4 PERFORMANCE-BASED METHODS SYSTEM (PBMS)

Determining the performance characteristics of individual methods enables agencies to share data to a certain extent by providing an estimate of the level of confidence in assessments from one method to the next. The purpose of this chapter is to provide a framework for measuring the performance characteristics of various methods. The contents of this chapter are taken liberally from Diamond et al. 1996, which is a refinement of the PBMS approach developed for ITFM (1995b). This chapter is best assimilated if the reader is familiar with data analysis for bioassessment. Therefore, the reader may wish to review Chapter 9 on data analysis before reading this PBMS material. Specific quality assurance aspects of the methods are included in the assemblage chapters.

Regardless of the type of data being collected, field methods share one important feature in common—they cannot tell whether the information collected is an accurate portrayal of the system of interest (Intergovernmental Task Force on Monitoring Water Quality [ITFM] 1995a). Properties of a given field sample can be known, but research questions typically relate to much larger spatial and temporal scales. It is possible to know, with some accuracy, properties or characteristics of a given sample taken from the field; but typically, research questions relate to much larger spatial and temporal scales. To grapple with this problem, environmental scientists and statisticians have long recognized that field methods must strive to obtain information that is representative of the field conditions at the time of sampling.

An accurate assessment of stream biological data is difficult because natural variability cannot be controlled (Resh and Jackson 1993). Unlike analytical assessments conducted in the laboratory, in which accuracy can be verified in a number of ways, the accuracy of macroinvertebrate assessments in the field cannot be objectively verified. For example, it isn't possible to "spike" a stream with a known species assemblage and then determine the accuracy of a bioassessment method. This problem is not theoretical. Different techniques may yield conflicting interpretations at the same sites, underscoring the question of accuracy in bioassessment. Depending on which methods are chosen, the actual structure and condition of the assemblage present, or the trends in status of the assemblage over time may be misinterpreted. Even with considerable convergence in methods used in the U.S. by states and other agencies (Southerland and Stribling 1995, Davis et al. 1996), direct sharing of data among agencies may cause problems because of the uncertainty associated with unfamiliar methods, misapplication of familiar methods, or varied data analyses and interpretation (Diamond et al. 1996).

4.1 APPROACHES FOR ACQUIRING COMPARABLE BIOASSESSMENT DATA

Water quality management programs have different reasons for doing bioassessments which may not require the same level or type of effort in sample collection, taxonomic identification, and data analysis (Gurtz and Muir 1994). However, different methods of sampling and analysis may yield comparable data for certain objectives despite differences in effort. There are 2 general approaches for acquiring comparable bioassessment data among programs or among states. The first is for everyone to use the same method on every study. Most water resource agencies in the U.S. have developed standard operating procedures (SOPs). These SOPs would be adhered to throughout statewide or regional areas

to provide comparable assessments within each program. The Rapid Bioassessment Protocols (RBPs) developed by Plafkin et al. (1989) and refined in this document are attempts to provide a framework for agencies to develop SOPs. However, the use of a single method, even for a particular type of habitat, is probably not likely among different agencies, no matter how exemplary (Diamond et al. 1996).

The second approach to acquiring comparable data from different organizations, is to encourage the documentation of performance characteristics (e.g., precision, sensitivity) for all methods and to use those characteristics to determine comparability of different methods (ITFM 1995b). This documentation is known as a performance-based method system (PBMS) which, in the context of biological assessments, is defined as a system that permits the use of any method (to sample and analyze stream assemblages) that meets established requirements for data quality (Diamond et al. 1996). Data quality objectives (DQOs) are qualitative and quantitative expressions that define requirements for data precision, bias, method sensitivity, and range of conditions over which a method yields satisfactory data (Klemm et al. 1990). The determination of DQOs for a given study or agency program is central to all data collection and to a PBMS, particularly, because these objectives establish not only the necessary quality of a given method (Klemm et al. 1990) but also the types of methods that are likely to provide satisfactory information.

In practice, DQO's are developed in 3 stages: (1) determine what information is needed and why and how that information will be used; (2) determine methodological and practical constraints and technical specifications to achieve the information desired; and (3) compare different available methods and choose the one that best meets the desired specifications within identified practical and technical limitations (USEPA 1984, 1986, Klemm et al. 1990, USEPA 1995a, 1997c). It is difficult to make an informed decision regarding which methods to use if data quality characteristics are unavailable. The successful introduction of the PBMS concept in laboratory chemistry, and more recently in laboratory toxicity testing (USEPA 1990c, American Society of Testing and Materials [ASTM] 1995), recommends adapting such a system for biological monitoring and assessment.

If different methods are similar with respect to the quality of data each produces, then results of an assessment from those methods may be used interchangeably or together. As an example, a method for sample sorting and organism identification, through repeated examination using trained personnel, could be used to determine that the proportion of missed organisms is less than 10% of the organisms present in a given sample and that taxonomic identifications (to the genus level) have an accuracy rate of at least 90% (as determined by samples verified by recognized experts). A study could require the above percentages of missed organisms and taxonomic accuracy as DQOs to ensure the collection of satisfactory data (Ettinger 1984, Clifford and Casey 1992, Cuffney et al. 1993a). In a PBMS approach, any laboratory sorting and identification method that documented the attainment of these DQOs would yield comparable data and the results would therefore be satisfactory for the study.

