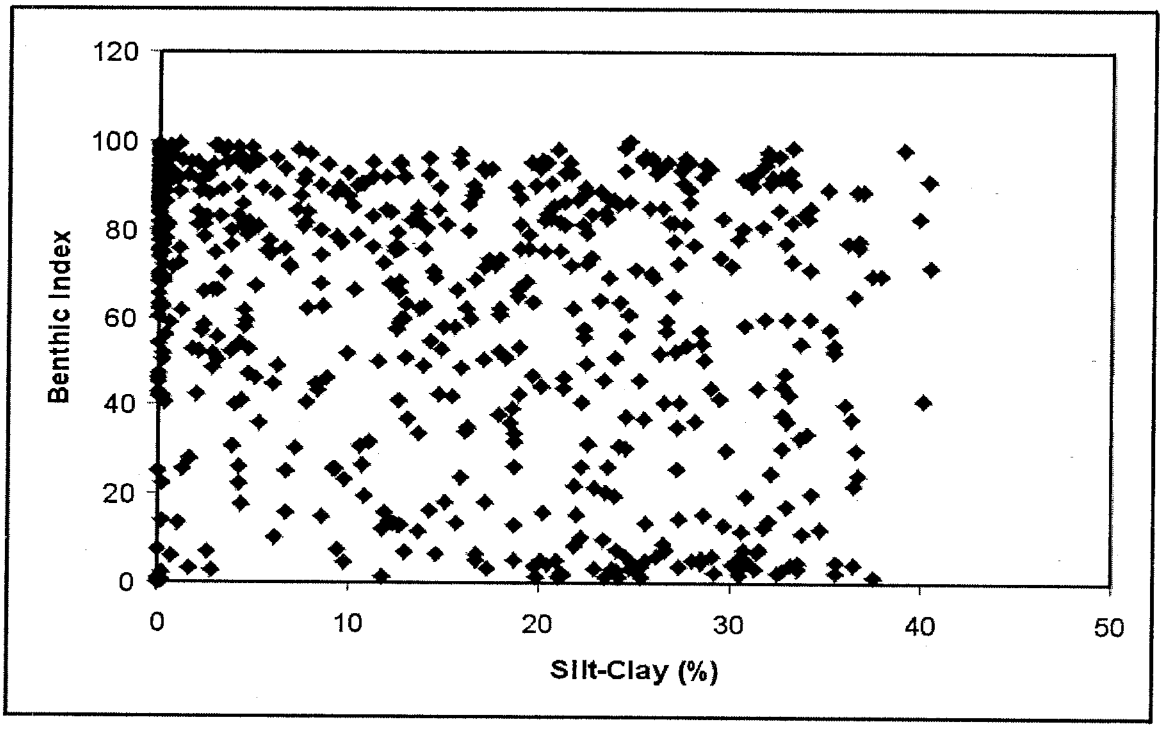


Figure 3-7. Benthic index versus silt-clay content of sediments for all sites from EMAP-E Louisianian Province, 1991-1994.

Phase 4: Interpretation and Utility

Guideline 13: Data Quality Objectives
The discriminatory ability of the indicator should be evaluated against program data quality objectives and constraints. It should be demonstrated how sample size, monitoring duration, and other variables affect the precision and confidence levels of reported results, and how these variables may be optimized to attain stated program goals. For example, a program may require that an indicator be able to detect a twenty percent change in some aspect of ecological condition over a ten-year period, with ninety-five percent confidence. With magnitude, duration, and confidence level constrained, sample size and extraneous variability must be optimized in order to meet the program's data quality objectives. Statistical power curves are recommended to explore the effects of different optimization strategies on indicator performance.

Traditional power analyses are employed in hypothesis testing to evaluate the probability of failing to reject the null hypothesis when the alternative hypothesis is true. This is called a "Type II error" and has been described as "not seeing enough in the data" by Anderson (1966 as cited in Keppel and Saufley 1980).

Power is then the probability that a correct decision is made (*i.e.*, the null hypothesis is rejected when the alternative hypothesis is true). Power analyses are recommended as part of the experimental design process to determine the number of samples needed in order to make a correct decision with a given level of power. Afterward, power analyses may be used to determine the probability that a correct decision was made given the number of samples and the sample variance.

Power analyses were included in the design of EMAP-E but were different from traditional analyses in that we were evaluating the power to detect a trend rather than the power to reject a null hypothesis. EMAP-E set a performance goal of detecting a 2% per year change in the province-wide proportion of area that exceeds a pre-specified indicator threshold value over a period of 12 years with a probability exceeding 0.8 (Larsen *et al.* 1995, Heimbuch *et al.* [in review]). In order to test whether EMAP-E is capable of meeting this target, power analyses were performed to construct scenarios for detecting a 1% to 3% change over 12 years.

The power to detect a 2% trend over 12 years is calculated by first estimating the variance. The overall variance consists of two components: a spatial component and a component that is dependent on the number of years over which you wish to estimate a trend. Spatial variance was estimated for the four years during which data were collected. We computed power curves for 1% to 3% change per year ($\beta = 0.01, 0.015, 0.02, 0.025, 0.03$) for years 6 to 12 at $\alpha = 0.10$ according to the methods presented in Heimbuch et al. (in review) for EMAP-E (Figure 3-8). According to the results presented here for the Louisianian Province, EMAP-E has met its performance goal of detecting a 2% change per year in the proportion of area with degraded benthic communities.

Guideline 14: Assessment Thresholds
To facilitate interpretation of indicator results by the user community, threshold values or ranges of values should be proposed that delineate acceptable from unacceptable ecological condition. Justification can be based on documented thresholds, regulatory criteria, historical records, experimental studies, or observed responses at reference sites along a condition gradient. Thresholds may also include safety margins or risk considerations. Regardless, the basis for threshold selection must be documented.

Benthic index values were calculated for the test data set by applying the appropriate weighting factors to the individual components according to the methods outlined in Guideline 3, Table 3-2. The benthic index values were then compared to the original classification of sites as degraded or undegraded. This comparison indicated an overlap in classification of degraded and undegraded sites between benthic index values of 3.5 and 4.5. To be conservative, we extended the threshold values for determining site classification using the benthic index such that index values ≤ 3.0 represented degraded sites, index values ≥ 5.0 represented undegraded sites, and index values between 3 and 5 represented sites with marginal conditions (these sites were misclassified by the original analysis). The distribution of benthic index values among the test sites (Fig. 3-9) shows the overlap of classification as degraded or undegraded by the benthic index with the symbols indicating the original classification of test sites using *a priori* criteria.

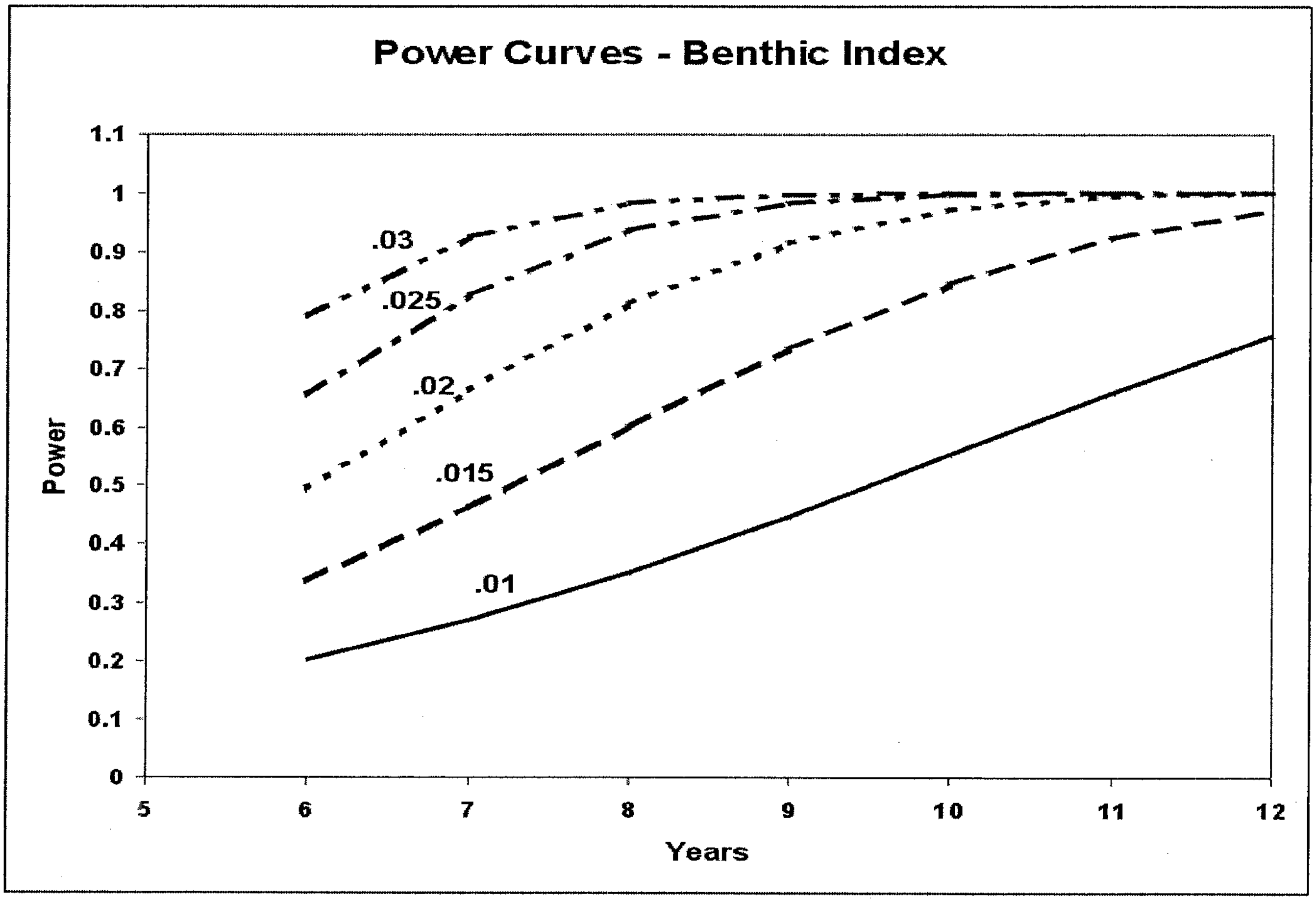


Figure 3-8. Power curves for detecting annual trends of 1% to 3% in the proportion of area with degraded benthic communities based on EMAP-E Louisianian Province data from 1991-1994.

Guideline 15: Linkage to Management Action

Ultimately, an indicator is useful only if it can provide information to support a management decision or to quantify the success of past decisions. Policy makers and resource managers must be able to recognize the implications of indicator results for stewardship, regulation, or research. An indicator with practical application should display one or more of the following characteristics: responsiveness to a specific stressor, linkage to policy indicators, utility in cost-benefit assessments, limitations and boundaries of application, and public understanding and acceptance. Detailed consideration of an indicator's management utility may lead to a re-examination of its conceptual relevance and to a refinement of the original assessment question.

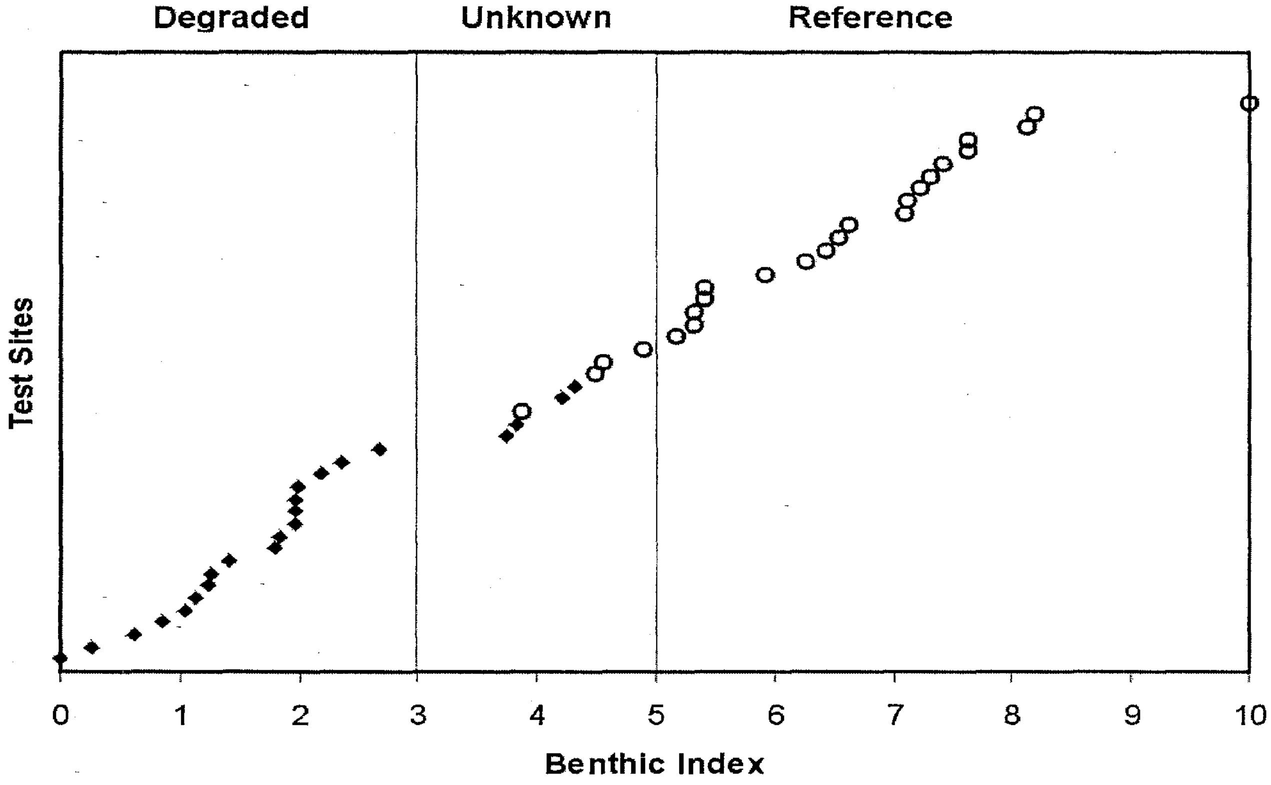


Figure 3-9. Distribution of benthic index values for the test sites where ◆ = degraded sites and o = undegraded sites determined by *a priori* criteria for dissolved oxygen, sediment chemistry, and sediment toxicity. The cutoff value for degraded sites determined by the benthic index is 3.0 and 5.0 is the cutoff value for undegraded sites determined by the benthic index.

The value of an index lies in its applicability across large geographical areas and its ability to provide regional assessments of ecological condition. The information derived from an index of environmental condition such as the benthic index is useful to environmental managers and policy and decision makers who want to identify areas of potential degradation and track the status of environmental condition over time. A benthic index can be used to answer questions about the health of benthic communities in the estuaries of a large geographical region, the spatial or temporal variation of degraded areas of benthic communities, and the status of benthic ecological conditions between the estuaries of different regions.

This benthic index, although developed for EMAP-Estuaries in the Louisianian Province, is easily applied to benthic data from other sampling programs in the northern Gulf of Mexico (see the examples in Engle and Summers [1998]). This benthic index has been successfully applied by others in order to assess benthic conditions in specific estuaries on the Gulf coast. Alabama's Department of Environmental Management effectively used this benthic index to assess the sediment quality in the estuaries of their state (Carlton *et al.* 1998). Similarly, this EMAP benthic index successfully discriminated degraded from undegraded sites in a regional assessment of environmental conditions in Galveston Bay, Texas (C. Gorham-Test, unpublished data). The Texas Natural Resource Conservation Commission applied this benthic index to sites in a targeted study of land use in Galveston Bay and found significant correlations with site rankings based on sediment toxicity tests and sediment chemical concentrations (G. Guillen and L. Broach, pers. comm.). Although the Louisianian Province is geographically widespread, we caution the application of the index outside of this biogeographic region. The environmental stresses affecting the benthos in Gulf of Mexico estuaries may differ from those affecting other regions (*e.g.,* the Mid-Atlantic or Pacific Northwest).

We have successfully synthesized benthic community information into a benthic index of ecological condition that provides environmental managers with an alternate way to assess the status of benthic communities over large geographical areas. A response indicator like the benthic index provides a numerical quantification of the response of the benthic communities to environmental stresses (Summers *et al.* 1995). Because the benthic index is scalable and the criteria for determining the classification of degraded or undegraded are numeric, the application of the benthic index to other estuaries is straightforward.

Summary

The results of this evaluation show that the benthic index is an appropriate indicator for determining the extent of degraded benthic communities in a regional monitoring program. The indicator is conceptually relevant to both the assessment question and ecological function. When part of the collection for a suite of indicators, collecting sediment for benthos is easily implemented and standardized methods are well-established. The greatest cost of implementing the benthic index is in the laboratory processing charges. Temporal and spatial variability have been minimized by both the EMAP-E sample design and rigorous training and QC of the field and laboratory personnel. Probably the greatest concern in the implementation of this indicator is the discriminatory ability of the benthic index. The percent efficiency of the benthic index to classify independent sites was adequate but not as high as we had hoped. Because the EMAP-E design is not limited to specific habitat types but characterizes a region as a whole, there is, inherently, a high degree of variability in the benthic communities in the Gulf of Mexico region. We may have sacrificed a level of precision in favor of a generalized index that is applicable across a wide geographic area.

Potential users have criticized this index approach because of various perceived difficulties in application. Several reviewers have expressed that indices of biotic integrity (IBIs) are easier to understand. We would agree that the IBIs modeled after Karr (1981) are more intuitive in that the models are forced to incorporate the conceptual framework of the developer. The EMAP benthic index employs the same generalized approach but assumes multi-stressor relationships and depends solely on the data to delineate which benthic parameters relate to the observed situation. The IBIs are also perceived to be easier to employ. Clearly, they may be easier to develop than the proposed benthic index but the scoring on multiple habitats of 4 to 7 parameters is certainly more involved than inserting 5 parameters into an equation. The normalization to force the range to be between 0 and 10 is for ease of interpretation and does not need to be done (*i.e.*, if not normalized, the cut-off between poor condition and marginal condition is 0.0). Finally, the index proposed here is applicable over a wide range of environmental conditions and geography and provides comparable scores over these gradients. While the metrics used for an IBI should be selected for broad applicability (*e.g.*, Kerans & Karr 1994), not all applications of the original IBI concept have followed this tenet.

The purpose of monitoring the condition of estuaries on a regional scale was to provide environmental managers with an estimate of baseline conditions. EMAP-E provides regional estimates of estuarine condition against which local managers can compare their estuaries. This further emphasizes the need to evaluate an indicator in the context of a specific monitoring program, as an indicator that may be ideal for one type of program may be inappropriate for another. In this case, the indicator itself could be applied to other, more spatially-specific, monitoring programs if alternative methods were employed. The benthic index in this context is of high value to environmental managers, especially to those concerned with estimating biotic or ecological condition in a quantitative manner.

References

Anderson, B.F. 1966. The Psychology Experiment. Wadsworth: Belmont, California. p. 72.

Baker, J.R. and G.D. Merritt. 1990. Environmental Monitoring and Assessment Program: Guidelines for Preparing Logistics Plans. EPA/600/4-91/001. U.S. Environmental Protection Agency, Las Vegas, Nevada.

Bilyard, G. R. 1987. The value of benthic infauna in marine pollution monitoring studies. *Marine Pollution Bulletin* 18:581-585.

Boesch, D. F. and N. N. Rabalais. 1991. Effects of hypoxia on continental shelf benthos: comparisons between the New York Bight and the northern Gulf of Mexico. p. 27-34 *In* R. V. Tyson and T. H. Pearson (eds.). Modern and Ancient Continental Shelf Anoxia. Geological Society Special Publ. No. 58, Geological Society, London.

Boesch, D. F. and R. Rosenberg. 1981. Response to stress in marine benthic communities, p. 179-200 *In* G. W. Barrett and R. Rosenberg (eds.), Stress Effects on Natural Ecosystems. Wiley-Interscience, New York.

Carlton, J., J.S. Brown, J.K. Summers, V.D. Engle, and P.E. Bourgeois. 1998. A Report on the Condition of the Estuaries of Alabama in 1993-1995: A Program in Progress. Alabama Department of Environmental Management, Mobile, Alabama. 20 pp.

Chapman, P. M. 1989. Current approaches to developing sediment quality criteria. *Environmental Toxicology and Chemistry* 8:589-599.

Dauer, D.M., R.M. Ewing, and A.J. Rodi. 1987. Macrobenthic distribution within the sediment along an estuarine salinity gradient. *Internationale revue der gesamten hydrobiologie* 72:529-538.

Engle, V.D. and J.K. Summers. 1998. Determining the causes of benthic condition. *Environmental Monitoring and Assessment* 51: 381-397.

Engle, V.D. and J.K. Summers. 1999. Refinement, validation, and application of a benthic condition index for Gulf of Mexico Estuaries. *Estuaries* 22:624-635.

Engle, V.D., J.K. Summers, G.R. Gaston. 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. *Estuaries* 17:372-384.

Flint, R. W., S. Rabalais, and R. D. Kalke. 1982. Estuarine benthos and ecosystem functioning, p. 185-201 *In* J. R. Davis(ed.), Proceedings of the Symposium on Recent Benthological Findings in Texas and Adjacent States. Aquatic Sciences Section, Texas Academy of Science, Austin.

Gaston, G. R. 1985. Effects of hypoxia on macrobenthos of the inner shelf off Cameron, Louisiana. *Estuarine, Coastal and Shelf Science* 20:603-613.

Gaston, G.R. and J.C.Nasci. 1988. Trophic structure of macrobenthic communities in the Calcasieu Estuary, Louisiana. *Estuaries* 11:201-211.

Gaston, G. R., P. A. Rutledge, and M. L. Walther. 1985. The effects of hypoxia and brine on recolonization by macrobenthos off Cameron, Louisiana (USA). *Contributions in Marine Science* 28:79-93.

Gaston, G.R. and J.C. Young. 1992. Effects of contaminants on macrobenthic communities in the Upper

Calcasieu Estuary, Louisiana. *Bulletin of Environmental Contamination and Toxicology* 49:922-928.

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

Harper, D. E. Jr., L. D. McKinney, R. R. Salzer, and R. J. Case. 1981. The occurrence of hypoxic bottom water off the upper Texas coast and its effects on the benthic biota. *Contributions in Marine Science* 24:53-79.

Heimbuch, D.G., S.J. Schwager, H.T. Wilson, and J.K. Summers. Power Analysis for Tests for Long Term Trend in Environmental Conditions of Estuaries. Contribution Number 948 of the U.S. Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL. (in review)

Heitmuller, P.T. and R. Valente. 1991. Environmental Monitoring and Assessment Program: EMAP-Estuaries Louisianian Province: 1991 Quality Assurance Project Plan. EPA/ERL-GB No. SR-120. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL 32561.

Holland, A. F., A. T. Shaughnessy, and M. H. Hiegel. 1987. Long-term variation in mesohaline Chesapeake Bay macrobenthos: spatial and temporal patterns. *Estuaries* 10:227-245.

Hyland, J.L., T.R. Snoots, and W.L. Balthis. 1998. Sediment quality of estuaries in the southeastern U.S. *Environmental Monitoring and Assessment* 51:331-343.

Karr, J.R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6:21-27.

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

Karr, J.R. 1993. Defining and assessing ecological integrity: beyond water quality. *Environmental Toxicology and Chemistry* 12:1521-1531.

Kerans, B.L. and J.R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* 4:768-785.

Keppel, G. and W.H. Saufley, Jr. 1980. Introduction to Design and Analysis. W. H. Freeman and Co.: New York, New York. p. 105

Lang, C. and O. Reymond. 1995. An improved index of environmental quality for Swiss rivers based on benthic invertebrates. *Aquatic Sciences* 57:172-180.

Lang, C., G. L'Eplattenier, and O. Reymond. 1989. Water quality in rivers of western Switzerland: application of an adaptable index based on benthic invertebrates. *Aquatic Sciences* 51:224-234.

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* 31:117-139.

Lenat, D.R. 1988. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. *Journal of the North American Benthological Society.* 7:222-233.

Macauley, J.M. 1991. Environmental Monitoring and Assessment Program-Near Coastal Louisianian Province: 1991 Monitoring Demonstration. Field Operations Manual. EPA/600/X-91/XXX. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL 32561.

Macauley, J.M., J.K. Summers, P.T. Heitmuller, V.D. Engle, G.T. Brooks, M. Babikow, and A.M. Adams. 1994. Annual Statistical Summary: EMAP-Estuaries Louisianian Province - 1992. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL 32561. EPA/620/R-94/002.

Macauley, J.M., J.K. Summers, P.T. Heitmuller, V.D. Engle, and A.M. Adams. 1996. Annual Statistical Summary: EMAP-Estuaries Louisianian Province - 1993. U.S. Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf Ecology Division, Gulf Breeze, FL 32561.

McManus, J. W. and D. Pauly. 1990. Measuring ecological stress: variations on a theme by R. M. Warwick. *Marine Biology* 106:305-308.

Messer, J.J., R.A. Linthurst, and W.S. Overton. 1991. An EPA program for monitoring ecological status and trends. *Environmental Monitoring and Assessment* 17:67-78.

Plafkin, J.L., M.T. Barbour, K.D. Porter, S.K. Gross, R.M. Hughes. 1989. Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. EPA/440/4-89/001. U.S. Environmental Protection Agency, Office of Water, Assessment and Watershed Protection Division, Washington, D.C.

Pearson, T. H. and R. Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology Annual Review* 16:229-311.

Rakocinski, C.F., S.S. Brown, G.R. Gaston, R.W. Heard, W.W. Walker, and J.K. Summers. 1997. Macrobenthic responses to natural an contaminant-related gradients in northern Gulf of Mexico estuaries. *Ecological Applications* 7:1278-1298.

Ranasinghe, J.A., S.B. Weisberg, J.B. Frithsen, D.M. Dauer, L.C. Schaffner, and R.J. Diaz. 1994. Chesapeake Bay Benthic Community Restoration Goals. Report CBP/TRS 107/94. U.S. Environmental Protection Agency, Chesapeake Bay Program, Annapolis, Maryland.

Reish, D. J. 1986. Benthic invertebrates as indicators of marine pollution: 35 years of study. *Oceans 86* 3:885-888.

Rosenberg, R. 1977. Benthic macrofaunal dynamics, production, and dispersion in an oxygen-deficient estuary of West Sweden. *Journal of Experimental Marine Biology and Ecology* 26:107-133.

Rygg, B. 1986. Heavy-metal pollution and log-normal distribution of individuals among species in benthic communities. *Marine Pollution Bulletin* 17:31-36.

Santos, S.L. and J.L. Simon. 1980. Response of soft-bottom benthos to annual catastrophic disturbance in a south Florida estuary. *Marine Ecology Progress Series* 3: 347-355.

Scott, K.J. 1990. Indicator strategy for near-coastal waters, *In*: Hunsaker, C.T. and D.E. Carpenter (eds.) Environmental Monitoring and Assessment Program: Ecological Indicators. EPA 600/3-90/060. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.

Sheehan, P.J. 1984. Effects on community and ecosystem structure and dynamics, pp. 52-99 *In*: Sheehan, P.J., D.R. Miller, G.C. Butler, and Ph. Bourdeau (eds.) Effects of Pollutants at the Ecosystem Level. John Wiley & Sons: Chichester. 443 p.

Stokes, M. E., C. S. Davis, and G. G. Kock. 1995. Categorical Data Analysis using the SAS System. SAS Institute Inc., Cary, North Carolina. 499 pp.

Summers, J.K., J.M. Macauley and P.T. Heitmuller. 1991. Environmental Monitoring and Assessment Program. Implementation Plan for Monitoring the Estuarine Waters of the Louisianian Province - 1991 Demonstration. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL 32561. EPA/600/5-91/228.

Summers, J.K., J.M. Macauley, V.D. Engle, G.T. Brooks, P.T. Heitmuller, A.M. Adams. 1993. Louisianian Province Demonstration Report: EMAP-Estuaries - 1991. EPA/600/R-94/001. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL.

Summers, J.K., J.M. Macauley, J.M., P.T. Heitmuller, V.D. Engle, A.M. Adams and G.T. Brooks. 1992. Annual Statistical Summary: EMAP-Estuaries Louisianian Province - 1991. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Gulf Breeze, FL 32561. EPA/600/R-93/001.

Summers, J.K., J.F. Paul, and A. Robertson. 1995. Monitoring the ecological condition of estuaries in the United States. *Toxicological and Environmental Chemistry* 49:93-108.

U.S. EPA. 1995. Environmental Monitoring and Assessment Program (EMAP): Laboratory Methods Manual - Estuaries, Volume 1: Biological and Physical Analyses. United States Environmental Protection Agency, Office of Research and Development, Narragansett, RI. EPA/620/R-95/008.

Van Dolah, R.F., J.L. Hyland, A.F. Holland, J.S. Rosen, and T.R. Snoots. 1999. A benthic index for assessing sediment quality in estuaries of the southeastern United States. *Marine Environmental Research* 48:269-283.

Warwick, R.M. 1986. A new method for detecting pollution effects on marine macrobenthic communities. *Marine Biology* 92:557-562.

Weisberg, S.B., J.A. Ranasinghe, D.M. Dauer, L.C. Schaffner, R.J. Diaz, and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. *Estuaries* 20:149-158.

Wilson, J.G. and D.W. Jeffrey. 1994. Benthic biological pollution indices in estuaries, p. 311-327. *In*: K.J.M. Kramer (ed.), Biomonitoring of Coastal Waters and Estuaries. CRC Press, Boca Raton, Florida.

Sources of Unpublished Materials:

Cynthia Gorham-Test, USEPA REGION 6, Fountain Place, 1445 Ross Avenue, Dallas, TX 75202-2733

George Guillen and Linda Broach, Texas Natural Resource Conservation Commission, 5425 Polk Ave., Suite H, Houston, Texas 77023

Chapter Four

Application of the Indicator Evaluation Guidelines to a Multimetric Indicator of Ecological Condition Based on Stream Fish Assemblages

Frank H. McCormick, U.S. EPA, National Exposure Research Laboratory, Ecological Exposure Research Division, Cincinnati, OH
David V. Peck, U.S. EPA, National Health and Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, OR

In this chapter, we employ the guidelines presented in Chapter 1 to evaluate a complex (*i.e.*, multiple components) indicator of ecological condition based on stream fish assemblages. This indicator is being modified and developed for implementation in a specific monitoring effort (the Mid-Atlantic Highlands Assessment, described below) designed to address specific regional-scale assessment questions.

This chapter does not provide complete documentation of the indicator or the process of its development. Our primary intent is to provide examples of the type of information that is appropriate to address each evaluation guideline for the indicator. In some cases, examples are presented based on hypothetical or simulated data, and in some cases, not all available information pertinent to a comprehensive evaluation is provided. More complete documentation of the development and evaluation of the suitability of this indicator exists or will be forthcoming in various scientific journals.

The indicator is a composite index, and its development is based on the multimetric Index of Biotic Integrity (IBI) originally developed by Karr (Karr 1981, Karr *et al.* 1986). The IBI was developed to assess the condition of water bodies by direct evaluation of biological attributes (Karr 1981, Karr and Dudley 1981, Karr 1991). Multimetric indicators such as the IBI are based on the premise that biological data represent a means to integrate various structural and functional attributes of an ecosystem and provide an overall assessment of ecosystem condition (Fausch *et al.* 1990, Karr 1991, Karr and Chu 1997). Biological and socioeconomic characteristics of stream fish assemblages, including the capability to integrate the effects of a variety of stressors across different time scales and levels of ecological organization, and the importance and familiarity of fishes to the general public, make them conducive to the development of an indicator of ecological condition (Table 4-1).

Some important features of the indicator are presented in Table 4-2. The development of the indicator is based on accepted ecological and mathematical principles. Various critical structural and functional attributes of the biotic components of an ecosystem (e.g., taxonomic richness, trophic structure) believed to respond predictably to increasing intensities of human disturbance are represented by different metrics (Karr 1986, Karr 1991, Barbour *et al.* 1995). Metrics are derived from species composition and relative abundance data of a particular ecological assemblage or community (stream fish in this example) collected at individual sampling sites. A final suite of metrics is selected for use in developing the indicator, based on responsiveness to biotic or abiotic conditions resulting from increasing human disturbance, and their biological importance. For each sampling site, the response value for each metric selected is transformed to a metric score. The score for each metric is based on the degree of deviation of the response value from that expected at a similar site under conditions of minimal human disturbance. The individual metric scores are then aggregated to produce an overall indicator score. A higher score indicates better ecological condition (*i.e.*, closer to the expected condition when human disturbance is minimal). More detailed descriptions of the general approach used to develop multimetric indices can be found in Fausch *et al.* (1984), Karr *et al.* (1986), Karr (1991), Plafkin *et al.* (1989), Gibson (1994), Barbour *et al.* (1995), Simon and Lyons (1995), and Karr and Chu (1997). Simon and Lyons (1995) and Karr and Chu (1997) summarize and address criticisms of the multimetric approach to developing indicators of condition.

Multimetric indicators based on modifications to the original IBI concept of Karr (Simon and Lyons 1995) have been developed for use in various geographic areas in the United States and elsewhere (Miller *et al.* 1988, Lyons *et al.* 1995, Yoder and Rankin 1995, Lyons *et al.* 1996, Hughes and Oberdorf 1999), various systems (Jordan *et al.* 1993) and taxa (Lenat 1993, Kerans and Karr 1994, Fore *et al.* 1994, DeShon 1995, Fore *et al.* 1996, Barbour *et al.* 1996). Many of these studies address evaluations of the indicator, approximating some of ORD's evaluation guidelines as outlined in Chapter 1, and use a variety of approaches to address a particular guideline.

Table 4-1. Rationale for indicators of ecological condition based on stream fish assemblages[1]

- Historical data available
- Autecology of most species described
- Integrates effects of stressors at various scales, time periods, and levels of organization
- Includes long-lived and mobile species
- Assemblage composed of populations, individuals
- Important resource to humans
- High level of familiarity with general public

[1] Compiled and summarized from Karr *et al.* (1986), Plafkin *et al.* (1989), Simon (1991), and Simon and Lyons (1995).

The stream fish assemblage indicator is being developed using data collected as part of the Mid-Atlantic Highlands Assessment (MAHA). This study was funded by the EPA Environmental Monitoring and Assessment Program (EMAP; *e.g.*, Whittier and Paulsen 1992) in conjunction with a Regional-EMAP (R-EMAP) effort (U.S. EPA 1997). The MAHA study represents a partnership between EMAP and EPA Region III to develop and demonstrate EMAP approaches such as probability-based survey designs and appropriate indicators of ecological condition to address specific assessment questions of interest to the Region.

The monitoring framework for MAHA consists of a regional-scale probability-based survey design to select sampling sites. This design permits unbiased inferences to be made with known certainty from the subset of sites where samples and data are collected to explicitly defined populations of ecological resource units (Larsen 1995, 1997, Diaz-Ramos *et al.* 1996). For MAHA, populations are defined based on the total length of streams. The design allows one to estimate the total length of streams in the target population (*e.g.*, all permanent streams appearing on a particular scale of map) which meet some criteria (*e.g.*, all first-order target streams, all target streams within a specific ecoregion). The distribution of indicator scores can then be examined for these defined populations to determine the estimated length of stream characterized by a particular set of indicator values, with associated uncertainty in these estimates represented by confidence bounds.

Multimetric indicators developed for a particular geographic area and scale of monitoring effort should not be applied to other scales of monitoring or other geographic areas without evaluation and modification. Because of differences in the biological assemblage structure and composition, and different expectations of conditions associated with minimal human disturbance, component metrics may require substitution and validation (Miller *et al.* 1988, Barbour *et al.* 1995). Thus, previously existing multimetric indicators may not be useful for the MAHA geographic region and monitoring framework. Likewise, this assemblage indicator should not be used in any other monitoring context and/or geographic area.

Table 4-2. Characteristics of indicator

Basic data from fish assemblages required:

- Presence/absence of species
- Abundance of individual species

Requires representative data on fish species composition and abundance be collected at each sampling site.

Metrics: Categories of species or individuals representing various aspects of ecology and life history in the assemblage

- Species richness and composition
- Abundance and individual condition
- Trophic function
- Reproductive function

Each fish species assigned to categories within each metric based on life history characteristics.