For the PBMS approach to be useful, 4 basic assumptions must be met (ITFM 1995b):

1. DQOs must be set that realistically define and measure the quality of the data needed; reference (validated) methods must be made available to meet those DQOs;

2. to be considered satisfactory, an alternative method must be as good or better than the reference method in terms of its resulting data quality characteristics;

3. there must be proof that the method yields reproducible results that are sensitive enough for the program; and

4. the method must be effective over the prescribed range of conditions in which it is to be used. For bioassessments, the above assumptions imply that a given method for sample collection and analysis produces data of known quality, including precision, the range of habitats over which the collection method yields a specified precision, and the magnitude of difference in data among sites with different levels or types of impairment (Diamond et al. 1996).

PERFORMANCE CHARACTERISTICS

- **Precision**
- **Bias**
- **Performance range**
- **Interferences**
- **Sensitivity**

Thus, for multimetric assessment methods, such as RBPs, the precision of the total multimetric score is of interest as well as the individual metrics that make up the score (Diamond et al. 1996). Several performance characteristics must be characterized for a given method to utilize a PBMS approach. These characteristics include method precision, bias, performance range, interferences, and sensitivity (detection limit). These characteristics, as well as method accuracy, are typically demonstrated in analytical chemistry systems through the use of blanks, standards, spikes, blind samples, performance evaluation samples, and other techniques to compare different methods and eventually derive a reference method for a given analyte. Many of these performance characteristics are applicable to biological laboratory and field methods and other prelaboratory procedures as well (Table 4-1). It is known that a given collection method is not equally accurate over all ecological conditions even within a general aquatic system classification (e.g., streams, lakes, estuaries). Therefore, assuming a given method is a "reference method" on the basis of regulatory or programmatic reasons does not allow for possible translation or sharing of data derived from different methods because the performance characteristics of different methods have not been quantified. One can evaluate performance characteristics of methods in 2 ways: (1) with respect to the collection method itself and, (2) with respect to the overall assessment process. Method performance is characterized using quantifiable data (metrics, scores) derived from data collection and analysis. Assessment performance, on the other hand, is a step removed from the actual data collected. Interpretive criteria (which may be based on a variety of approaches) are used to rank sites and thus, PBMS in this case is concerned with performance characteristics of the ranking procedures as well as the methods that lead to the assessment.

Table 4-1. Progression of a generic bioassessment field and laboratory method with associated examples of performance characteristics.

Step	Procedure	Examples of Performance Characteristics
1	Sampling device	***Precision***—repeatability in a habitat. ***Bias***—exclusion of certain taxa (mesh size). ***Performance range***—different efficiency in various habitat types or substrates. ***Interferences***—matrix or physical limitations (current velocity, water depth).
2	Sampling method	***Precision***—variable metrics or measures among replicate samples at a site. ***Bias***—exclusion of certain taxa (mesh size) or habitats. ***Performance range***—limitations in certain habitats or substrates. ***Interferences***—high river flows, training of personnel.

Table 4-1. Progression of a generic bioassessment field and laboratory method with associated examples of performance characteristics. (Continued)

Step	Procedure	Examples of Performance Characteristics
3	Field sample processing (subsampling, sample transfer, preservation)	***Precision***—variable metrics among splits of subsamples. ***Bias***— efficiency of locating small organisms. ***Performance range***—sample preservation and holding time. ***Interferences***—Weather conditions. **Additional characteristics:** ***Accuracy***—of sample transfer process and labeling.
4	Laboratory sample processing (sieving, sorting)	***Precision***—split samples. ***Bias***—sorting certain taxonomic groups or organism size. ***Performance range***—sorting method depending on sample matrix (detritus, mud). ***Interferences***—distractions; equipment. **Additional characteristics:** ***Accuracy***—sorting method; lab equipment.
5	Taxonomic enumeration	***Precision***—split samples. ***Bias***—counts and identifications for certain taxonomic groups. ***Performance range***—dependent on taxonomic group and (or) density. ***Interferences***—appropriateness of taxonomic keys. ***Sensitivity***— level of taxonomy related to type of stressor **Additional characteristics:** ***Accuracy***—identification and counts.

Data quality and performance characteristics of methods for analytical chemistry are typically validated through the use of quality control samples including blanks, calibration standards, and samples spiked with a known quantity of the analyte of interest. Table 4-2 summarizes some performance characteristics used in analytical chemistry and how these might be translated to biological methods.

The collection of high-quality data, particularly for bioassessments, depends on having adequately trained people. One way to document satisfactory training is to have newly trained personnel use the method and then compare their results with those previously considered acceptable. Although field crews and laboratory personnel in many organizations are trained in this way (Cuffney et al. 1993b), the results are rarely documented or quantified. As a result, an organization cannot assure either itself or other potential data users that different personnel performing the same method at the same site yield comparable results and that data quality specifications of the method (e.g., precision of metrics or scores) are consistently met. Some of this information is published for certain bioassessment sampling methods, but is defined qualitatively (see Elliott and Tullett 1978, Peckarsky 1984, Resh et al. 1990, Merritt et al. 1996 for examples), not quantitatively. Quantitative information needs to be more available so that the quality of data obtained by different methods is documented.

Table 4-2. Translation of some performance characteristics, derived for laboratory analytical systems, to biological laboratory systems (taken from Diamond et al. 1996).

Performance Characteristics	Analytical Chemical Methods	Biological Methods
Precision	Replicate samples	Multiple taxonomists identifying 1 sample; split sample for sorting, identification, enumeration; replicate samples within sites; duplicate reaches
Bias	Matrix-spiked samples; standard reference materials; performance evaluation samples	Taxonomic reference samples; "spiked" organism samples
Performance range	Standard reference materials at various concentrations; evaluation of spiked samples by using different matrices	Efficiency of field sorting procedures under different sample conditions (mud, detritus, sand, low light)
Interferences	Occurrence of chemical reactions involved in procedure; spiked samples; procedural blanks; contamination	Excessive detrital material or mud in sample; identification of young life stages; taxonomic uncertainty
Sensitivity	Standards; instrument calibration	Organism-spiked samples; standard level of identification
Accuracy	Performance standards; procedural blanks	Confirmation of identification, percentage of "missed" specimens

It is imperative that the specific range of environmental conditions (or performance range) is quantitatively defined for a sampling method (Diamond et al. 1996). As an example, the performance range for macroinvertebrate sampling is usually addressed qualitatively by characterizing factors such as stream size, hydrogeomorphic reach classification, and general habitat features (riffle vs. pool, shallow vs. deep water, rocky vs. silt substrate; Merritt et al. 1996). In a PBMS framework, different methods could be classified based on the ability of the method to achieve specified levels of performance characteristics such as data precision and sensitivity to impairment over a range of appropriate habitats. Thus, the precision of individual metrics or scores obtained by different sampling methods can be directly and quantitatively compared for different types of habitats.

4.2 ADVANTAGES OF A PBMS APPROACH FOR CHARACTERIZING BIOASSESSMENT METHODS

Two fundamental requirements for a biological assessment are: (1) that the sample taken and analyzed is representative of the site or the assemblage of interest and, (2) that the data obtained are an accurate reflection of the sample. The latter requirement is ensured using proper quality control (QC) in the laboratory including the types of performance characteristics summarized in Table 4-2. The first requirement is met through appropriate field sampling procedures, including random selection of sampling locations within the habitat type(s) of interest, choice of sampling device, and sample preservation methods. The degree to which a sample is representative of the environment depends on the type of sampling method used (including subsampling) and the ecological endpoint being measured. For example, many benthic samples may be needed from a stream to obtain 95% confidence intervals that are within 50% of the mean value for macroinvertebrate density, whereas fewer benthic samples may be needed to determine the dominant species in a given habitat type at a particular time (Needham and Usinger 1956, Resh 1979, Plafkin et al. 1989).

Several questions have been raised concerning the appropriateness or "accuracy" of methods such as RBPs, which take few samples from a site and base their measures or scores on subsamples. Subsampling methods have been debated relevant to the "accuracy" of data derived from different methods (Courtemanch 1996, Barbour and Gerritsen 1996, Vinson and Hawkins 1996). Using a PBMS framework, the question is not which subsampling method is more "accurate" or precise but rather what accuracy and precision level can a method achieve, and do those performance characteristics meet the DQOs of the program? Looking at bioassessment methods in this way, (including subsampling and taxonomic identification), forces the researcher or program manager to quantitatively define beforehand the quality control characteristics necessary to make the type of interpretive assessments required by the study or program.

Once the objectives and data quality characteristics are defined for a given study, a method is chosen that meets those objectives. Depending on the data quality characteristics desired, several different methods for collecting and sorting macroinvertebrates may be suitable. Once data precision and "accuracy" are quantified for measures derived from a given bioassessment method, the method's sensitivity (the degree of change in measures or endpoints between a test site and a control or reference site that can be detected as a difference) and reliability (the degree to which an objectively defined impaired site is identified as such) can be quantified and compared with other methods. A method may be modified (e.g., more replicates or larger samples taken) to improve the precision and "accuracy" of the method and meet more stringent data requirements. Thus, a PBMS framework has the advantage of forcing scientists to focus on the ever-important issue: what type of sampling program and data quality are needed to answer the question at hand?

A second advantage of a PBMS framework is that data users and resource managers could potentially increase the amount of available information by combining data based on known comparable methods. The 305(b) process of the National Water Quality Inventory, (USEPA 1997c) is a good example of an environmental program that would benefit from a PBMS framework. This program is designed to determine status and trends of surface water quality in the U.S. A PBMS framework would make explicit the quality and comparability of data derived from different bioassessment methods, would allow more effective sharing of information collected by different states, and would improve the existing national database. Only those methods that met certain DQOs would be used. Such a decision might encourage other organizations to meet those minimum data requirements, thus increasing the amount of usable information that can be shared. For example, the RBPs used by many state agencies for water resources (Southerland and Stribling 1995) could be modified for field and laboratory procedures and still meet similar data quality objectives. The overall design steps of the RBPs, and criteria for determining useful metrics or community measures, would be relatively constant across regions and states to ensure similar quality and comparability of data.

4.3 QUANTIFYING PERFORMANCE CHARACTERISTICS

The following suggested sampling approach (Figure 4-1) need only be performed once for a particular method and by a given agency or research team; it need not be performed for each bioassessment study. Once data quality characteristics for the method are established, limited quality control (QC) sampling and analysis should supplement the required sampling for each bioassessment study to ensure that data quality characteristics of the method are met (USEPA 1995a). The additional effort and expense of such QC are negligible in relation to the potential environmental cost of producing data of poor or unknown quality.

The first step is to define precision of the collection method, also known as "measurement error". This is accomplished by replicate sampling within sites (see Hannaford and Resh 1995). The samples

collected are processed and analyzed separately and their metrics compared to obtain a more realistic measure of the method precision and consistency. Repeated samples within sites estimate the precision of the entire method, comprising variability due to several sources including small-scale spatial variability within a site; operator consistency and bias; and laboratory consistency. Finally, it is desirable to sample a range of site classes (stream size, habitat type) over which the method is likely to be used. This kind of sampling, processing, and analysis should reveal potential biases.