Score for a particular metric based on comparison of observed responses to expectations under conditions of minimal human disturbance.

- Each metric is assigned a score of between 0 and 10 based on the similarity of the observed response to expectations.

Indicator score is computed as the sum of individual metric scores.

- Metrics are equally weighted
- Indicator score is re-scaled to be between 0 and 100

  0 = no fish collected at site

  100 = site meets all expectations for conditions under minimal human disturbance

Assessment question(s) addressed by extrapolating indicator values from probability sample of sites to entire target resource population (stream length).

Public Perception

- Indicator score easily understood
- Once developed and validated, indicator does not require sophisticated technical expertise to interpret

The example indicator is modified from other multimetric indicators developed previously to tailor it to a specific geographic region (mid-Atlantic highlands), the characteristic ichthyofauna of the region, and the proposed monitoring framework (regional scale survey design). Some metrics are modified to be more generic (*i.e.*, to include more species) to account for the fact that not all fish species are distributed throughout the target region. Expectations for the responses of various metrics are modified to be more appropriate for the geographic region and extant ichthyofauna.

In this presentation of a stream fish assemblage indicator for ecological condition, we have re-stated each individual guideline presented in Chapter 1 for convenience and easy reference. In order to demonstrate the application of relevant information to the guideline, we have developed Performance Objectives, or brief descriptions of our interpretation of the specific needs for each guideline based on the specific type of indicator and the proposed monitoring framework. After presentation and discussion of pertinent information, we offer a summary of findings regarding the suitability of the indicator with respect to each guideline.

Phase1: Conceptual Relevance

Guideline 1: Relevance to the Assessment
Early in the evaluation process, it must be demonstrated in concept that the proposed indicator is responsive to an identified assessment question and will provide information useful to a management decision. For indicators requiring multiple measurements (indices or aggregates), the relevance of each measurement to the management objective should be identified. In addition, the indicator should be evaluated for its potential to contribute information as part of a suite of indicators designed to address multiple assessment questions. The ability of the proposed indicator to complement indicators at other scales and levels of biological organization should also be considered. Redundancy with existing indicators may be permissible, particularly if improved performance or some unique and critical information is anticipated from the proposed indicator.

Performance objectives

1. Demonstrate that the indicator is linked to an identified assessment question
2. Discuss its role in contributing information to address multiple assessment questions
3. Demonstrate the complementarity and minimal redundancy with other potential indicators

The design of the MAHA study was driven in part by a series of specific assessment questions that collectively would provide the means to determine the status and extent of the condition of stream resources in the region with respect to the societal value of biological integrity (as defined by Karr and Dudley 1981). The principal questions pertaining to stream fish assemblages are presented in Table 4-3 (U.S. EPA 1997). The nature of these questions suggests that an appropriate indicator should focus at the assemblage level and consist of multiple components to address the various aspects of the questions. In addition, to yield representative estimates of status and extent of stream resources with respect to biological integrity, a monitoring framework based on a probability-based survey design is required. The example multimetric indicator, applied in conjunction with the appropriate sampling design, meets all of the requirements to address the principal assessment question. The indicator can address all three components of the principal assessment question by including appropriate metrics (*e.g.*, the number of species considered to be sensitive to human disturbance).

The indicator is also useful in that the basic fish species and abundance data used to develop it can also be used with little or no additional effort to address other assessment questions of interest (Table 4-3) (U.S. EPA

1997). These subsidiary questions are relevant to a separate societal value of interest to the MAHA study, fishery health.

There are several possible complementary relationships with other indicators (Table 4-4). It can be used as part of a suite of indicators to address multiple assessment questions, or can provide a more complete assessment of biological integrity when combined with other indicators using biological assemblages.

Component metrics of the indicator are selected based on their hypothesized response to stressors which are monitored at different scales and incorporate information from different levels of biological organization. Possible causes of poor condition as determined by the indicator can be identified (although specific cause-effect relationships cannot always be ascertained) by examining correlations between the indicator or component metrics and various measures of ecosystem stress (measurement variables or multi-component indicators). Finally, the potential exists that the indicator may provide information that is highly redundant with indicators derived from stream macroinvertebrate assemblages. This has yet to be evaluated for the MAHA study. DeShon (1995) reported that multimetric indicators for fish and benthic macroinvertebrates provided complementary, rather than redundant, information when compared in Ohio streams.

Table 4-3. Assessment questions driving development of indicator

Principal Question Relevant to the Societal Value of Biological Integrity

What % of stream miles (and spatial distribution) have fish assemblages that differ from "reference" condition as measured by:

- Species richness?
- Number of species and/or % individuals of species sensitive to human disturbance?
- Cumulative index of biotic integrity based on fish assemblage?

Subsidiary Questions Relevant to the Societal Value of Fishery Health

What % of stream miles have game fish?
Which species are most abundant or widely distributed?
What is the % of stream miles with game fish classified by stream order?
What % of stream miles support coldwater *vs.* warmwater fisheries as determined by the fish species?

Specific assemblages of interest include:

- Cold water (*e.g.*, salmon, trout)
- Cool water (*e.g.*, smallmouth bass)
- Warm water (*e.g.*, largemouth bass, sunfish)

Table 4-4. Relationship to other indicators

Complementarity

Potential to be combined with other condition indicators to provide a more complete idea of overall biotic integrity or sustainability

- Macroinvertebrate assemblages
- Periphyton assemblages
- Index of Well Being (Gammon 1976)
- Abiotic condition indicators (*e.g.*, habitat quality or chemical quality)

Metrics can be selected that are linked to stressors that can be monitored at different scales: site level, watershed level, and landscape level.

Indicator incorporates information at various levels of biological organization[1]

- Assemblage (community): species richness, trophic composition, habitat guilds
- Population: abundance, life history/reproductive strategies

Associations between indicator (or component metrics) and other stressor indicators (e.g., habitat disturbance, chemical water quality) can be examined to identify possible causes of impairment.

Redundancy

Potentially redundant with other condition indicators based on assemblages (*e.g.*, macroinvertebrates, periphyton), but this has not been empirically demonstrated.

[1] Modified from Table 7 in U.S. EPA (1997)

Summary

The indicator and associated monitoring framework are linked to a specific assessment question developed for use in the mid-Atlantic highlands as part of a program to determine ecological condition of freshwater streams. Ancillary questions related to separate societal values (*e.g.*, fishery health) can also be addressed with the indicator, its components, or its basic measurement data. The indicator, in conjunction with other condition or stressor indicators monitored at other scales or levels of biological organization, can contribute information to address multiple assessment questions or provide a capability to diagnose possible causes of impairment. The potential for providing redundant information with other condition indicators based on different types of assemblages or communities is identified, but has not been empirically demonstrated or evaluated for this monitoring program.

Guideline 2: Relevance to Ecological Function
It must be demonstrated that the proposed indicator is conceptually linked to the ecological function of concern. A straightforward link may require only a brief explanation. If the link is indirect or if the indicator itself is particularly complex, ecological relevance should be clarified with a description, or conceptual model. A conceptual model is recommended, for example, if an indicator is comprised of multiple measurements or if it will contribute to a weighted index. In such cases, the relevance of each component to ecological function and to the index should be described. At a minimum, explanations and models should include the principal stressors that are presumed to impact the indicator, as well as the resulting ecological response. This information should be supported by available environmental, ecological and resource management literature.

Performance Objectives

1. Demonstrate conceptual linkages between ecosystem components and principal stressors.
2. Demonstrate conceptual linkages between principal stressors believed to cause impairment with respect to the societal value of interest, ecosystem responses to these stressors, and the indicator (or its components).

Basic relationships between major structural components and processes can be graphically represented (Figure 4-1, modified from Hughes *et al.* 1994) to illustrate possible routes of exposure from anthropogenic stressors. This diagram also points out the location and functional roles of fish assemblages to demonstrate those stressor-response relationships that can be effectively monitored with a fish assemblage indicator. Fish assemblages can be used to assess condition both in the water column and bottom habitats, and can provide information from multiple trophic levels.

More specific hypotheses have been developed regarding the relationship of indicator metrics with anthropogenic stressors (Fig. 4-2). This approach is based on a model originally conceived by Karr *et al.* (1986). We have modified the model to organize it by types of major stressors (following terminology presented in U.S. EPA 1997). This representation shows direct linkages between individual metrics and each type of stressor, and helps to illustrate the diagnostic capability of the indicator since different scores for individual components can be associated with responses to certain groups of stressors.

The suite of candidate metrics represent those selected after a screening process to eliminate those that were not responsive to hypothesized stressors of interest, were redundant in their information content, or were otherwise not suited for application in the proposed monitoring framework and/or geographic region of interest. The metrics shown in Table 4-5 provide information about the ecological relevance of each component metric; consequently Table 4-5 could be considered one type of conceptual "model" of the indicator. It demonstrates anticipated responses of component metrics to various types of stressors, based on characteristics of fish assemblages in environmentally degraded systems described by Fausch *et al.* (1990).

Further demonstrating the diagnostic capability and potential discriminatory ability of the indicator (addressed more completely as part of Guideline 12), Figure 4-3 shows the range of response for each individual metric, based on summaries in Angermeier and Karr (1986) and Karr (1991). Some metrics (*e.g.*, species richness) will exhibit a response over the entire range of condition, while others help to discriminate either very good condition (*e.g.*, sensitive species richness) or very poor condition (*e.g.*, proportion of tolerant individuals).

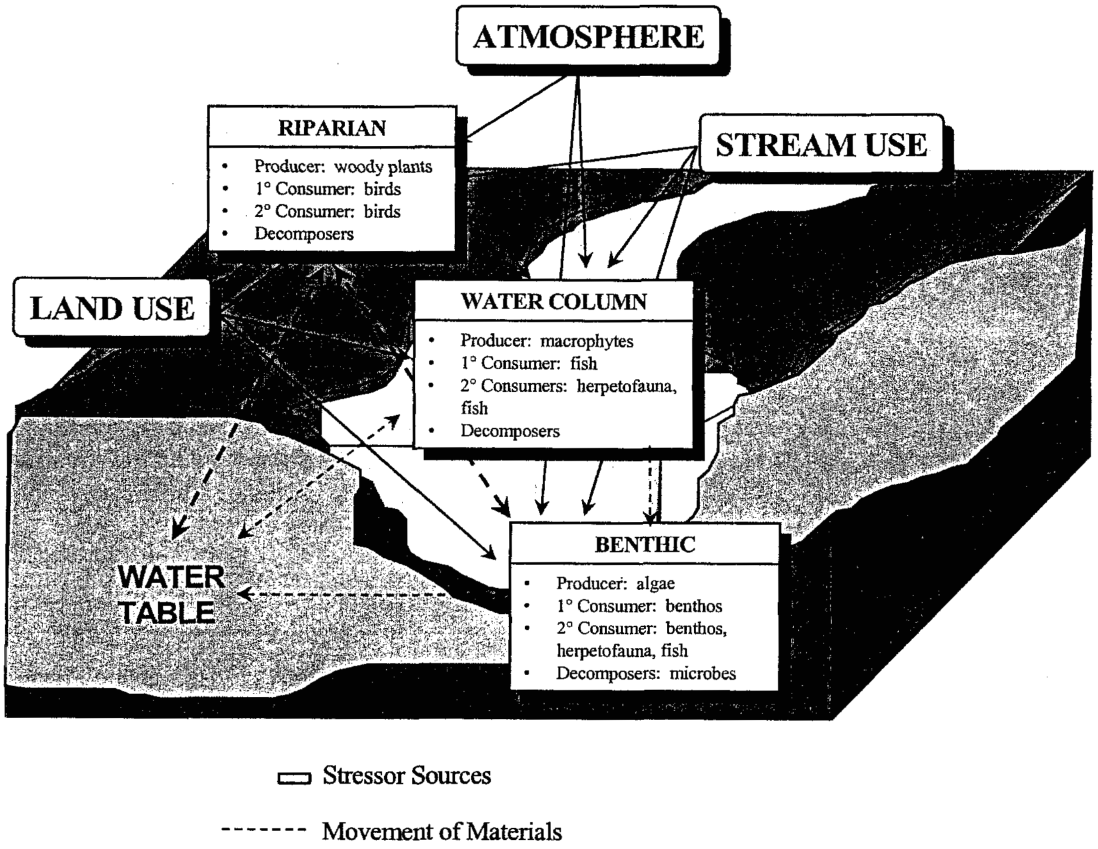
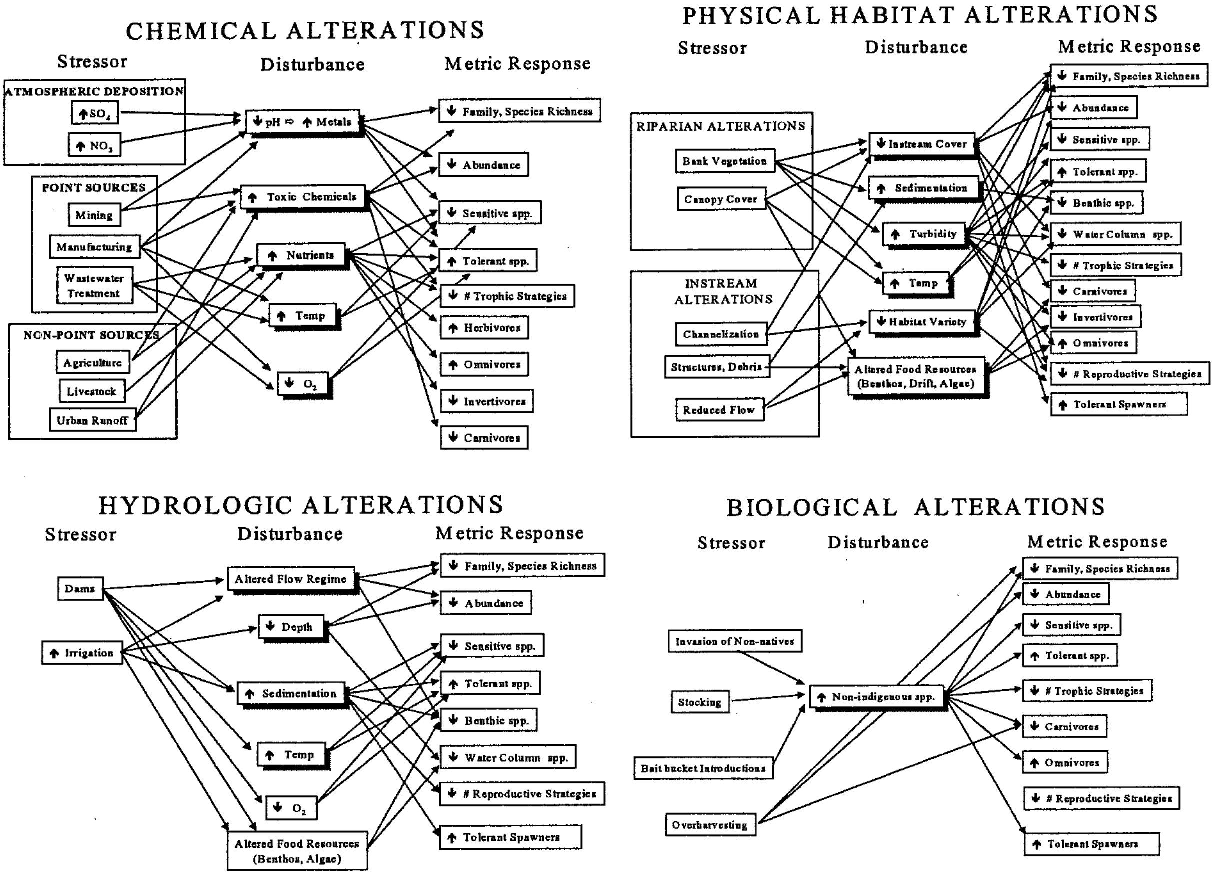


Figure 4-1. Conceptual model of major structural and functional components of a stream ecosystem modified from Hughes *et al.* (1994).



CHEMICAL ALTERATIONS

PHYSICAL HABITAT ALTERATIONS

HYDROLOGIC ALTERATIONS

BIOLOGICAL ALTERATIONS

Figure 4-2. Conceptual model of indicator, showing linkages between various types and classes of stressors and component metrics [Derived from Karr (1985), Fausch *et al.* (1990), and U.S. EPA (1997)].