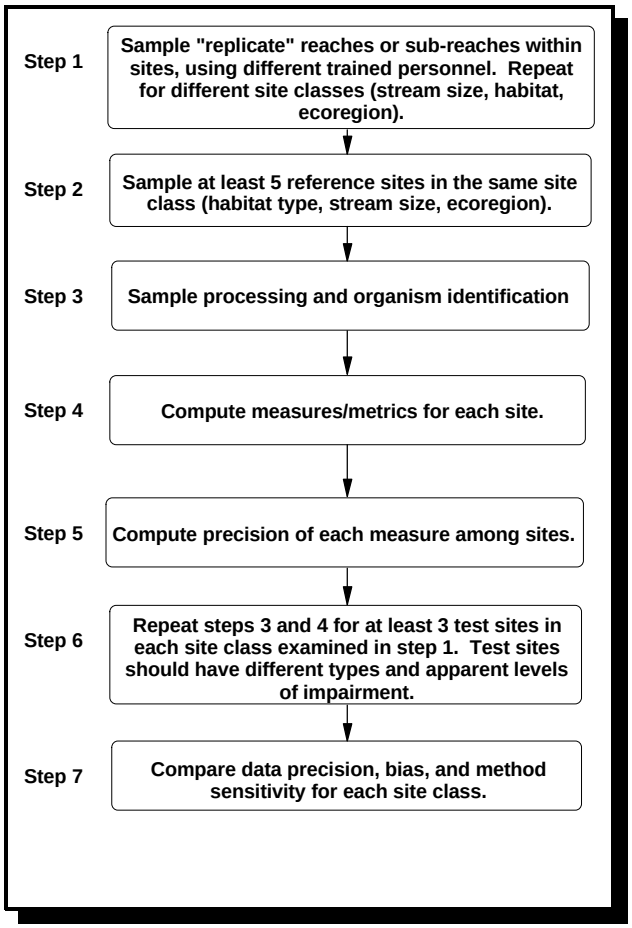

Figure 4-1. Flow chart summarizing the steps necessary to quantify performance characteristics of a bioassessment method (modified from Diamond et al. 1996).

Once the precision of the method is known, one can determine the actual variability associated with sampling "replicate" reference sites within an ecoregion or habitat type. This is known as sampling error, referring to the sample (of sites) drawn from a subpopulation (sites in a region). The degree of assemblage similarity observed among "replicate" reference streams, along with the precision of the collection method itself, will determine the overall precision, accuracy, and sensitivity of the bioassessment approach as a whole. This kind of checking has been done, at least in part, by several states (Bode and Novak 1995; Yoder and Rankin 1995a; Hornig et al. 1995; Barbour et al. 1996b), some USEPA programs (Gibson et al. 1996), and the U.S. Geological Survey (USGS) National Water Quality Assessment Program (Cuffney et al. 1993b, Gurtz 1994). Evaluation of metric or score variability among replicate reference sites can result in improved data precision and choices of stream classification. For example, the Arizona Department of Environmental Quality (DEQ) determined that macroinvertebrate assemblage structure varied substantially within ecoregions resulting in large metric variability among reference sites and poor classification (Spindler 1996). Using detrended correspondence and cluster analysis, the state agency determined that discrimination of sites by elevation and watershed area, corresponding to montane upland, desert lowland, and transition zones, resulted in much lower variability among reference sites and a better classification scheme to measure sensitivity to impairment.

If multiple reference sites are sampled in different site classes (where the sampling method is judged to be appropriate), several important method performance characteristics can be quantified, including: (1) precision for a given metric or assessment score across replicate reference sites within a site class; (2) relative precision of a given metric or score among reference sites in different classes; (3) range of classes over which a given method yields similar precision and "accuracy"; (4) potential interferences to a given method that are related to specific class characteristics and qualities; and (5) bias of a given metric, method, or both, owing to differences in classes (Diamond et al. 1996).

A study by Barbour et al. (1996b) for Florida streams, illustrates the importance of documenting method performance characteristics using multiple reference sites in different site classes. Using the same method at all sites, fewer taxa were observed in reference sites from the Florida Peninsula (one site class) compared to the Florida Panhandle (another site class), resulting in much lower reference values for taxa richness metrics in the Peninsula. Although metric precision was similar among reference sites in each site class, method sensitivity (i.e., the ability of a metric to discern a difference between reference and stressed sites) was poorer in the Peninsula for taxa richness. Thus, bioassessment "accuracy" may be more uncertain for the Florida Peninsula; that is, the probability of committing a Type II error (concluding a test site is no different from reference — therefore minimally impaired — when, in fact, it is) may be greater in the Peninsula region. In the context of a PBMS, the state agency can recognize and document differences in method performance characteristics between site classes and incorporate them into their DQOs. The state in this case can also use the method performance results to identify those site classes for which the biological indicator (index, metric, or other measurement endpoint) may not be naturally sensitive to impairment; i.e., the fauna is naturally species-poor and thus less likely to reflect impacts from stressors. If the state agency desires greater sensitivity than the current method provides, it may have to develop and test different region-specific methods and perhaps different indicators.

In the last step of the process, a method is used over a range of impaired conditions so as to determine the method's sensitivity or ability to detect impairment. As discussed earlier, sites with known levels of impairment or analogous standards by which to create a calibration curve for a given bioassessment method are lacking. In lieu of this limitation, sampling sites are chosen that have known stresses (e.g., urban runoff, toxic pollutants, livestock intrusion, sedimentation, pesticides). Because different sites may or may not have the same level of impairment within a site class (i.e., they are not replicate sites), precision of a method in impaired sites may best be examined by taking and analyzing multiple samples from the same site or adjacent reaches (Hannaford and Resh 1995).

The quantification of performance characteristics is a compromise between statistical power and cost while maintaining biological relevance. Given the often wide variation of natural geomorphic conditions and landscape ecology, even within supposedly "uniform" site classes (Corkum 1989, Hughes 1995), it is desirable to examine 10 or more reference sites (Yoder and Rankin 1995a, Gibson et al. 1996). More site classes in the evaluation process would improve documentation of the performance range and bias for a given method. Using the sampling design suggested in Figure 4-1, data from at least 30 sites (reference and test sites combined), sampled within a brief time period (so as to minimize seasonal changes in the target assemblage), are needed to define performance characteristics. An alternative approach might be to use bootstrap resampling of fewer sites to evaluate the nature of variation of these samples (Fore et al. 1996).

A range of "known" stressed sites within a site class is sampled to test the performance characteristics of a given method. It is important that stressed sites meet the following criteria: (1) they belong to the same site class as the reference sites examined; (2) they clearly have been receiving some chemical, physical, or biological stress(es) for some time (months at least); and (3) impairment is not obvious without sampling; i.e., impairment is not severe.

The first criterion is necessary to reduce potential interferences owing to class differences between the test and reference sites. Thus, the condition of the reference site will have high probability of serving as a true blank as discussed earlier. For example, it is clearly inappropriate to use high gradient mountain streams as references for assessing plains streams.

The second criterion, which is the documented presence of potential stresses, is necessary to ensure the likelihood that the test site is truly impaired (Resh and Jackson 1993). A potential test site might include a body of water that receives toxic chemicals from a point-source discharge or from nonpoint sources, or a water body that has been colonized by introduced or exotic "pest" species (for example, zebra mussel or grass carp). Stresses at the test site should be measured quantitatively to document potential cause(s) of impairment.

The third criterion, that the site is not obviously impaired, provides a reasonable test of method sensitivity or "detection limit." Severe impairment (e.g., a site that is dominated by 1 or 2 invertebrate species, or a site apparently devoid of aquatic life) generally requires little biological sampling for detection.

4.4 RECOMMENDED PROCESS FOR DOCUMENTATION OF METHOD COMPARABILITY

Although a comparison of methods at the same reference and test sites at the same time is preferable (same seasons and similar conditions), it is not essential. The critical requirement when comparing different sampling methods is that performance characteristics for each method are derived using similar habitat conditions and site classes at similar times/seasons (Diamond et al. 1996). This approach is most useful when examining the numeric scores upon which the eventual assessment is based. Thus, for a method such as RBP that sums the values of several metrics to derive a single score for a site, the framework described in Figure 4-1 should use the site scores. If one were interested in how a particular multimetric scoring system behaves, or one wishes to compare the same metric across methods, then individual metrics could be examined using the framework in Figure 4-1. For multivariate assessment methods that do not compute metric scores, one could instead examine a measure of community similarity or other variable that the researcher uses in multivariate analyses (Norris 1995).

Method comparability is based on 2 factors: (1) the relative magnitude of the coefficients of variation in measurements within and among site classes, and (2) the relative percent differences in measurements between reference and test sites. It is important to emphasize that comparability is not based on the measurements themselves, because different methods may produce different numeric scores or metrics and some sampling methods may explicitly ignore certain taxonomic groups, which will influence the metrics examined. Instead, detection of a systematic relationship among indices or the same measures among methods is advised. If 2 methods are otherwise comparable based on similar performance characteristics, then results of the 2 methods can be numerically related to each other. This outcome is a clear benefit of examining method comparability using a PBMS framework.

Figure 4-1 summarizes a suggested test design, and Table 4-3 summarizes recommended analyses for documenting both the performance characteristics of a given method, and the degree of data comparability between 2 or more methods. The process outlined in Figure 4-1 is not one that is implemented with every study. Rather, the process should be performed at least once to document the limitations and range of applicability of the methods, and should be cited with subsequent uses of the method(s).

The following performance characteristics are quantified for each bioassessment method and compared: (1) the within-class coefficient of variation for a given metric score or index by examining reference-site data for each site class separately (e.g., CV_{A1r} and CV_{B1r}; Fig. 4-1); (2) difference or bias in precision related to site class for a given metric or index (by comparing reference site coefficient of

variation from each class: CV_{A1r}/CV_{B1r}; Table 4-3); and (3) estimates of method sensitivity or discriminatory power, by comparing test site data with reference site data

Table 4-3. Suggested arithmetic expressions for deriving performance characteristics that can be compared between 2 or more methods. In all cases, $\bar{x}$ = mean value, X = test site value, s = standard deviation. Subscripts are as follows: capital letter refers to site class (A or B); numeral refers to method 1 or 2; and lower case letter refers to reference (r) or test site (t) (modified from Diamond et al. 1996).