Table 4-5. Description of component metrics of example indicator and conceptual linkages to stressors

Metric	Description, Rationale, and Linkage to Stressors	Response To Increased Human Disturbance[1]
	Species Richness and Composition Metrics[2]	
Native Species Richness	Number of different native species collected. Measure of biodiversity.	Decrease
Native Family Richness	Number of different families represented in sample. Assess degree to which stream reach supports families represented by only one or a few species.	Decrease
Native Benthic Species Richness	Species adapted for utilizing bottom habitat. Measure of substrate habitat quality and quality of riffle habitats, affected by sedimentation.	Decrease
Native Water Column Species Richness	Species adapted for utilizing surface and midwater habitats. Measure of habitat quality (esp. pools); affected by increased turbidity.	Decrease
	Abundance and Condition Metrics[2]	
Total Abundance	Number of individuals collected. Measure of relative productivity.	Decrease
	Indicator Species Metrics[2]	
Proportion of Individuals as Non-native Species	Measure of biological "pollution," also a biological stressor to native fish populations.	Increase
Sensitive Species Richness	First species lost following human disturbance, last to recover following restoration	Decrease
Proportion of Individuals from Tolerant Species	Individuals that are tolerant of disturbance or extremes in environmental restoration conditions.	Decrease
	Trophic Function Metrics[2]	
Number of Trophic Strategies	Excluding omnivores; measure of trophic/food web complexity of fish assemblage.	Decrease
Proportion of Individuals as Top Carnivores	Measure of ability of food chain to support top level; affected by toxics, turbidity.	Decrease
Proportion of Individuals as Invertivores	Measure of capacity of system to support primary consumers (major trophic group of fishes).	Decrease
Proportion of Individuals as Omnivores	Trophic generalists feeding on variety of plant and animal material.	Increase
Proportion of Individuals as Herbivores	Taxa that feed exclusively on plants and algae.	Increase
	Reproductive Function Metrics[2]	
Number of Reproductive Strategies	Excludes strategies tolerant to siltation. Measure of ability of stream reach to support a variety of reproductive strategies; affected by toxics, turbidity, sedimentation.	Decrease
Proportion of Individuals as Tolerant Spawners	Species that can reproduce under a broad range of habitat conditions with no special requirements for spawning to occur.	Increase

[1] Based on Fausch *et al.* (1990) and Hoefs and Boyle (1992)

[2] Terminology based on Simon and Lyons (1995)

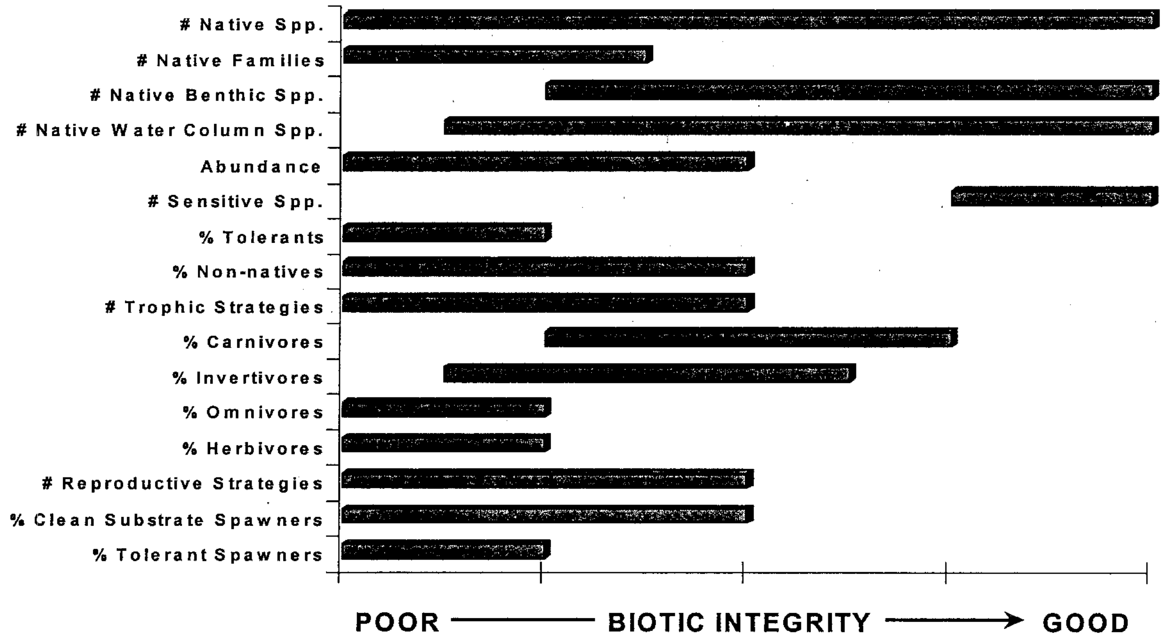


Figure 4-3. Range of ecological condition (as expressed by biotic integrity) over which individual metrics comprising the indicator are expected to respond [Modified from Karr and Angemeier (1986) and Karr (1991)].

Summary

Conceptual linkages between air, land and stream stressors of stream ecosystems with riparian, water column and benthic receptors are presented. Principal stressor types include chemical, biological, hydrologic and physical habitat alterations. Two different approaches were used to demonstrate conceptual linkages between principal stressors believed to cause impairment with respect to the societal value of interest (biological integrity), ecosystem responses to these stressors, and the indicator (or its components).

Phase 2: Feasibility of Implementation

Guideline 3: Data Collection Methods
Methods for collecting all indicator measurements should be described. Standard, well-documented methods are preferred. Novel methods should be defended with evidence of effective performance and, if applicable, with comparisons to standard methods. If multiple methods are necessary to accommodate diverse circumstances at different sites, the effects on data comparability across sites must be addressed. Expected sources of error should be evaluated.

Methods should be compatible with the monitoring design of the program for which the indicator is intended. Plot design and measurements should be appropriate for the spatial scale of analysis. Needs for specialized equipment and expertise should be identified.

Sampling activities for indicator measurements should not significantly disturb a site. Evidence should be provided to ensure that measurements made during a single visit do not affect the same measurement at subsequent visits or, in the case of integrated sampling regimes, simultaneous measurements at the site. Also, sampling should not create an adverse impact on protected species, species of special concern, or protected habitats.

Performance Objectives

1. Clearly describe all methods required to obtain field measurement data for the indicator. Demonstrate performance and compatibility with standard methods if necessary.
2. Demonstrate that the plot design (*e.g.*, reach length, index period) associated with methods is appropriate for proposed monitoring framework.
3. Describe equipment and technical expertise required to successfully implement methods.
4. Demonstrate that the proposed sampling design and methods have a low impact on the environment and other potential indicator measurements.
5. Identify and evaluate sources of error associated with implementing a particular method.

Information regarding the basic procedure used at each sampling site to obtain values for the indicator is presented in Table 4-6. Collection of field data at an individual sampling site is based on standard approaches for stream fish assemblages (McCormick and Hughes 1998). References are presented in Table 4-6 that more completely document the procedures, and point out possible compatibilities with other monitoring efforts. Laboratory methods include confirming field identifications of fish species; confirmation should be conducted by a recognized taxonomic expert on the regional ichthyofauna. Other activities include compiling available life history information on each fish species in preparation for making assignments to individual metric categories (*e.g.*, sensitive species, trophic function, type of reproductive strategy). Finally, information regarding the composition and structure of fish assemblages in the region of interest that might be expected under conditions of minimal human disturbance must be obtained to develop expectations for each metric.

We also present analytical methods used to develop the indicator from measurements of fish assemblages. Most of these procedures are based on standard approaches published for multimetric indicators. We point out deviations from the standard approach and their rationale. We also point out how indicator values are coupled with the proposed monitoring framework to produce a distribution of indicator scores applicable to stream resource populations as described in the Introduction.

Table 4-6. Summary of procedures to obtain measurement data and indicator values at each sampling site.

Field Procedures:

Standard sampling gears and techniques:

"Best Effort" sampling using a combination of gear types, standardized sampling times and distances (40 times mean channel width (Lyons 1992) or 150 m, whichever is greater).

- Electrofishing
- Seining

Identify individual fish to species and enumerate.

Documented Protocols:

Lazorchak *et al.* (1998): EMAP Surface Waters-Streams Methods Manual
Similar protocols used by other large-scale monitoring programs.

- Meador *et al.* (1993): National Water Quality Assessment Program (NAWQA)

Laboratory Procedures:

Confirm species identifications from voucher specimens.
Compile autecological information for each species from published sources of life history data.
Obtain information regarding assemblage composition and structure under conditions of minimal human disturbance.

Data Analysis Procedures:

Standard approaches for multimetric-based indicators (Karr *et al.* 1986, Plafkin *et al.* 1989, Klemm *et al.* 1993, Barbour *et al.* 1995, Simon and Lyons 1995).

Compute response values for each metric (*e.g.*, species richness, proportion of tolerant individuals) based on abundance data and autecological information for each fish species.

Determine expected conditions for each metric response.

- Metrics based on numbers of species require calibration for stream size.

  Develop "maximum species richness lines" (Fausch *et al.* 1984)
  Data from all sites used, rather than just "reference sites" (Simon and Lyons 1995).

- Metrics based on proportion of individuals require expectations based on composition of a fish assemblage under conditions of minimal human disturbance.

  Expectations modified when possible, based on knowledge of historical assemblages prior to European settlement (Hughes 1995).

Compute scores for each metric based on deviation of response from expectations.

- 0 to 10 scale: Differs from standard approach (1,3, 5): Provides more continuous distribution of scores (Hughes *et al.* 1998).

Sum metric scores to produce indicator value.

- Rescale to be between 0 and 100: differs from standard approach (no rescaling)

Use EMAP techniques (*e.g.*, Diaz-Ramos *et al.* 1995) to calculate distribution of indicator values in estimated resource populations based on probability sample.

Characteristics and issues associated with the application of methods to an individual sampling site within the proposed monitoring framework are presented in Table 4-7 (Hughes 1993, Lazorchak *et al.* 1998). "Plot design" refers to the approach required to obtain representative data on the fish assemblage from an individual sampling site. Plot design involves considerations such as when to sample, where to sample within a designated site, and how many individual samples are required from each site. Probability-based survey designs result in sites being selected at random, without regard to ease of access or other aspects of location.

Table 4-7. Features of monitoring framework and plot design

Monitoring Framework

Target resource is wadeable streams (Strahler order 1 through 3).
Survey design provides synoptic information about spatial distribution and extent of condition in target resource populations.

Limited utility in describing condition at an individual site due to lack of replication.

Indicator specifically developed for use in mid-Atlantic highlands region.
Indicator is principally retrospective; anticipatory capability is low, although some metrics (*e.g.*, number of sensitive species) may provide an early warning of potential degradation.

Plot Design

Characteristics:

When to sample (Index period): Once per year during the summer baseflow period.
Where to sample: All habitats within a defined length of stream (based on mean width).
Number of samples per visit: One composite sample of fish assemblage is created from collections made with appropriate sampling gear.

Issues and constraints within proposed monitoring framework:

Site Inaccessibility:

- Access rights to private land
- Remote locations away from roads

Small proportion of sampling units are potentially affected by

- Restrictions Imposed on State and Federal Scientific Collection Permits
- Species: threatened, endangered, economically valuable
- Sites: Wilderness areas, parks, preserves
- Gear types allowed at sites

Accuracy and consistency of field identifications among field crews.
Control measures include:

- Consistent training
- Performance evaluations against experts
- Use of experienced personnel (Federal, state and university)
- Consistent protocol for vouchering specimens for confirmation of species identifications to allow for data correction when necessary
- Field audits

Thus, the possibility exists that some proportion of sampling sites will be located within protected areas or within ranges of protected fish species. Eventual access to these sites may be denied or possibly restricted (in terms of when sampling can occur or how samples may be obtained) by State or Federal agencies when scientific collecting permits are issued. Finally, a large-scale sampling effort requiring multiple field crews requires consistent implementation of the sampling and data acquisition procedures to permit robust comparisons of data across sites sampled by different crews. Table 4-7 presents measures used to control for crew differences.

Specialized equipment needs and technical requirements for field and laboratory personnel are presented in Table 4-8. No specialized sampling or analytical equipment or instrumentation is required for the indicator. Some level of technical expertise is required to support the collection of data for the indicator, especially in the areas of ichthyology, fisheries biology and aquatic ecology. Some of this expertise can be gained through specialized training programs, or addressed through staffing schemes considered under Guideline 4. Karr (1991) points out the need for experienced professional fisheries biologists in the initial development and subsequent interpretation of the indicator values.

Table 4- 8. Equipment and technical expertise requirements

Specialized Field or Analytical Equipment

Backpack, bank or boat-mounted electrofishing unit, seines, nets.

Technical Expertise

Field:

Ability to identify majority of common fish species in field

- Especially state-listed species, larger species and sport fish (on which sampling restrictions may be placed) which are identified and released

Ability to operate different types of sampling gear safely and effectively

- Electrofishing
- Seining

Laboratory:

Ability to identify all fish species in region from preserved specimens

- Especially small, non-game fish (*e.g.*, minnows)

Ability to review literature , compile information, and categorize individual fish species regarding life history characteristics, tolerance to disturbance, etc.

Data analysis and interpretation:

Critical that professional fisheries biologists and ecologists be involved in the selection and evaluation of metrics, the determining of expectations for each metric, and the assignment of threshold values associated with different classes of ecological condition.

Potential consequences of the proposed plot design and collection methods are identified and presented in Table 4-9. Adverse effects on fish populations are generally low, and the fish assemblage has sufficient time to recover between sampling visits. Methods for the indicator can be a component of an integrated sampling regime for several different indicators.

A critical aspect of obtaining a representative sample of the fish assemblage under the proposed plot design is determining the length of stream that must be sampled at each site. For this indicator, a sample of the assemblage must be collected from a single pass through a prescribed length of stream (Karr *et al.* 1986, McCormick and Hughes 1998). Repeated sampling of a stream reach is neither practical nor representative. Thus, it is imperative that the length of stream to be sampled maximizes the number of species collected. A small pilot study on a few selected streams was conducted to make this determination. Based on this study (Fig. 4-4), a stream length equal to 40 times the mean channel width was selected as the area to be sampled at each stream. This length of stream is sufficient to obtain approximately 90 percent of the fish species inhabiting the reach. Sampling additional lengths of a stream does not substantially increase the number of species obtained. Lyons (1992) reported similar results.

Table 4-9. Effects of sampling

Effects on stream fish assemblages and protected organisms, populations, habitats:

- Minimal impacts under normal conditions
- Most individuals collected released alive (some retained as voucher specimens)
- Some mortality due to electrofishing or seining, especially if large numbers are collected and processed
- Some impact due to physical disturbance of stream channel during seining, but scale of sampling is small relative to entire watershed
- May be minimal in some areas due to collection permit restrictions (no sampling)

Effects of a single visit on subsequent visits:

- Collecting replicate samples during a single visit not practical nor accurate
- Entire reach is disturbed during sampling
- Most individuals collected are released alive, but need time to recover and redistribute after collection
- Studies indicate that fish assemblages recover after natural disturbances such as floods or extended drought

Effects on concurrent measurement of other indicators:

- Methods are used in an integrated sampling regime for a variety of different indicators.
- No impact if proper sampling sequence is followed (*e.g.*, collect chemical samples prior to obtaining fish assemblage samples)

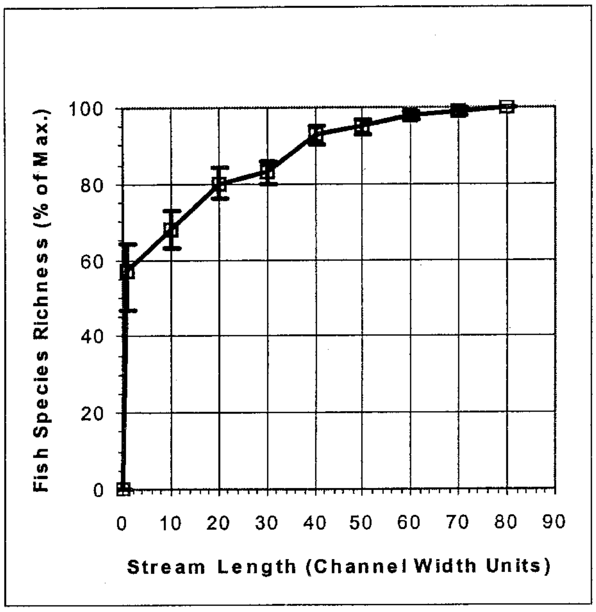


Figure 4-4. Effort-return curve of fish species richness versus length of stream sampled. (McCormick, unpublished data).

Summary

Standard sampling and data analysis methods for this indicator are shown to be compatible with the proposed plot design and monitoring framework, and are based on standard techniques for documenting stream fish assemblages. Possible constraints associated with the proposed plot design using the monitoring framework include site accessibility and consistency among crews. Equipment and technical expertise requirements are defined, and include the need for experienced fisheries biologists during initial phases of a monitoring program and for data analysis and interpretation. Field methods used for the indicator have little impact on either fish populations or habitat, and are compatible with an integrated sampling regime for multiple indicators. A pilot study was performed to determine the optimal stream length for sampling (*i.e.*, to assure that the stream length sampled is sufficient to obtain at least 90% of the fish species inhabiting the reach).

> ***Guideline 4: Logistics***
> *The logistical requirements of an indicator can be costly and time-consuming. These requirements must be evaluated to ensure the practicality of indicator implementation, and to plan for personnel, equipment, training, and other needs. A logistics plan should be prepared that identifies requirements, as appropriate, for field personnel and vehicles, training, travel, sampling instruments, sample transport, analytical equipment, and laboratory facilities and personnel. The length of time required to collect, analyze and report the data should be estimated and compared with the needs of the program.*

Performance Objective

1. Demonstrate the feasibility of data acquisition with respect to the proposed scale and intensity of monitoring in terms of staffing, training, travel, equipment, laboratory facilities, and data turnaround time.

There are several critical features of the proposed monitoring framework (Table 4-10) that must be considered in developing the logistics plan (*e.g.,* Baker and Merritt 1991) for data acquisition. These features require that a large number of sites be visited across a broad geographic area in a relatively brief time period each year. Table 4-10 also presents field logistics issues that must be addressed in the context of the constraints imposed by the monitoring program. Considerable effort is required to locate and obtain permission from landowners who must be contacted to access sampling sites. A considerable amount of lead time is also required to apply for and obtain all required scientific collecting permits. This is because of the number of different states included in the mid-Atlantic highlands region and the large number of sites that must be reviewed individually for presence of protected species. As mentioned previously (Table 4-7), restrictions may be placed on collecting at individual sites that harbor protected species.

Based on experience from the MAHA study, it is feasible to implement the indicator as part of an even larger-scale, long-term monitoring program under the proposed monitoring framework (Hughes 1993, Lazorchak *et al.* 1998). Several field crews are required to accomplish all sampling within the required time period. The location of sites in different states imposes certain constraints that must be considered in determining the best source of personnel; if State personnel are used, they may be restricted to travel within their home state. Use of State personnel has advantages including: 1) shortening the process for obtaining scientific permits, and 2) providing more familiarity with staging areas, access points, landowners, streams, and fishes in the region. A crew of 3 or 4 people can accomplish all collecting from a stream in less than 4 hours, and complete all processing activities within a single day. A crew of this size can also obtain data for other indicators during the same visit. During the MAHA study, crews obtained samples or data for 8 additional indicators during a single site visit (Lazorchak *et al.* 1998). Four to five streams a week can be visited by a single crew, allowing for one day of travel between sites that are not necessarily close together because of the random selection process. Because of the level of technical expertise required (Table 4-8), the use of volunteers is not recommended for the indicator unless they can be included on crews with other personnel having sufficient technical background and experience. Additional training beyond basic instruction in collection procedures includes safety training associated with electrofishing and a workshop on field identification of the regional fish fauna. If State personnel are used, training may be less intensive, as they will be more familiar with basic collecting procedures and may have more experience with the identification of fishes in the field.