Performance Characteristic	Parameters for Quantifying Method Comparability	Desired Outcome
Relative *precision* of metric or index *within* a site class	CV_{A1r} and CV_{A2r} ; CV_{B1r} and CV_{B2r}	Low values
Relative *precision* of metric or index *between* sites (population of samples at a site) or site classes (population of sites)	$\frac{CV_{A1r}}{CV_{B1r}}$; $\frac{CV_{A2r}}{CV_{B2r}}$	High ratio
Relative *sensitivity* or "detection limit" of metric or index *within* a site class. Comparison of those values between methods reveals the most sensitive method	$\frac{\bar{x}_{A1r}-X_{A1t}}{s_{A1r}}$; $\frac{\bar{x}_{A2r}-X_{A2t}}{s_{A2r}}$ $\frac{\bar{x}_{B1r}-X_{B1t}}{s_{B1r}}$; $\frac{\bar{x}_{B2r}-X_{B2t}}{s_{B2r}}$	High ratio
Relative *sensitivity* of metric or index *between* site classes	$\frac{\bar{x}_{A1r}-X_{A1t}}{s_{A1r}}$; $\frac{\bar{x}_{B1r}-X_{B1t}}{s_{B1r}}$ $\frac{\bar{x}_{A2r}-X_{A2t}}{s_{A2r}}$; $\frac{\bar{x}_{B2r}-X_{B2t}}{s_{B2r}}$	High ratio

within each site class as a function of reference site variability (Table 4-3), e.g.,

$$\frac{\bar{x}_{A1r}-X_{A1t}}{s_{A1r}}$$

A method that yields a smaller difference between test and reference sites in relation to the reference site variability measured (Table 4-3) would indicate less discriminatory power or sensitivity; that is, the test site is erroneously perceived to be similar to or better than the reference condition and not impaired (Type II error).

Relatively few methods may be able to consistently meet the above data quality criterion and also maintain high sensitivity to impairment because both characteristics require a method that produces relatively precise, accurate data. For example, if the agency's intent is to screen many sites so as to prioritize "hot spots" or significant impairment in need of corrective action, then a method that is inexpensive, quick, and tends to show impairment when significant impairment is actually present

(such as some volunteer monitoring methods) (Barbour et al. 1996a) can meet prescribed DQOs with less cost and effort. In this case, the data requirements dictate high priority for method sensitivity or discriminatory power (detection if impaired sites), understanding that there is likely also to be a high Type I error rate (misidentification of unimpaired sites).

Relative accuracy of each method is addressed to the extent that the test sites chosen are likely to be truly impaired on the basis of independent factors such as the presence of chemical stresses or suboptimal habitat. A method with relatively low precision (high variance) among reference sites compared with another method may suggest lower method accuracy. Note that a method having lower precision may still be satisfactory for some programs if it has other advantages, such as high ability to detect impaired sites with less cost and effort to perform.

Once performance characteristics are defined for each method, data comparability can be determined. If 2 methods are similarly precise, sensitive, and biased over the habitat types sampled, then the different methods should produce comparable data. Interpretive judgements could then be made concerning the quality of aquatic life using data produced by either or both methods combined. Alternatively, the comparison may show that 2 methods are comparable in their performance characteristics in certain habitats or regions and not others. If this is so, results of the 2 methods can be combined for the type for the types of habitats in which data comparability was demonstrated, but not for other regions or habitat types.

In practice, comparability of bioassessment methods would be judged relative to a reference method that has already been fully characterized (using the framework summarized in Figure 4-1) and which produces data with the quality needed by a certain program or agency. The qualities of this reference method are then defined as method performance criteria. If an alternative method yields less precision among reference sites within the same site class than the reference method (e.g., $CV_{A1r} > CV_{A2r}$ in Table 4-3), then the alternative method probably is not comparable to the reference method. A program or study could require that alternative methods are acceptable only if they are as precise as the reference method. A similar process would be accomplished for other performance characteristics that a program or agency deems important based on the type of data required by the program or study.

4.5 CASE EXAMPLE DEFINING METHOD PERFORMANCE CHARACTERISTICS

Florida Department of Environmental Protection (DEP) has developed a statewide network for monitoring and assessing the state's surface waters using macroinvertebrate data. Florida DEP has rigorously examined performance characteristics of their collection and assessment methods to provide better overall quality assurance of their biomonitoring program and to provide defensible and appropriate assessments of the state's surface waters (Barbour et al. 1996b, c). Much of the method characterization process developed for Florida DEP is easily communicated in the context of a PBMS approach.

In addition to characterizing data quality and method performance based on ecoregional site classes, Florida DEP also characterized their methods based on season (summer vs. winter sampling index periods), and size of subsample analyzed (100, 200, or 300-organism subsample). In addition, analyses were performed on the individual component metrics which composed the Florida stream condition index (SCI). For the sake of brevity, the characterization process and results for the SCI in the summer index period and the Peninsula and Northeast bioregions are summarized. The same process was used for other bioregions in the state and in the winter index period.

Performance Criteria Characteristics of Florida SCI (see Figure 4-1 for process)

Characterize Measurement Error (Method Precision Within a Site)—A total of 7 sites in the Peninsula bioregion were subjected to multiple sampling (adjacent reaches). The DEP observed a mean SCI = 28.4 and a CV (within a stream) = 6.8%. These data suggest low measurement error associated with the method and the index score. Given this degree of precision in the reference condition SCI score, power analysis indicated that 80% of the time, a test site with an SCI 5 points less (based on only a single sample at the test site) than the reference criterion, could be distinguished as impaired with 95% confidence. This analysis also indicated that if duplicate samples were taken at the test site, a difference of 3 points in the SCI score between the test site and the reference criterion could be distinguished as impaired with 95% confidence.