Several issues related to equipment and supplies (Table 4-10) should be considered in selecting the proposed indicator for use in a monitoring program. For example, the random selection of sites with no regard for location or ease of access will result in a number of sites located in remote areas of the mid-Atlantic highlands region. Experience with the MAHA program revealed these sites were accessible only by 4-wheel drive vehicles, by foot, or by a combination of the two. Sources and availability of leased 4-wheel drive vehicles are usually limited in many areas, and a long lead time may be required to obtain appropriate vehicles. If State or Federal personnel are used, appropriate government vehicles may be available. Accessing sites by foot may affect the crew size (*i.e.*, additional people may be required to transport all the required equipment), and possibly even the choice of equipment based on its portability. The use of hazardous material (formalin, gasoline) requires knowledge of and compliance with all regulations related to personal protection and transport. Depending on the sampling scenario developed, this usage may involve additional training requirements and purchase of appropriate shipping and packaging materials.

Laboratory issues (Table 4-10) relate primarily to the selection of a qualified facility to confirm identifications of voucher specimens and provide long-term archival of vouchers. A key constraint is the length of time that may be required before results of confirmatory identifications are available. Confirmation may affect the time required to complete validation of the data (Guideline 5) and report results for the indicator value. Under most circumstances, the 9-month timeline specified in Table 4-10 should be achievable for the indicator.

Table 4-10. Logistical considerations

Monitoring Requirements

50 to 200 site visits per year; sites to be revisited every 4 years
Sites selected at random with no consideration for location (*e.g.*, public *vs.* private ownership) or ease of access (*e.g.*, bridge crossings *vs.* wilderness areas)
Sites located across mid-Atlantic highlands region (multiple States)
All site visits must be completed within summer baseflow (July-September)
Each stream visit is limited to a single day
Results (*e.g.*, indicator values) should be reported within 9 months of collection

Field Logistics Issues

Site Access

- Landowner permission required
- Scientific collection permits required
  - Federal Endangered Species Permits
  - State permits may include limits for listed Threatened/Endangered species

Field Crew

- 3-4 person crew
- Personnel sources
  - Partnerships with State/Federal Agencies
  - Contract
  - Volunteers not recommended

Time/Effort Requirements

- 1 site per day
- Total sampling time: 45 min. - 3 hrs
  - Processing time varies based on catch
- 4-5 sites sampled per week

Additional training requirements

- Electrofishing requires safety training, first aid, CPR
- Field identification of fishes

Equipment and Supplies

- 4-wheel drive vehicles may be necessary to access remote sites (unmaintained roads)
- Equipment may need to be transported by foot over rugged terrain
- Transport and handling of hazardous materials required (formalin, gasoline)

Laboratory Logistics Issues

Sources

- Partnerships with State/Federal Agencies
- Contracts with regional museum facilities

Time/Effort Requirements

- Identification of voucher specimens: up to several months
- Long-term archival of voucher specimens

Summary

Data collection activities for the indicator are described within the constraints imposed by the proposed monitoring framework. Considerable lead time is required to obtain permission to access sampling sites and to obtain the required scientific collecting permits. Partnerships with local agencies can streamline collection efforts. Various staffing options are presented to provide the required expertise; some safety training may be necessary for crews. Equipment and supplies may have to be transported over harsh terrain. Validation of fish identifications can be achieved by a competent local museum.

Guideline 5: Information Management
Management of information generated by an indicator, particularly in a long-term monitoring program, can become a substantial issue. Requirements should be identified for data processing, analysis, storage, and retrieval, and data documentation standards should be developed. Identified systems and standards must be compatible with those of the program for which the indicator is intended and should meet the interpretive needs of the program. Compatibility with other systems should also be considered, such as the internet, established federal standards, geographic information systems, and systems maintained by intended secondary data users.

Performance Objectives

1. Identify requirements for data processing, review, analysis, and storage, and demonstrate compatibility with those capabilities available to the proposed monitoring program.
2. Describe the metadata necessary for primary and secondary users to access the data, to reproduce the results, or to use the data in other types of analytical and interpretive activities.

There are important information management requirements and issues related to supporting the routine use of the indicator within the proposed monitoring framework (Table 4-11). Experience with the MAHA study has indicated that a fairly lengthy time period is required to complete review and validation of the measurement data prior to their use in computing metric responses, scores, and the indicator value. This process may inhibit the ability to achieve the desired reporting timeframe (9 months; Table 4-10). We anticipate this time will be reduced as experience with the data is gained and automated routines are developed to facilitate review and validation activities.

The requirements for hardware and software (Table 4-11) were selected to be compatible with nearly all potential participants in the proposed monitoring program. Some programming support may be needed to develop the routines for computing metric responses, scores, and indicator values from validated measurement data. Diaz-Ramos *et al.* (1996) provide statistical algorithms needed to compute resource population estimates in spreadsheet-compatible format.

The critical data sets and metadata required to support the development of the indicator and its component metrics (Table 4-11) are few in number and fairly straightforward. A critical component of archival activities for the indicator is the incorporation of voucher specimens into a permanent museum collection.

Table 4-11. Information management requirements

Time to Validate and Analyze Data

Six-twelve months: Will be reduced with experience in the region and development of automated check routines.

Hardware and Software Requirements

Hardware: High-end personal computer (PC)

- Capable of performing all calculations required for indicator development and evaluation and resource population estimates.
- Measurement data files (species ID and abundance data) can become large; sufficient storage capacity is required.

Software: Data management

- Relational database software useful, but not required
- Spreadsheets can be used to perform calculations and manage data files, but may be cumbersome

Statistical analysis software, graphics software

- SAS has been used to develop indicator in MAHA
- Spreadsheet capable of computing basic statistics, regressions
- Multivariate analyses can be beneficial in evaluating metrics and IBI, but not required

Critical Data Sets

Validated species ID and count data for each site visit

Autecological data for each species, including taxonomic information, habitat, tolerance class, feeding class and reproductive strategy

Individual metric values, scores, and IBI value for each site visit

File containing locational information for each site, site classification information (*e.g.*, ecoregion, drainage), and inclusion probability values to calculate resource population estimates

Ancillary databases to allow metric and indicator evaluation of response to disturbance (chemistry, physical habitat, watershed stressors, landscape-level data)

Metadata Requirements

Methods documentation

Sources of autecological information

Documentation for how individual metrics are computed from basic count and ID data, and how a final score is assigned for each metric

Database documentation regarding files and variables

Data and Sample Archival

Field data forms (paper or electronic)

Voucher specimens cataloged into permanent museum collection

Summary

Information management requirements for the indicator are currently time consuming (6-12 months to develop indicator values from the raw data), but time should be reduced with automated routines. The length of time currently required to validate measurement data may affect the desired turnaround time established for the proposed monitoring program. No specialized hardware, software, or programming support is required, and data storage is compatible with other systems for retrieval. Critical data sets and associated metadata are not extensive or complicated.

Guideline 6: Quality Assurance
For accurate interpretation of indicator results, it is necessary to understand their degree of validity. A quality assurance plan should outline the steps in collection and computation of data, and should identify the data quality objectives for each step. It is important that means and methods to audit the quality of each step are incorporated into the monitoring design. Standards of quality assurance for an indicator must meet those of the targeted monitoring program.

Performance Objective

1. Demonstrate that the critical components of an appropriate quality assurance program are established for the indicator, and that techniques are available to monitor and control important sources of error in the measurement data for the indicator.

The scale and time frame of the proposed monitoring framework and the need for multiple field crews (Table 4-10) require a rigorous quality assurance (QA) program to ensure consistency in data collection and interpretation of indicator values (*e.g.*, Chaloud and Peck 1994). There are important considerations (Table 4-12) for developing an appropriate QA program for EMAP-related studies. Resources are available, in the form of guidance documents and existing quality assurance plans, that can be adapted or modified to other types of monitoring efforts. No additional research is required to develop appropriate standards or other techniques to monitor and control data quality. All field and laboratory procedures associated with the indicator are amenable to the development of performance criteria and to internal or external audits by qualified personnel. Measurement related errors can be identified (Guideline 8) and compared against established performance criteria. The use of a qualified museum facility to confirm field identifications of voucher specimens and as a permanent repository provides a means to control and correct for a critical source of error. Examination of data from sites visited more than once during a single index period (Table 4-7) can be used to evaluate the consistency and performance of collection methods and field personnel. Concurrent identification of fish species in the field by a recognized authority in fish taxonomy can provide rapid identification and correction of errors. Finally, a variety of procedures are available to provide a rigorous review and validation of data to identify and correct for entry errors, erroneous species identification, and abundance values.

Summary

An appropriate quality assurance program can be developed and implemented for the indicator and monitoring framework using available resources and techniques. No additional research is required to provide appropriate performance standards or other techniques to monitor and control data quality. Measurement errors can be identified and evaluated at each critical step of indicator measurement.

Table 4-12. Quality assurance considerations

QA program guidance available:

- Environment Canada (1991)
- Draft guidance documents from U.S. EPA National Center for Environmental Research and Quality Assurance
- EMAP-Surface Waters integrated QA project plan (Chaloud and Peck 1994)

Controls and audits can be established for all field and laboratory protocols. Performance evaluations can be accomplished using repeat visits or by comparisons to results obtained by recognized experts.

Data review procedures available:

- Comparison of observed locations of species to known geographic range
- Exploratory analysis to identify outliers and suspicious values
- Internal consistency of counts

Guideline 7: Monetary Costs
Cost is often the limiting factor in considering to implement an indicator. Estimates of all implementation costs should be evaluated. Cost evaluation should incorporate economy of scale, since cost per indicator or cost per sample may be considerably reduced when data are collected for multiple indicators at a given site. Costs of a pilot study or any other indicator development needs should be included if appropriate.

Performance Objective

1. Provide information regarding costs associated with implementing the indicator within the proposed monitoring framework. Compare these costs, if possible, to similar costs associated with other indicators that could be implemented within the proposed monitoring framework.

There are a variety of costs associated with implementing data collection activities for the indicator under the proposed monitoring framework (Table 4-13). These costs are based on collection activities within the MAHA study, using private contract field crews. Also included are costs associated with permanent archival of voucher specimens. Equipment costs are presented on a per-crew basis, under the assumption that new equipment is required. Karr (1991) presents cost-related information that suggests that sampling fish assemblages may be more economical than other types of biological or chemical samples. Yoder and Rankin (1995) present costs required to implement a similar indicator within a statewide network of hand-selected monitoring sites in Ohio. They also show that fish assemblage data are less expensive to collect and analyze than quantitative macroinvertebrate samples, chemical samples, or various types of bioassays. Their costs are based on the capability for a small-sized crew (3) to sample 3 to 6 sites per day. Costs might increase when field logistics or accessibility are difficult (Guideline 4, Table 4-10), but under the proposed monitoring framework this is offset to some extent by visiting fewer sites.

Table 4-13. Cost information

Sampling costs (per site)= $1,540

Field Crews: $1200

- Includes salary, benefits, per diem, and lodging for 4 persons
- Based on 1 site per day collecting data for multiple indicators
- Does not include costs associated with vehicles
- Includes cost of acquisition of specimens for fish tissue, biomarkers, and genetics indicators

Laboratory Costs: $325

- Data analysis and data management
- Identification, verification, and archiving voucher specimens

Supplies: $15

- Jars, waterproof paper, formalin

Field equipment (per crew): $3,515

- Backpack electrofishing unit: $3,000 (one-time cost for multiple years' use; annual maintenance = $300)
- Dip nets and seines: $200
- Measuring board: $65
- Miscellaneous equipment: $250
- Estimate 15% annual maintenance and replacement cost

Summary

The greatest cost for this indicator is salary for a field crew. Per-site costs may depend on field logistics and accessibility of sites. Much of the cost for supplies and equipment can be distributed over several years (for the life of the equipment or duration of the monitoring program). Similarly, costs for field crews and site visits could be distributed to other indicator measurements made at the same sites. Examination of costs for this indicator, using estimates presented here and with similar information obtained from other published sources, suggests that data collection and analysis may be less expensive for this indicator than for other biological or chemical indicators that might be implemented within the proposed monitoring.

Phase 3: Response Variability

> ***Guideline 8: Estimation of Measurement Error***
> *The process of collecting, transporting, and analyzing ecological data generates errors that can obscure the discriminatory ability of an indicator. Variability introduced by human and instrument performance must be estimated and reported for all indicator measurements. Variability among field crews should also be estimated, if appropriate. If standard methods and equipment are employed, information on measurement error may be available in the literature. Regardless, this information should be derived or validated in dedicated testing or a pilot study.*

Performance Objective

1. Provide estimates of important measurement-related errors associated with the indicator, and compare them to established performance criteria for the proposed monitoring framework.

Several different types of errors can affect either the measurement data or the development of indicator values from measurement data (Table 4-14). Measurement-related errors of field collection data, in terms of number of species collected, species composition, and number of individuals cannot be estimated directly for the indicator by collecting replicate samples during a single visit to a site (Table 4-9, Fore *et al.* 1994). In terms of repeatability, other published studies may not be applicable to the entire mid-Atlantic highlands region or to the proposed monitoring framework. Measurement-related errors can be treated as a part of temporal indicator variability (Guideline 9) and can be indirectly evaluated as part of "within-year extraneous variance." The other critical source of error in measurement data is incorrect identifications of fish species. Various means of controlling this source of error have been presented previously, including the collection and confirmation of voucher specimens (Table 4-7), using personnel experienced in fish identification (Table 4-8) and additional training in field identification of regional fishes (Table 4-10).

Table 4-14. Potential sources of measurement error

Data Collection

Poor repeatability in number of species collected, species composition, and counts

- Cannot be estimated directly using replicate samples
- Controlled by standardized protocols and methods

Incorrect identification of species by field crews or data recording errors

- Performance objective is < 10% errors
- Controlled through training, field audits and performance checks, and confirmation of voucher specimens

See Table 4-9 for additional details

A more quantitative evaluation of errors related to identification of fish species by field crews was derived from 3 years of sampling for the MAHA study (Fig. 4-5). Five types of error are investigated. Transcription errors occur when the wrong species code is recorded on the field data form. The remaining four relate to actual errors in taxonomy. They include a cumulative estimate of errors for all species, errors specific to two groups of fishes that are difficult to identify to species in the field (sculpins, genus *Cottus*, and a cyprinid genus *Nocomis*), and errors at the genus level. Over the 3-year period, improvements were made to field data forms, crew training, and the procedure for collecting voucher specimens. Performance improved each year, with the virtual elimination of transcription errors, misidentification of sculpins, and misidentifications at the genus level (which are potentially more serious than species level identification in terms of the impact on metric responses). The remaining error levels for overall species misidentifications and identification of *Nocomis* species declined to well below the performance objective initially established (< 10%; Table 4-14).

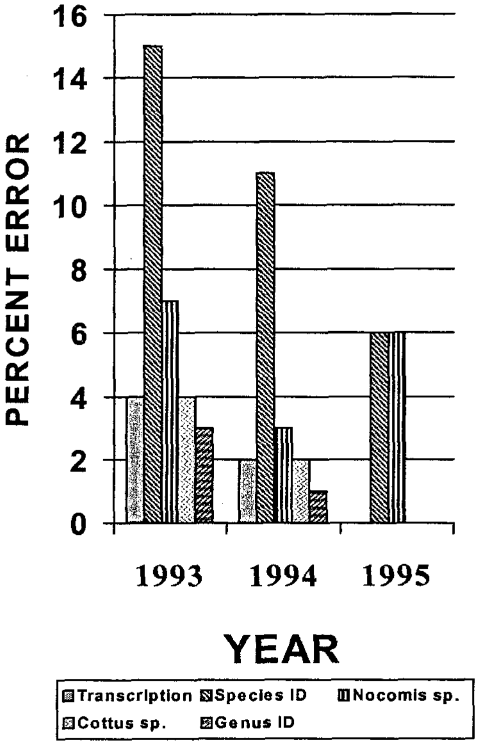


Figure 4-5. Comparison of different errors in the identification of fish species (McCormick, unpublished data).

Summary

Measurement error is difficult to assess because repeatability is not possible using single-sample data collection methods. However, relevant information can be obtained indirectly through the estimation of other variance components. Performance criteria established to control important sources of measurement error in the indicator can be achieved with the implementation of appropriate control measures. Control measures applied to field identifications of fishes resulted in a substantial reduction of errors to within the performance criteria initially established for the monitoring program.

Guideline 9: Temporal Variability - Within the Field Season
It is unlikely in a monitoring program that data can be collected simultaneously from a large number of sites. Instead, sampling may require several days, weeks, or months to complete, even though the data are ultimately to be consolidated into a single reporting period. Thus, within-field season variability should be estimated and evaluated. For some monitoring programs, indicators are applied only within a particular season, time of day, or other window of opportunity when their signals are determined to be strong, stable, and reliable, or when stressor influences are expected to be greatest. This optimal time frame, or index period, reduces temporal variability considered irrelevant to program objectives. The use of an index period should be defended and the variability within the index period should be estimated and evaluated.

Guideline 10: Temporal Variability - Across Years
Indicator responses may change over time, even when ecological condition remains relatively stable. Observed changes in this case may be attributable to weather, succession, population cycles or other natural inter-annual variations. Estimates of variability across years should be examined to ensure that the indicator reflects true trends in ecological condition for characteristics that are relevant to the assessment question. To determine inter-annual stability of an indicator, monitoring must proceed for several years at sites known to have remained in the same ecological condition.

Combined Performance Objectives for Guidelines 9 and 10

1. Identify important components of variance based on the proposed monitoring framework and sampling design.
2. Demonstrate that the magnitude of individual components is within performance criteria established for the proposed monitoring program.

The use of a probability-based survey design as the monitoring framework requires a modified approach to defining and estimating important components of an indicator's spatial and temporal variance. In the case of a multi-metric index, it is also necessary to determine variability in individual candidate metrics in order to select the final suite.

Important sources of variation for the indicator within the proposed monitoring framework and target performance criteria for EMAP have been identified (Table 4-15). Note the inclusion of "population variance," which is the variation due to the relationships between the probability sample and the survey design. It is solely a function of the number of probability-based samples used to estimate the resource population of interest. It has been determined that 50 samples provides population estimates with 90 percent confidence bounds that are approximately ±10 percent of the proportion (Larsen *et al.* 1995, Larsen 1997). The survey design is flexible in allowing one to define *a posteriori* various resource subpopulations of interest within the constraints of sample size.

Sources of temporal variation in the indicator value (or metric score variable) are included in "extraneous variance" (Table 4-15). These components and descriptions are based on Larsen *et al.* (1995) and Urquhardt *et al.* (1998), for indicators associated with monitoring frameworks similar to that of the proposed indicator. Within-year variability (Guideline 9) is estimated as index period variance. Variability across years (Guideline 10) for this indicator is addressed by two separate components of variance: the coherent variability of all sites across years, and the among-year variability of individual sites. Within-year variability also includes "measurement-related" errors described under Guideline 8. Because the proposed monitoring framework emphasizes regional scales and populations of sites, rather than patterns at individual sites, the importance of measurement-related error is reduced. Measurement-related errors are considered in detail only when within-year variability is unacceptably large relative to the total extraneous variance of the indicator. In such cases, it must be determined whether the variability is due primarily to temporal variability or to measurement-related errors.

Variance components were estimated using all 298 sites in the mid-Atlantic highlands region (including repeat visits within and across years), weighted by the appropriate population expansion factor. These expansion factors are used to extrapolate the results from each site in the survey sample to the entire resource population

Table 4-15. Principal variance components for proposed monitoring framework

Among-site variance: Variation due to differences in the indicator value among a sample of stream sites. This component represents the environmental "signal" to be detected and interpreted with respect to an ecological condition.

- Function of number of sites sampled (inclusion probability)
- Calculated based on routines in Diaz-Ramos *et al.* (1996)
- Performance Objective: Target subpopulation sample size of 50, with minimum of 30

Extraneous variance: Remaining temporal, spatial and measurement-related variation. Collectively, these components represent "noise" that inhibit the ability to detect and interpret the environmental "signal." Extraneous variance is characterized using a randomly-selected subset of the probability sample sites. These sites are revisited across and within years.

- Components of extraneous variance:
  - ***Coherent variance across years:*** Amount that all sites in a region vary in common due to regional-scale effects (*e.g.*, climate, hydrology); important component in ability to detect trends in regional population of sites
  - ***Among-year variance:*** Interaction of site and annual variability; amount an individual site varies among years
  - ***Within-year variance:*** Temporal variance at a site within the defined index period. Also contains measurement-related error, crew errors, *etc.* Important component in determining status of resource population

    Approach: If temporal variance is significantly less than total observed variance, then measurement-related components are not important. If temporal variance contributes substantially to total variance, then examine measurement-related components for possible sources.
  - ***Spatial variance:*** Variance among different ecological subregions (see Guideline 11)
- Calculated based on 2-factor analysis of variance model (sites, coherent variance across years), with an interaction term representing among-year variability at an individual site
- Performance Objectives:
  - Variance within the index period should be approximately 10% of total variance to minimize effect on status estimation and maximize discriminatory ability.
  - Variance between years must be minimal relative to total; target capability is to detect a 2% change per year in a regional population mean with a Type I[a] error of 0.1 and a Type II[b] error of 0.2.

[a] Type I error (false positive) is the probability of concluding that a trend is present when in truth it is not.

[b] Type II error (false negative) is the probability of concluding that a trend is absent when in truth it is present.

of interest, and are based on the probability that an individual site will be selected as part of the survey sample from the universe of potential target population sites. The relative magnitudes of different components of variability for the indicator and each candidate metric were identified (Fig. 4-6), following Urquhardt *et al.* (1998). With respect to estimating the status of resource populations using the proposed indicator, within-year (index period) variability should comprise 10 percent or less of the total extraneous variance (Larsen *et al.* 1995, Larsen 1997). The indicator itself achieved this target, with index variability contributing approximately 5% to the total extraneous variance. However, performance of individual metrics was mixed: Six metrics

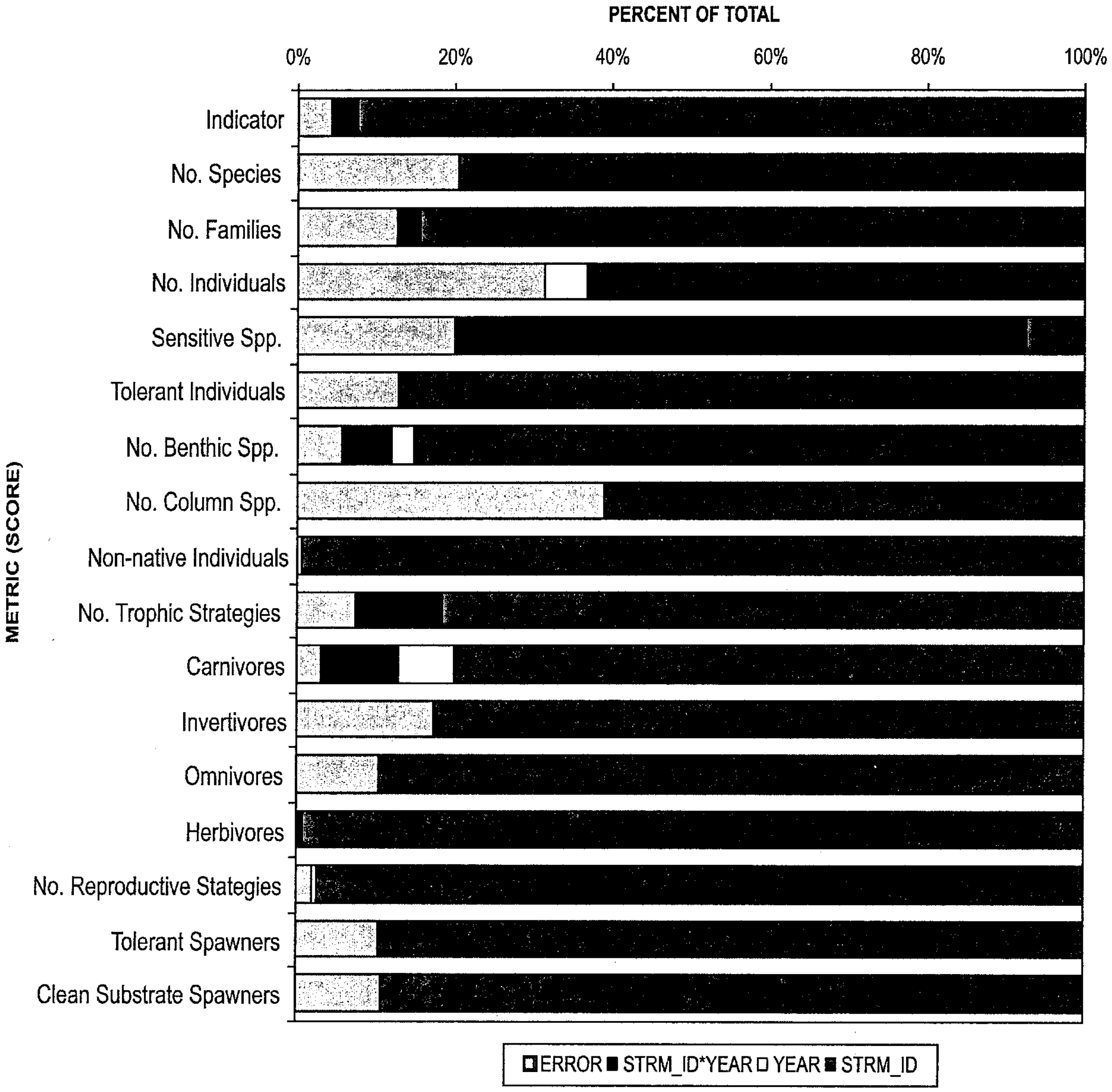


Figure 4-6. Relative contributions of important variance components for the proposed indicator and for candidate metric score variables. **ERROR** = within year (index) variability. **STRM_ID*YEAR** = among year variability of individual sites. **YEAR** = coherent variability of all sites across years. **STRM_ID** = among site variability.

were within 10%, five exceeded this only slightly, and five were 20-40% of the total extraneous variance. The latter group requires additional evaluation to determine if error variability can be reduced by modification of the scoring approach, alteration of the fish species in a particular group, or possibly altering the index period itself. Also, these estimates are based on only two years and a relatively small number of sites; more precise estimates will be possible with additional years of repeat sampling. Some of these metrics represent the most widely-used attributes for characterizing aquatic communities; deleting any of them for purely statistical reasons might diminish the utility of the indicator (Karr and Chu 1977).

For trend detection capability, the coherent variability across years component should not be large relative to the total extraneous variance. This cannot be completely evaluated at the present time because the coherent variation component is estimated from only two years of data, and thus probably underestimates the true coherent variation of all sites across years. Several years of data from repeat sampling are required to rigorously estimate this component. Trend detection capability is further evaluated as part of Guideline 13 (Data Quality Objectives).

Summary

Important components of variability, particularly within-year variability, were estimated for the indicator and candidate metric score variables. The indicator and most individual metrics achieved or nearly achieved the performance objective (contributing < 10% to the total extraneous variance). Five metrics were well above this and should be further evaluated. Performance of the indicator and candidate metrics with respect to trend detection cannot be evaluated at this time, as several years of data are required to provide rigorous estimates of coherent annual variability.

> ***Guideline 11: Spatial Variability***
> *Indicator responses to various environmental conditions must be consistent across the monitoring region if that region is treated as a single reporting unit. Locations within the reporting unit that are known to be in similar ecological condition should exhibit similar indicator results. If spatial variability occurs due to regional differences in physiography or habitat, it may be necessary to normalize the indicator across the region, or to divide the reporting area into more homogeneous units.*

Performance Objective

1. Demonstrate that indicator response will be consistent across the monitoring region of interest.

The geographic scale of the proposed monitoring framework is such that differences might be expected in species composition and potential richness, general structure of stream fish assemblages, and general abiotic characteristics of stream ecosystems. Aquatic ecoregions (Omernik 1987), along with consideration of zoogeographic factors affecting fish distribution patterns, can serve as a basis for determining if normalizing the indicator across the region of interest is necessary (Table 4-16). If major differences in the response variables associated with individual candidate metrics (*e.g.*, potential species richness, percent of carnivorous individuals) are observed among ecoregions (or aggregates of similar ecoregions), the indicator will require some type of normalization. Normalization can be attained by adjusting expectations (addressed under Guideline 14) for individual metrics within ecoregions as necessary. For example, the expectation for the percent of tolerant individuals may be 10% or less in one ecoregion, but be 20% or less in another because of the natural occurrence of more tolerant species. The final indicator value remains consistent with this approach, but its derivation is altered (*i.e.*, an indicator value of 60 means the same across the entire region of interest).

Table 4-16. *Use of aquatic ecoregions to evaluate regional consistency in interpretation of indicator*

Aquatic ecoregions (*e.g.*, Omernik 1987, Omernik and Griffith 1991, Omernik 1995) can serve as a regional framework to classify stream ecosystems in a target resource population

- Based on overall similarity in several natural features (*e.g.*, climate, soils, vegetation, physiography, land use).

Ecoregions correspond to spatial patterns in fish assemblages and abiotic characteristics of streams (*e.g.*, Pflieger 1975, Larsen *et al.* 1986, Rohm *et al.* 1987, Whittier *et al.* 1988, Hughes and Larsen 1988, Jenkins and Burkhead 1993).

Ecoregions have been shown to be useful in improving the consistency of interpretation of other multimetric indicators applied over large geographic scales (*e.g.,* Yoder and Rankin 1995, Barbour *et al.* 1996).

Ecoregions serve as a basis to account for natural differences in potential biotic integrity under minimal human disturbance.

- Can be used to define different expectations for individual metrics, or different thresholds for indicator value (*e.g.*, Yoder and Rankin 1995).
- Metric-based adjustment is more suitable for EMAP indicators because of focus on regional resource population estimates.

Two examples (Fig. 4-7) are provided to demonstrate an evaluation of differences in fish assemblage characteristics across the region of interest. The distributions of metric response variables across two levels of aquatic ecoregion aggregations are examined using box-and-whisker plots. Regions showing restricted or expanded distributions in comparison to others should be considered for possible adjustment in metric expectations. For both examples (Figure 4-7 (A), number of water column species; and (B) proportion of individuals of tolerant species), examination of the boxplots suggests that no substantial differences exist in the range or general distribution of response values across the two levels of ecoregion aggregation. For these two metrics, adjustments of expectations do not appear to be necessary. Similar analyses applied to other candidate metrics have provided similar results, and at present, the indicator is being developed without normalization of component metrics.

Summary

Aquatic ecoregions, evaluated in the context of historical zoogeography affecting fish distributions, can be used to assess the natural variation in metric responses. These results can be used to adjust the expectations for individual metrics. Preliminary examination suggests that normalization is not necessary for the component metrics or the indicator, but additional analyses, performed in conjunction with assessing the responsiveness of the indicator (Guideline 12), are needed.

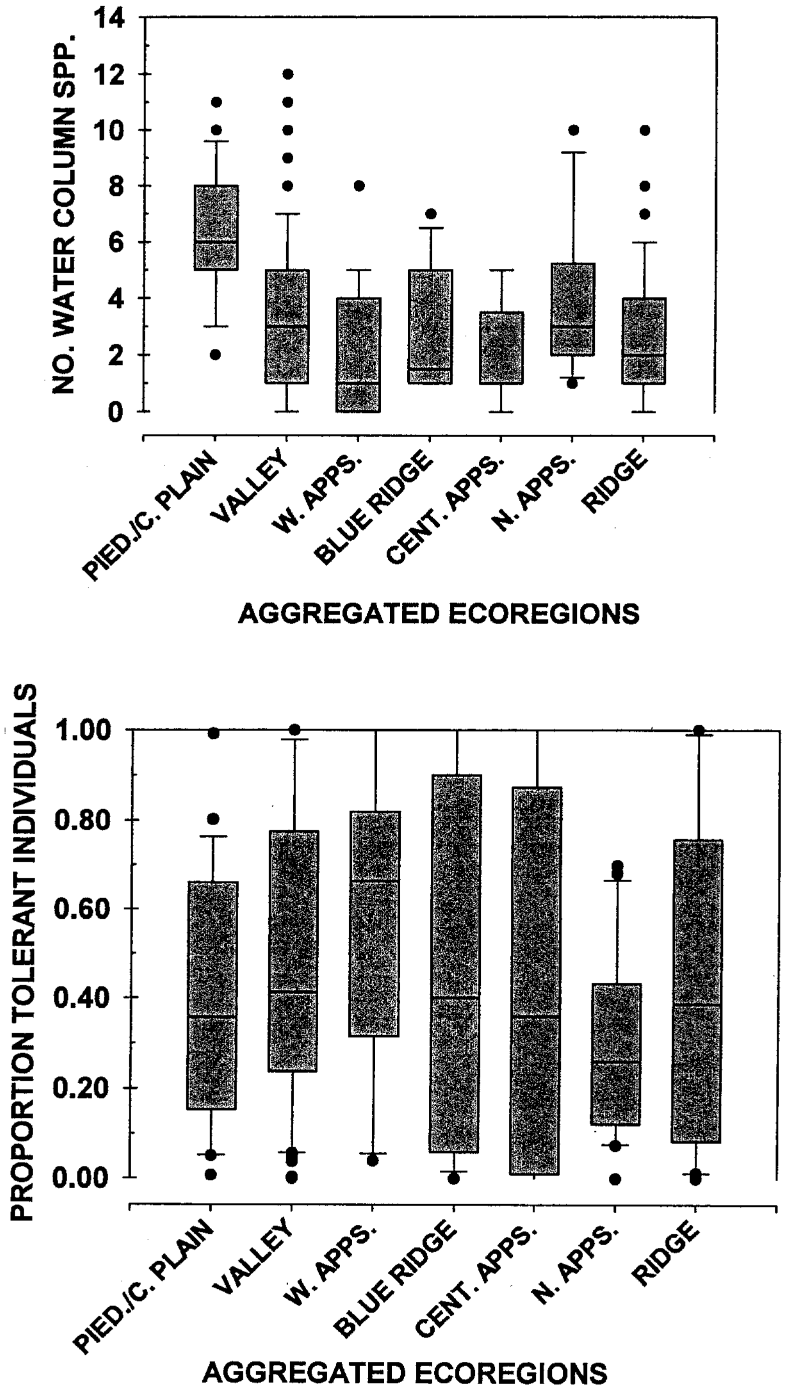


Figure 4-7. Examples of evaluating possible differences in metric expectations across MAHA subdivided by two levels of ecoregion aggregation. (A) number of water column species, a metric based on species richness; (B) proportion of individuals of tolerant species, a metric based on the proportion of individuals collected.

Guideline 12: Discriminatory Ability
The ability of the indicator to discriminate differences among sites along a known condition gradient should be critically examined. This analysis should incorporate all error components relevant to the program objectives, and separate extraneous variability to reveal the true environmental signal in the indicator data.

Performance Objective

1. Demonstrate responsiveness of the indicator and its component metrics to individual stressors or to the cumulative effects of multiple stressors.

Conceptual relationships between the indicator and its component metrics and various types of stressors have been addressed (Guideline 2). Other studies using similar multimetric indicators have demonstrated the potential responsiveness of the indicator (Table 4-17). For this indicator, a large number of sites, representing a range of stressor intensities, are used rather than an experimental-based approach using sites of known stress intensity. The proposed evaluation approach for this guideline is graphic (Fig. 4-8), rather than statistical (Fore *et al.* 1996, Karr and Chu 1997). Indicator values or individual metric scores are plotted against individual stressor variables, and/or against new variables derived from multivariate analyses of suites of stressor variables (*e.g.*, Hughes *et al.* 1998).

Table 4-17. Responsiveness of other multimetric fish assemblage indicators to stressors

- Karr *et al.* (1985): Chlorine
- Steedman (1988): Gradient of urban to forest land use
- Rankin (1995): Habitat quality in Ohio (Correlation coefficients between 0.45 and 0.7)
- Wang *et al.* (1997): Land use in Wisconsin
- Hughes *et al.* (1998): Intensity of human disturbance

Summary

Individual metrics respond predictably to specific stressors, though in some cases those specific responses are weak. The individual metrics and the indicator exhibit the predicted responsiveness to a multivariate "disturbance" variable derived from several individual chemical, habitat, and watershed stressor variables. Individual metric responses to specific stressor variables, as well as expectations for scoring metrics, should be examined to determine if responsiveness of the indicator to suites of stressor variables can be improved.