Characterize Sampling Error (Method Precision on a Population of Reference Sites)—A total of 56 reference sites were sampled in the Peninsula bioregion (Step 1, Figure 4-1). The SCI score could range from a minimum of 7 to a theoretical maximum of 31 based on the component metric scores. However, in the Peninsula, reference site SCI scores generally ranged between 21 and 31. A mean SCI score of 27.6 was observed with a CV of 12.0%.

Determine Method and Index Sensitivity—Distribution of SCI scores of the 56 reference sites showed that the 5th percentile was a score of 20. Thus, 95% of Peninsula reference sites had a score >20. Accuracy of the method, using known stressed sites, indicated that approximately 80% of the test sites had SCI scores $\leq$ 20 (Fig. 4-2). In other words, a stressed site would be assessed as impaired 80% of the time using the collection method in the Peninsula bioregion in the summer, and an impairment criterion of the 5th percentile of reference sites. The criterion could also be raised to, say, the 25th percentile of reference sites, which would increase accuracy of correctly classifying stressed sites to approximately 90%, but would decrease accuracy of correctly assessing unimpaired sites to 75%.

Determination of Method Bias and Relative Sensitivity in Different Site Classes—A comparative analysis of precision, sensitivity, and ultimately bias, can be performed for the Florida DEP method and the SCI index outlined in Table 4-3. For example, the mean SCI score in the Panhandle bioregion, during the same summer index period, was 26.3 with a CV = 12.8% based on 16 reference sites. Comparing this CV to the one reported for the Peninsula in the previous step, it is apparent that the precision of this method in the Panhandle was similar to that observed in the Peninsula bioregion.

The 5th percentile of the Panhandle reference sites was an SCI score of 17, such that actual sensitivity of the method in the Panhandle was slightly lower than in the Peninsula bioregion (Figure 4-2). An impaired site would be assessed as such only 50% of the time in the Panhandle bioregion in the summer as opposed to 80% of the time in the Peninsula bioregion during the same index period. Part of the difference in accuracy of the method among the 2 bioregions can be attributed to differences in sample size. Data from only 4 "known" impaired sites were available in the Panhandle bioregion while the Peninsula bioregion had data from 12 impaired sites. The above analyses show, however, that there may be differences in method performance between the 2 regions (probably attributable to large habitat differences between the regions) which should be further explored using data from additional "known" stressed sites, if available.