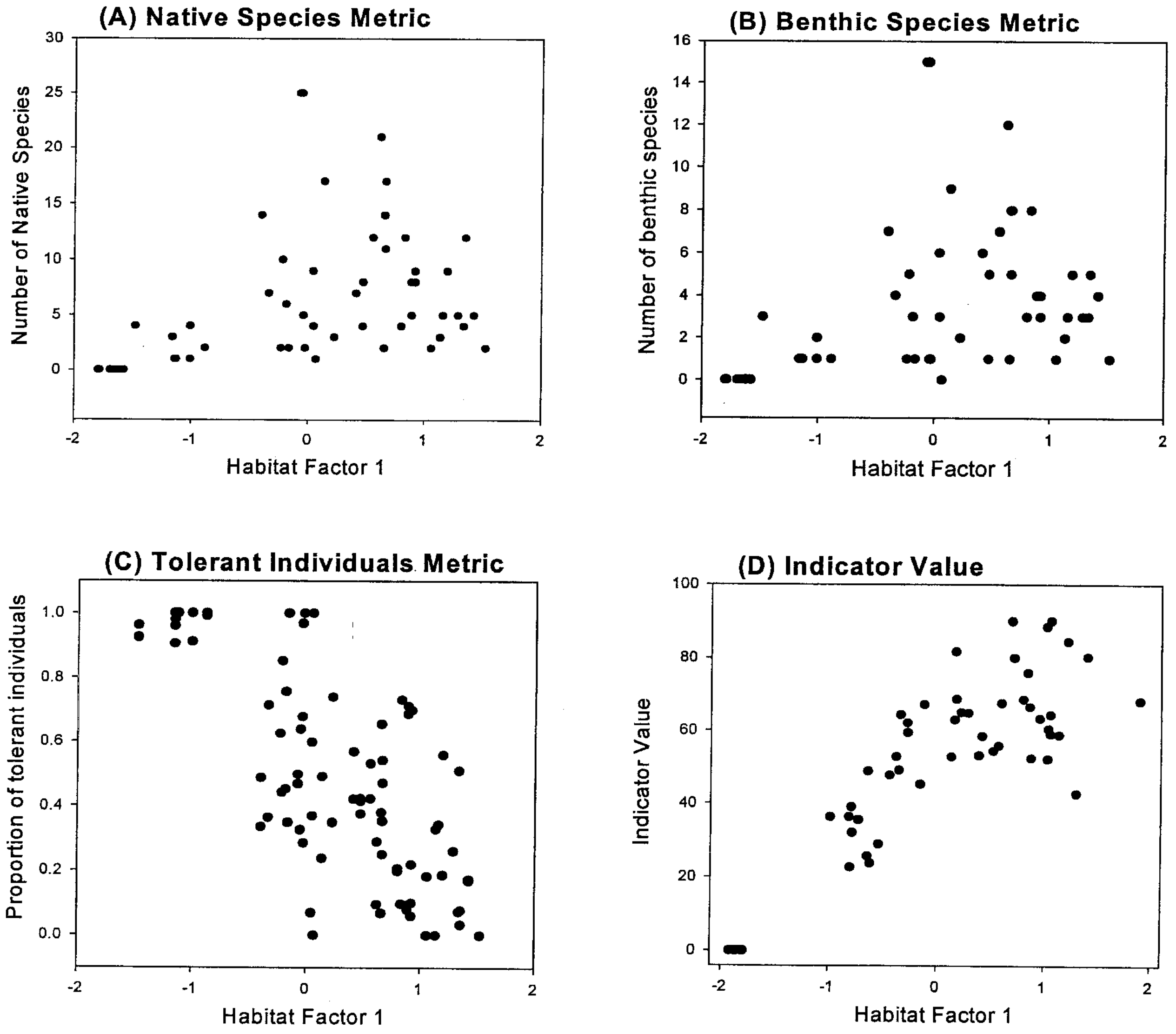


Figure 4-8. Examples of metric variables and indicator values plotted against a multivariate "disturbance intensity" factor derived from individual chemistry, habitat, and watershed stressor variables.

Phase 4. Interpretation and Utility

Guideline 13: Data Quality Objectives
The discriminatory ability of the indicator should be evaluated against program data quality objectives and constraints. It should be demonstrated how sample size, monitoring duration, and other variables affect the precision and confidence levels of reported results, and how these variables may be optimized to attain stated program goals. For example, a program may require that an indicator be able to detect a twenty percent change in some aspect of ecological condition over a ten-year period, with ninety-five percent confidence. With magnitude, duration, and confidence level constrained, sample size and extraneous variability must be optimized in order to meet the program's data quality objectives. Statistical power curves are recommended to explore the effects of different optimization strategies on indicator performance.

Performance Objectives

1. Demonstrate the capability of the indicator to distinguish classes of ecological condition within the proposed monitoring framework.
2. Demonstrate the capability of the indicator to detect trend in condition change within the proposed monitoring framework.

The capacity to estimate status and detect trend in condition is primarily a function of variability. Variability is due in part to natural differences that occur across a set of sampling sites (Guidelines 8 through 11), and also to differences in the intensity of human disturbance across those sites (Guideline 12). An indicator can have low variability (and thus high statistical power), but poor discriminatory capability because it cannot discern differences in intensities of human disturbance. However, high variability serves to reduce the discriminatory capability of an indicator.

Specific performance criteria for the indicator to detect trend in ecological condition have been developed for the proposed monitoring framework (Table 4-18). These criteria were examined using several power curves for the indicator to evaluate the effects of coherent variation across years, magnitude of trend, and sample size (Fig. 4-9). These curves were developed using the initial variance component estimates from the 1993-1994 MAHA study (Guidelines 10 and 11) and the approach described by Larsen *et al.* (1995) and Urquhardt *et al.* (1998). Derived estimates of the coherent variation across years were not used because they are based on only two years of data. Instead, to provide a range of possible scenarios, values of coherent variation (S^2_{year}) were substituted to range from 0-100, where 100 is approximately 1.7 times the within-year variance ($S^2_{residual}$). Four different magnitudes of trend were also evaluated, ranging from 0.5 to 2 indicator points per year (equal to 0.5 - 4% per year for an indicator score of 50 points). This represents a potential trend in the indicator score of 5 to 20 points over a 10-year period.

With respect to estimating status, the indicator satisfies the performance criterion (Table 4-18) under the conditions specified in Figure 4-9 (A). After 4 years of monitoring, the standard error of the indicator score ranges between 1 and 2 points (depending on sample size), which would provide 95% confidence intervals of about ±2 to ±4 points (which is less than 10% of the proposed impairment threshold of 50 points [Table 4-18]). Intervals computed for $\alpha = 0.1$(90% confidence intervals) would be smaller. With continued monitoring, the standard error of the estimate is stable through time.

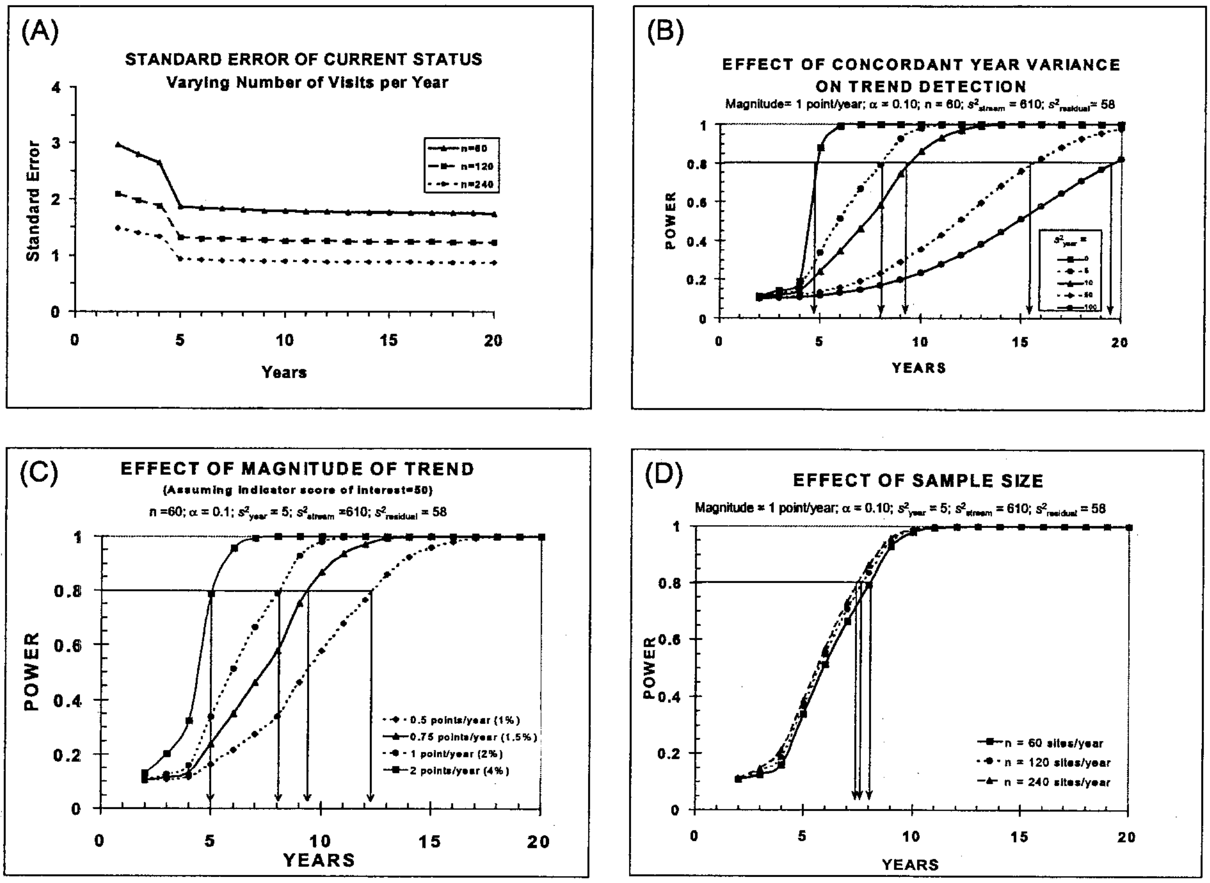


Figure 4-9. Statistical power curves for indicator. (A) Effect of annual sample size on standard estimate of indicator score; (B) Effect of the magnitude of coherent across-year variance (indicator score units) on trend detection; (C) Capability to detect different magnitudes of trend; (D) Effect of annual sample size on trend detection.

Figure 4-9(B) illustrates that, under the conditions specified, it would take between 5 and 20 years (equal to 1 to 5 sampling cycles) for the indicator to detect the specified magnitude of change (2% per year if the regional median = 50 points) with a power of 0.8, given a coherent variance across years of between 0 and 100 points^2 of indicator score. Figure 4-9(C) shows that, under the specified conditions, it would take between 5 and 13 years to detect various changes in indicator scores, representing 1 to 4% change per year in a regional population median score of 50 points. Figure 4-9(D) shows that, under the specified conditions, it would take between 7 and 9 years to detect the specified trend, depending on the number of sites visited per year. This series of figures points out that the capability of the indicator to detect trend is affected most by the magnitude of coherent across-year variance and by the desired magnitude of change that the monitoring program is expected to detect, and is affected to a lesser degree by the sample size. These analyses need to be repeated once a more robust estimate of coherent variance across years is obtained from several years' worth of data to determine which of the scenarios presented in Figure 4-9 is the most realistic. If the coherent across-year variance in the indicator score is relatively small (< 10 points^2) the indicator should meet the performance criteria for both status and trend established for EMAP-related monitoring frameworks.

Table 4-18. Statistical power capabilities

Power to discriminate among classes of ecological condition

Proposed monitoring framework: a minimum of 3, and preferably 4 classes of impairment in condition are desired:

- Fore *et al.* (1996): Analysis of similar multimetric indicator suggests 5-6 classes of condition can be distinguished at $\alpha = 0.05$ and $\beta = 0.2$
- Similar approach, using sites with repeat visits and/or resampling methods such as bootstrap procedures, is potentially feasible with indicator

Performance criteria for proposed monitoring framework:

- 90% confidence interval should be < 10% of the estimated proportion of a resource that is at or below a threshold value designating impairment

Power to detect trend in condition

Performance criteria for proposed monitoring framework:

- Magnitude of trend: 2% per year change in regional population median indicator score (= 20% change over a 10-year period)
- Sample size=50 to 200 sites monitored in region per year
- Probability of false positive (α) = 0.1
- Probability of false negative (β) = 0.2 (Power = 0.8)

Summary

Results from a previous study imply that 3 or 4 classes of condition can be distinguished over the potential range of indicator scores. Preliminary analyses indicate performance criteria for both status and trend detection can be met if across-year variance is relatively small compared to within-year variance. These analyses must be repeated after more robust estimates of coherent variability among sites are obtained from several years of data collection, and after the responsiveness of the indicator (Guideline 12) has been adequately established. Power curves were used to demonstrate the effects of alternative monitoring requirements, especially the importance of coherent across-year variance and desired magnitude of change.

Guideline 14: Assessment Thresholds
To facilitate interpretation of indicator results by the user community, threshold values or ranges of values should be proposed that delineate acceptable from unacceptable ecological condition. Justification can be based on documented thresholds, regulatory criteria, historical records, experimental studies, or observed responses at reference sites along a condition gradient. Thresholds may also include safety margins or risk considerations. Regardless, the basis for threshold selection must be documented.

Performance Objectives

1. Present and justify approach used to describe expected conditions under a regime of minimal human disturbance.
2. Present and justify proposed threshold values for the indicator to distinguish among classes of ecological condition.

The approach to scoring individual metrics is based on comparison of an observed metric response at a sampling site to the response expected under conditions of minimal human disturbance (see Table 4-6). Expectations for individual metrics (Table 4-19) that are based on measures of species richness are derived from a large number of sample sites from the MAHA study, as opposed to using a set of representative "reference" sites believed to be minimally impacted by human activities. For metrics based on the percentage of individuals, expectations are based primarily on values developed for similar indicators in other areas (*e.g.*, Karr 1986, Yoder and Rankin 1995).

Initial threshold values of the final indicator score have been proposed to classify different states of ecological condition (Table 4-20). Four classes of condition are proposed, based in part on the examination of the distribution of values within resource populations of the 1993-1994 MAHA study. Impaired condition was operationally defined as any score less than 50, which represents a level of biotic integrity less than one-half of that score expected under minimal human disturbance. This number of classes is consistent with the potential power of the indicator to distinguish differences in condition (Table 4-18). These thresholds are also somewhat consistent with those proposed by other groups using similar multimetric indicators (*e.g.*, Fore *et al.* 1996).

These threshold values have not been quantitatively examined, a process that requires a better understanding of indicator responsiveness (Guideline 12). Independent confirmation of appropriate threshold values is also necessary to achieve the performance objectives established for this guideline and implement this indicator in the proposed monitoring framework. Confirmation can be achieved by applying the indicator to an independent set of sites of known levels of impairment. Peer review of the proposed thresholds by professional ecologists and resource managers familiar with the development and interpretation of multimetric indicators is also required to complete the evaluation of the indicator with respect to this guideline.

Table 4-19. Thresholds defining expectations of indicator and metrics under minimal human disturbance

Expected conditions based on large number of sample sites, as opposed to a set of defined "reference" sites (Simon and Lyons 1995).

Expectations for metrics based on number of species calibrated for stream size or type (watershed area, gradient, cold *vs.* warm water) (Fausch *et al.* 1984).

Taxonomic composition and abundance metrics
- Number of native species: Varies with watershed area
- Number of native families: Varies with watershed area
- Total Abundance: ≥500 individuals collected in standard effort sample

Indicator species metrics:
- Percent of non-native individuals: 0%
- Sensitive spp. richness: Varies with watershed area
- Percent tolerant individuals:≤20%

Habitat metrics
- Number of benthic species: Varies with watershed area
- Number of water column species: Varies with watershed area

Trophic metrics
- Number of trophic strategies: 1 to 5 (varies with watershed area)
- Percent individuals as carnivores: ≥5%
- Percent individuals as invertivores: ≥50%
- Percent individuals as omnivores: ≤20%
- Percent individuals as herbivores: ≤10%

Reproductive guild metrics
- Number of reproductive strategies: 1 to 4 (varies with watershed area)
- Percent individuals as tolerant spawners: ≤ 20%

Table 4-20. Threshold values for classifying condition

Range of indicator values = 1 to 100
Excellent: > 85
Acceptable: 70 to 85
Marginal: 50 to 69.9
Impaired: < 50

Summary

The approach to defining expected conditions for individual metrics under a regime of minimal human disturbance is presented, and is based on standard documented approaches established for other multimetric indicators. Thresholds for the final indicator score are proposed for four classes of ecological condition. These thresholds are consistent with the potential capability of the indicator to distinguish among condition states, and with schemes developed for similar multimetric indicators. Additional research on the expectation for individual metrics remains, subsequent to achieving a better understanding of indicator responsiveness. These threshold values should be confirmed, either empirically through application to sites representing a known range of impairment, and/or through peer review by professional ecologists and resource managers.

Guideline 15: Linkage to Management Action
Ultimately, an indicator is useful only if it can provide information to support a management decision or to quantify the success of past decisions. Policy makers and resource managers must be able to recognize the implications of indicator results for stewardship, regulation, or research. An indicator with practical application should display one or more of the following characteristics: responsiveness to a specific stressor, linkage to policy indicators, utility in cost-benefit assessments, limitations and boundaries of application, and public understanding and acceptance. Detailed consideration of an indicator's management utility may lead to a re-examination of its conceptual relevance and to a refinement of the original assessment question.

Performance Objective

1. Demonstrate how indicator values are to be interpreted and used to make management decisions related to relative condition or risk.

Data derived from this indicator have not been assembled for management use, but EMAP has advanced an approach *(e.g.*, Paulsen *et al.* 1991, U.S. EPA 1997) to present information regarding the status of resource populations with respect to ecological condition (Fig. 4-10). Procedures are available (Diaz-Ramos *et al.* 1996) for developing cumulative distribution functions (cdfs) that show the proportion of a target resource population (estimated as lengths of target stream resource) that is at or below any specific value of the indicator (*e.g.*, a threshold value for impaired condition). Additional information regarding uncertainty is presented by computing confidence bounds about the cdf curve (*e.g.*, Diaz-Ramos *et al.* 1996, Stewart-Oaten 1996). In the example (Fig. 4-10), a threshold value of 50 (see Guideline 14, Table 4-20) is used to distinguish impaired condition. Approximately 30 percent (with 95 percent confidence bounds of approximately ±8 percent) of the target resource population has indicator values at or below the threshold value.