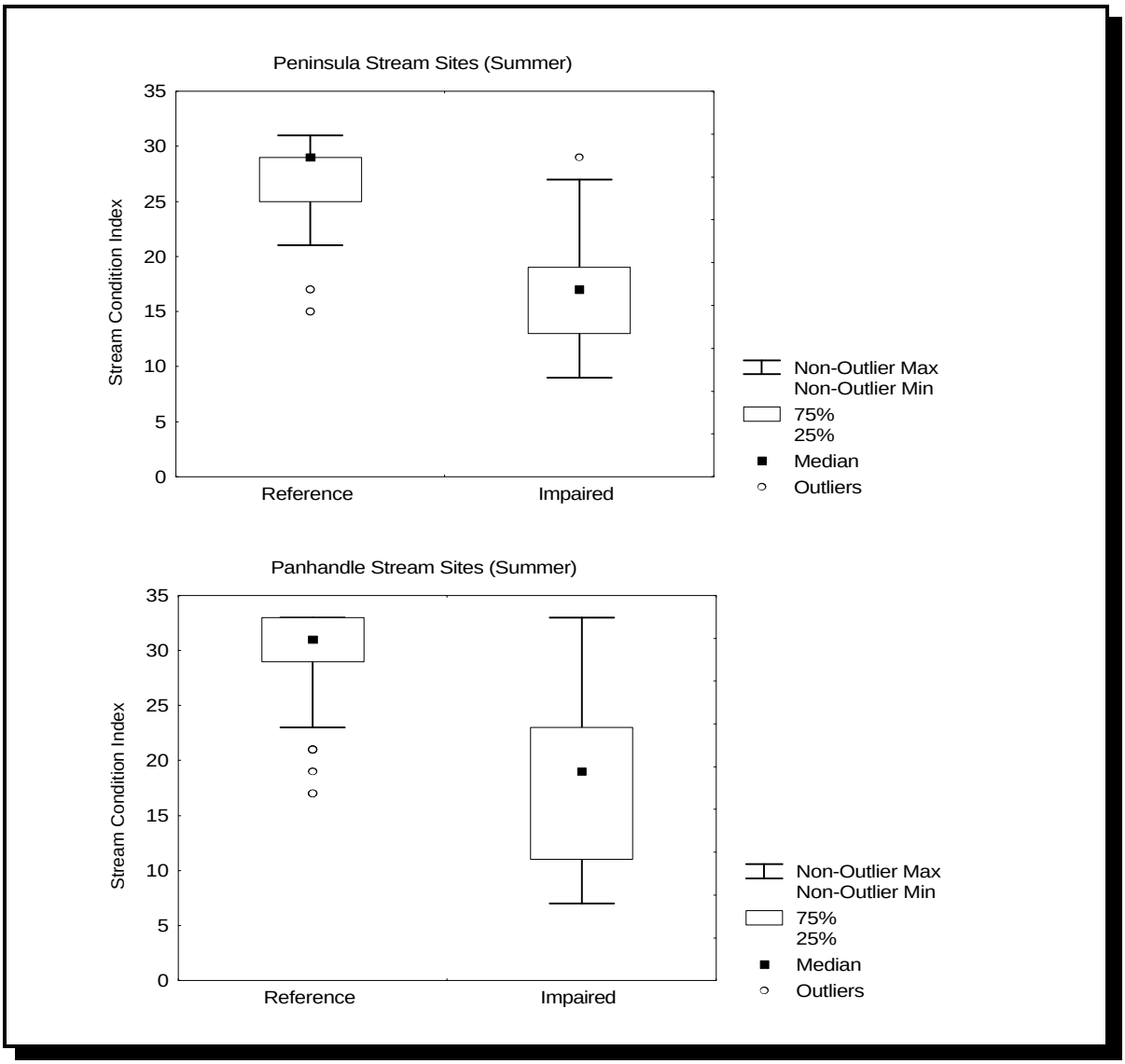

Figure 4-2. Comparison of the discriminatory ability of the SCI between Florida's Peninsula and Panhandle Bioregions. Percentiles used (not $\bar{x}$, sd) to depict relationship.

4.6 APPLICATION OF THE PBMS

The PBMS approach is intended to provide information regarding the confidence of an assessment, given a particular method. By having some measure of confidence in the endpoint and the subsequent decision pertinent to the condition of the water resource, assessment and monitoring programs are greatly strengthened. Three primary questions can be identified that enable agencies to ascertain the value and scientific validity of using information derived from different methods. Use of PBMS is necessary for these questions to be answered.

Question 1 — How rigorous must a method be to accurately detect impairment?

The analyses of Ohio EPA (1992) reveal that the power and ability of a bioassessment technique to accurately portray biological community performance and ecological integrity, and to discriminate even finer levels of aquatic life use impairments, are directly related to the data dimensions (i.e., ecological

complexity, environmental accuracy, discriminatory power) produced by each (Barbour et al. 1996b). For example, a technique that includes the identification of macroinvertebrate taxa to genus and species will produce a higher attainment of data dimensions than a technique that is limited to family-level taxonomy. In general, this leads to a greater discrimination of the biological condition of sites.

Some states use one method for screening assessments and a second method for more intensive and confirmatory assessments. Florida DEP uses a BioRecon (see description in Chapter 7) to conduct statewide screening for their watershed-based monitoring. A more rigorous method based on a multihabitat sampling (see Chapter 7) is used for targeted surveys related to identified or suspected problem areas. North Carolina Water Quality Division (WQD) has a rapid EPT index (cumulative number of species of Ephemeroptera, Plecoptera, Trichoptera) to conduct screening assessments. Their more intensive method is used to monitor biological condition on a broader basis.

Use of various methods having differing levels of rigor can be examined with estimates of precision and sensitivity. These performance characteristics will help agencies make informed decisions of how resulting data can be used in assessing condition.

Question 2 — How can data derived from different methods be compared to locate additional reference sites?

Many agencies are increasingly confronted with the issue of locating appropriate reference sites from which to develop impairment/unimpairment thresholds. In some instances, sites outside of jurisdictional boundaries are needed to refine the reference condition. As watershed-based monitoring becomes implemented throughout the U.S., jurisdictional boundaries may become impediments to effective monitoring. County governments, tribal associations, local environmental interest groups, and state water resource agencies are all examples of entities that would benefit from collaborative efforts to identify common reference sites.

In most instances, all of the various agencies conducting monitoring and assessment will be using different methods. A knowledge of the precision and sensitivity of the methods will allow for an agency to decide whether the characterization of a site as reference or minimally impaired by a second agency or other entity fits the necessary criteria to be included as an additional reference site.

Question 3 — How can data from different methods be combined or integrated for increasing a database for assessment?

The question of combining data for a comprehensive assessment is most often asked by states and tribes that want to increase the spatial coverage of an assessment beyond their own limited datasets. From a national or regional perspective, the ability to combine datasets is desirable to make judgements on the condition of the water resource at a higher geographical scale. Ideally, each dataset will have been collected with the same methods.

This question is the most difficult to answer even with a knowledge of the precision and sensitivity. Widely divergent methodologies having highly divergent performance characteristics are not likely to be appropriate for combining under any circumstances. The risk of committing error in judgement of biological condition from a combined dataset of this sort would be too high.

Divergent methodologies with similar or nearly identical performance characteristics are plausible candidates for combining data at metric or index levels. However, a calibration of the methods is

necessary to ensure that extrapolations of data from one method to the other is scientifically valid. The best fit for a calibrated model is a 1:1 ratio for each metric and index. Realistically, the calibration will be on a less-than-perfect relationship; extrapolations may be via range of values rather than absolute numbers. Thus, combining datasets from dissimilar methods may be valuable for characterizing severe impairment or sites of excellent condition. However, sites with slight to moderate impairment might not be detected with a high level of confidence.

For example, a 6-state collaborative study was conducted on Mid-Atlantic coastal plain streams to determine whether a combined reference condition could be established (Maxted et al. in review). In this study, a single method was applied to all sites in the coastal plain in all 6 states (New Jersey, Delaware, Maryland, Virginia, North Carolina, and South Carolina). The results indicated that two Bioregions exist for the coastal plain ecoregion—the northern portion, including coastal plain streams in New Jersey, Delaware, and Maryland; and the southern portion that includes Virginia, North Carolina, and South Carolina. In most situations, agencies have databases from well-established methods that differ in specific ways. The ability to combine unlike datasets has historically been a problem for scientific investigations. The usual practice has been to aggregate the data to the least common denominator and discard data that do not fit the criteria.

This Page Intentionally Left Blank