Information regarding relative risks from different stressors can be obtained using a similar approach (*i.e.*, developing cdf curves and evaluating the proportion of the target resource population that is at or below some threshold of impairment). Figure 4-11 presents an example showing the relative ranking of different stressors, based on the 1993-1994 MAHA study. Introduced fish species (based on presence) and watershed-level disturbances are the most regionally extensive stressors in the MAHA region, whereas acidic deposition, a larger-scale stressor, has a much lower impact across the region than might be expected. Once a suitably responsive indicator has been developed, association or contingency analysis of indicator values (or condition classes) and regionally important stressor variables (or impact classes) can be used to identify potential sources of impairment in condition. These analyses have not yet been conducted for the indicator, pending further research to improve the responsiveness of the indicator.

Summary

Approaches developed for EMAP can be used to graphically present results relating the distribution of indicator values (and corresponding condition classes) across a target resource population. Relative impact of various stressors on resource populations can also be determined and presented graphically. The combination of these two tools allows for the estimation of the status of resource populations with respect to ecological condition, and provides some indication of potential causes of impaired condition. Results from this indicator of biotic integrity can be used in the development of resource policy.

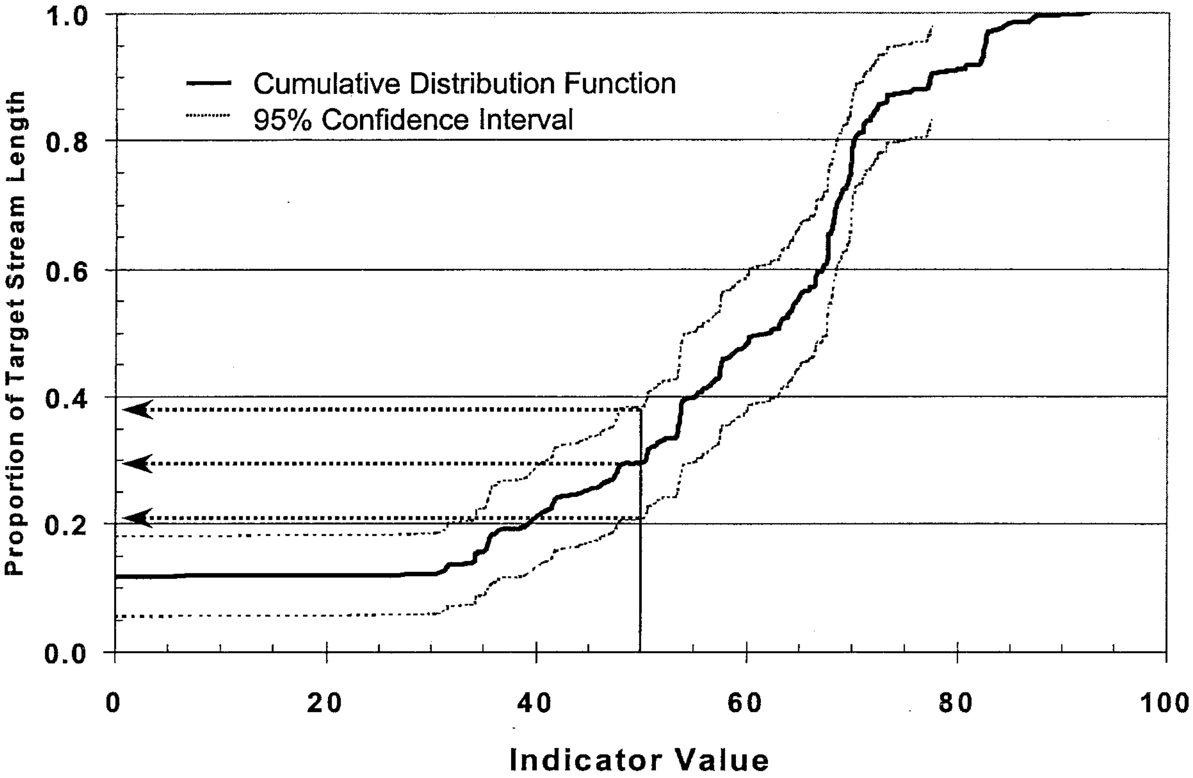


Figure 4-10. Hypothetical example showing how results from indicator values and monitoring framework will be used to estimate status of resource population.

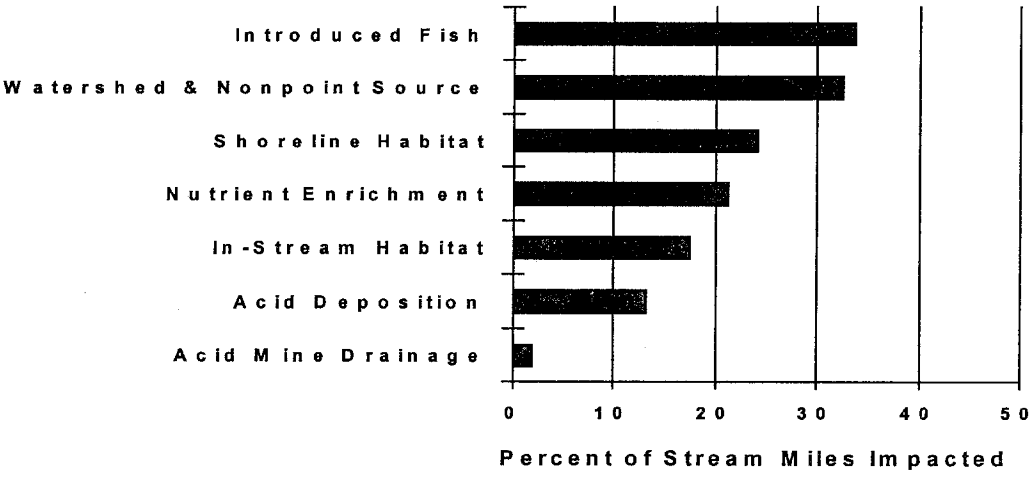


Figure 4-11. Relative ranking of stressor variables, based on the proportion of target resource population impacted (Source: 1993-1994 MAHA study).

References

Angermeier, P.A. and J.R. Karr. 1986. Applying an index of biotic integrity based on stream fish communities: considerations in sampling and interpretation. *North American Journal of Fisheries Management* 6:418-429.

Baker, J.R and G.D. Merritt. 1991. Guidelines for preparing logistics plans. EPA 600/4-91/001. U.S. Environmental Protection Agency, Office of Research and Development, Las Vegas, Nevada.

Barbour, M.T., S.B. Stribling, and J.R. Karr. 1995. Multimetric approach for establishing biocriteria and measuring biological condition. Pp. 63-77 *In*: W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Barbour, M.T., J. Gerrittsen, 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 macro-invertebrates. *Journal of the North American Benthological Society* 15(2):185-211.

Chaloud, D.J. and D.V. Peck (eds.). 1994. Environmental Monitoring and Assessment Program: Integrated quality assurance project plan for the Surface Waters Resource Group. EPA/600/X-91/080, Revision 2.00. U.S. Environmental Protection Agency, Las Vegas, Nevada

DeShon, J.E. 1995. Development and application of the invertebrate community index (ICI). Pages 217-243. In: W.S. Davis and T.P. Simon (eds.). Biological assessment and criteria: Tools for water resource planning and decision making. Lewis Publishers, Boca Raton, Florida.

Diaz-Ramos, S., D.L. Stevens, Jr., and A.R. Olsen. 1996. EMAP Statistical Methods Manual. EPA 620/R-96/002. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

Environment Canada. 1991. Quality Assurance Guidelines for Biology in Aquatic Environment Protection. National Water Research Institute, Burlington, Ontario, Canada.

Fausch, K.D., J.R. Karr and P.R. Yant. 1984. Regional application of an index of biotic integrity based on stream fish communities. *Transactions of the American Fisheries Society* 113: 39-55.

Fausch, K.D., J. Lyons, J.R. Karr and P.L. Angermeier. 1990. Fish communities as indicators of environmental degradation. Pages 123-144. In S.M. Adams (ed.). Biological indicators of stress in fish. American Fisheries Society Symposium 8. Bethesda, MD.

Fore, L.S., J.R. Karr and L.L. Conquest. 1994. Statistical properties of an index of biological integrity used to evaluate water resources. *Canadian Journal of Fisheries and Aquatic Sciences* 51:1077-1087.

Fore, L.S., J.R. Karr, and R.W. Wisseman. 1996. Assessing invertebrate responses to human activities, evaluating alternative approaches. *Journal of the North American Benthological Society* 15(2):212-231.

Gammon, J.R., 1976. The fish populations of the middle 340km of the Wabash River. Purdue University Water Research Center Technical Report 86. Lafayette, IN.

Gibson, G.R. (Editor). 1994. Biological Criteria: Technical Guidance for Streams and Small Rivers. EPA 822/B-94/001. U.S. Environmental Protection Agency, Office of Science and Technology, Washington, DC.

Hoefs, N.J. and T.P. Boyle. 1992 Contribution of fish community metrics to the index of biotic integrity in two Ozark rivers. Pages 283-303 *In*: D.H. McKenzie, D.E. Hyatt, and V.J. MacDonald (eds.), Ecological Indicators, Volume 1. Elsevier Applied Science, New York.

Hughes, R.M. and D.P. Larsen. 1988. Ecoregions: an approach to surface water protection. *Journal of the Water Pollution Control Federation* 60:486-493.

Hughes, R.M. 1993. Stream Indicator and Design Workshop. EPA/R-93/138. U.S. Environmental Protection Agency, Corvallis, Oregon.

Hughes, R.M, D.P. Larsen, and S.G. Paulsen. 1994. A strategy for developing and selecting biological condition indicators for EMAP-Surface Waters. Unpublished draft report, U.S. Environmental Protection Agency, Corvallis, Oregon.

Hughes, R.M. 1995. Defining acceptable biological status by comparing with reference conditions. Pages 31-48 *in* W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Hughes, R.M, P.R. Kaufmann, A.T. Herlihy, T.M. Kincaid, L. Reynolds, and D.P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. *Canadian Journal of Fisheries and Aquatic Sciences* 55: 1618-1631.

Hughes, R.M., and T. Oberdorff. 1999. Applications of IBI concepts and metrics to waters outside the United States and Canada. Pages 79-93 *in* T.P. Simon, editor. Assessing the sustainability and biological integrity of water resources using fish communities. Lewis Press, Boca Raton, FL.

Jenkins, R.E., and N.M. Burkhead. 1993. Freshwater Fishes of Virginia. American Fisheries Society. Bethesda, MD.

Jordan, S.J., J. Carmichael and B. Richardson. 1993. Habitat measurements and index of biotic integrity based on fish sampling in northern Chesapeake Bay. *In*: G.R. Gibson, Jr., S. Jackson, C. Faulkner, B. McGee and S. Glomb (eds.). Proceedings: Estuarine and Near Coastal Bioassessment and Biocriteria Workshop, November 18-19, 1992, Annapolis, MD. U.S. Environmental Protection Agency, Office of Water, Washington, D.C.

Karr, J.R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6:21-27.

Karr, J.R., K.D. Fausch, P.L. Angermeier, P.R. Yant, and I.J. Schlosser. 1986. Assessing biological integrity in running waters: a method and its rationale. Illinois Natural History Survey Special Publication 5. Champaign, IL.Karr, J.R. 1991. Biological integrity, a long neglected aspect of water resource management. *Ecological Applications* 1:66-84.

Karr, J.R., and D.R. Dudley. 1981. Ecological perspective on water quality goals. *Environmental Management* 5:55-68.

Karr, J.R., R.C. Heidinger, and E.H. Helmer. 1985. Sensitivity of the index of biotic integrity to changes in chlorine and ammonia levels from wastewater treatment facilities. *Journal of the Water Pollution Control Federation* 57:912-915.

Karr, J.R. and E.W. Chu. 1997. Biological Monitoring and Assessment: Using Multimetric Indexes Effectively. EPA 235/R97/001. University of Washington, Seattle.

Kerans, B.L., and J.R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of theTennessee Valley. *Ecological Applications* 4: 768-785.

Klemm, D.J., Q.J. Stober, and J.M. Lazorchak. 1993. Fish Field and Laboratory Methods for Evaluating the Biological Integrity of Surface Waters. EPA 600/R-92-111. U.S. Environmental Protection Agency, Office of research and Development, Cincinnati, Ohio.

Larsen, D.P., J.M. Omernik, R.M. Hughes, C.M. Rohm, T.R. Whittier, A.J. Kinney, A.L. Gallant, and D.R. Dudley. 1986. Correspondence between spatial patterns in fish assemblages in Ohio streams and aquatic ecoregions. *Environmental Management* 10:815-828.

Larsen, D.P. 1995. The role of ecological sample surveys in the implementation of biocriteria. Pages 287-300 *In*: W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Larsen, D.P. 1997. Sample survey design issues for bioassessment of inland aquatic ecosystems. *Human and Ecological Risk Assessment* 3(6):979-991.

Larsen, D.P. N.S. Urquhart, and D.L. Kugler. 1995. Regional scale trend monitoring of indicators of trophic condition in lakes. *Water Resources Bulletin* 31(1):117-139.

Lazorchak, J.M., D.J. Klemm, and D.V. Peck (eds.). 1998. Environmental Monitoring and Assessment Program-Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams. U.S. Environmental Protection Agency, Cincinnati, Ohio.

Lenat, D.R. 1993. A biotic index for the southeastern United States: derivation and list of tolerance values with criteria for assigning water-quality ratings. *Journal of the North American Benthological Society* 12:279-290.

Leonard, P.M., and D.J. Orth. 1986. Application and testing of an index of biotic integrity in small, coolwater streams. *Transactions of the American Fisheries Society* 115:401-414.

Lyons, J., 1992. Using the index of biotic integrity (IBI) to measure environmental quality in warmwater streams of Wisconsin. Gen. Tech. Rep. NC-149, U.S. Forest Service, North Central Forest Experiment Station, St. Paul, MN.

Lyons, J., Navarro-Perez, S, P.A.. Cochran, E. Santana C., and M. Guzman-Arroyo. 1995. Index of biotic integrity based on fish assemblages for the conservation of streams and rivers in West-Central Mexico. *Conservation Biology* 9(3):569-584.

Lyons, J., L. Wang and T.D. Simonson. 1996. Development and validation of an index of biotic integrity for coldwater streams in Wisconsin. *North American Journal of Fisheries Management* 16: 241-256.

McCormick, F.H. and R.M. Hughes. 1998. Aquatic Vertebrate Indicator. *In*: Klemm, D.J., J.M. Lazorchak, and D.V. Peck (eds.). Environmental Monitoring and Assessment Program-Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams. U.S. Environmental Protection Agency, Cincinnati, Ohio.

Meador, M.R., T.F. Cuffney, and M.E. Gurtz. 1993. Methods for Sampling Fish Communities as Part of the National Water-Quality Assessment Program. U.S. Geological Survey Open-File Report 93-104

Miller, D.L., P.M. Leonard, R.M. Hughes, J.R. Karr, P.B. Moyle, L.H. Schrader, B.A. Thompson, R.A. Daniels, K.D. Fausch, G.A. Fitzhugh, J.R. Gammon, D.B. Halliwell, P.L. Angermeier, and D.M. Orth. 1988. Regional applications of an index of biotic integrity for use in water resource management. *Fisheries* 13(5):12-20.

Oberdorf, T., and R.M. Hughes . 1992. Modification of an index of biotic integrity based on fish assemblages to characterize rivers of the Seine Basin, France. *Hydrobiologia* 228: 117-130.

Omernik, J.M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77:118-125.

Omernik, J.M. and G. E. Griffith. 1991. Ecological regions versus hydrologic units: Frameworks for managing water quality. *Journal of Soil and Water Conservation* 46:334-340.

Omernik, J.M. 1995. Ecoregions, a spatial framework for environmental management. Pages 49-62 *In*: W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Pflieger, W.F. 1975. The Fishes of Missouri. Missouri Department of Conservation. Jefferson City, MO.

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

Rankin, E.T. 1995. Habitat indices in water resource quality assessments. Pages 181-208 *In*: W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Rohm, C.M., J.W. Giese, and C.C. Bennett. 1987. Evaluation of an Aquatic ecoregion classification of streams in Arkansas. *Journal of Freshwater Ecology* 4:127-139.

Simon, T.P. 1991. Development of ecoregion expectations for the index of biotic integrity. I. Central Corn Belt Plain. EPA 905/9-90-005. U.S Environmental Protections Agency, Region V Environmental Sciences Division, Chicago, Illinois.

Simon, T.P. and J. Lyons. 1995. Application of the index of biotic integrity to evaluate water resources integrity in freshwater ecosystems. Pages 245-262 *in* W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.

Steedman, R.J. 1988. Modification and assessment of an index of biotic integrity to quantify stream quality in southern Ontario. *Canadian Journal of Fisheries and Aquatic Sciences* 45: 492-501.

Stewart-Oaten, A. 1996. Goals in Environmental Monitoring. Pages 17-28 *in* R.J. Schmitt and C.W. Osenberg (eds.) Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats. Academic Press, San Diego, CA. 399 pp.

U.S. EPA. 1997. Environmental Monitoring and Assessment Program (EMAP): Research Plan 1997. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

Urquhart, N.S., S.G. Paulsen, and D.P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. *Ecological Applications* 8(2):246-257.

Wang, L., J. Lyons, P. Kanehl, and R. Gatti. 1997. Influences of watershed land use on habitat quality and biotic integrity in Wisconsin Streams. *Fisheries* 22(6): 6-12.

Whittier, T.R., R.M. Hughes, and D.P. Larsen. 1988. Correspondence between ecoregions and spatial patterns in stream ecosystems in Oregon. *Canadian Journal of Fisheries and Aquatic Sciences* 45:1264-1278.

Whittier, T.R., and S.G. Paulsen. 1992. The surface waters component of the Environmental Monitoring and Assessment Program (EMAP): an overview. *Journal of Aquatic Ecosystem Health* 1:119-126.

Yoder, C.O. and E.T. Rankin. 1995. Biological criteria program development and implementation in Ohio. Pages 109-144 *in* W.S. Davis and T.P. Simon (eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, Florida.