

Chapter 3.0 Study Design, Data Quality, and the Performance-Based Methods System

This chapter...

- reviews regional, site-specific, and gradient study designs
- describes methods for ensuring data quality objectives are met

Study design...

- consists of a variety of approaches depending on study questions
- is critical for an assessment program to meet its objectives and its data quality goals

3.1 Types of Study Designs

A variety of monitoring designs exist for biological assessment, all of which depend on the question(s) to be addressed. The major designs used for most assessment programs are commonly based on one of three objectives organized by spatial scale: regional assessments, site-specific assessments, and gradient studies.

In defining an objective, it is important to characterize the population of interest, whether it

includes all reaches or segments (stretches between tributaries of a given size) in a region, all hydrologic unit codes (HUCs) of a certain size, all segments below publicly owned treatment works (POTW) sites, or the segment below one specific discharge. Once the objective and population are defined, it is easier to select a study design. Important aspects of statistically powerful designs include the concepts of random sampling and sample allocation. Stratified random sampling reduces variances, allowing a more precise measure of the variable of interest, and is therefore ideal for statistical rigor. But not all levels of a study need employ random selection. The important point is to randomize at the level of the question. For example, if we are interested in the average number of invertebrate taxa found in rivers of a particular region using a specified protocol, taking 100 random samples from one river reach would not be an appropriate allocation of effort because only that one river reach would be characterized. It would be more appropriate to take one sample from each of 100 randomly selected river reaches throughout the basin because the river is the level of the question. Suggestions of pseudoreplication are relevant and potentially valid criticisms only in the context of the questions being asked. To avoid trying to answer a particular question with inappropriate data, it is advisable to work with statisticians or analytical staff familiar with study design and data analysis. The purpose of Section 3.1 and its subsections are to provide an overview of the different sampling designs that could be used for large river systems.

3.1.1 Sampling Effort and Sampling Reach Length

It is challenging to balance required sampling effort with available resources, while simultaneously maintaining focus on programmatic goals and objectives. While increased sampling effort can be justified for increases in precision, there often are substantial increases in the cost of sampling (Reynolds et al. 2003, Hughes et al. 2002, Lyons 1992b, Cao et al. 2001). As Angermeier and Smogor (1995) point out, comparisons of estimates based on insufficient sampling effort can be confounded because real differences in assemblage structure may be indistinguishable from method error. In a bioassessment context, this can translate to a

decreased ability to distinguish among sites of varying condition (Patton et al. 2000). However, identifying the most appropriate sampling effort in assemblage-level studies is often ignored.

A number of issues have emerged that are worthy of discussion regarding sampling reach length for non-wadeable rivers. On these systems, sampling reach lengths are generally larger than those in wadeable systems - a result of the scaling-up to accommodate the magnitude of the resource. The approach used can result in relatively long (i.e., kilometers) or short sampling reaches (< 1 kilometer). Long reaches may mask small scale habitat conditions and impairments that may be of interest to resource managers. They may also weaken the ability of the data to detect linkages between local river conditions and the drivers of those conditions. In designs where a long sampling reach is warranted, however, several small sub-reaches, and thus multiple data points, can be used to estimate spatial variability. Such short reach lengths highlight small scale conditions which may, simultaneously, reduce their utility for estimating broader-scale characteristics. Both perspectives are justified at times, and selection of the appropriate reach length for a study should depend on the questions being addressed by the study.

The development of a scientifically-sound sampling design for large rivers must include some discussion of the sampling effort to be exerted at a given sampling location and along the river (Lyons 1992b, Angermeier and Smogor 1995, Paller 1995, Peterson and Rabeni 1995, Patton et al. 2000, Cao et al. 2001, Cao et al. 2002, Hughes et al. 2002, Dauwalter and Pert 2003, Reynolds et al. 2003, Maret and Ott 2004, Fayram et al. 2005, Flotemersch and Blocksom 2005, Hughes and Herlihy [accepted]). Any description of sampling effort includes specifying the spatial scale over which the sample(s) will be collected (channel length), the amount and types of habitats that will be sampled within that length, and the field sampling method to be used (Reynolds et al. 2003). Further, the field sampling method is typically described by detailing gear, specific habitat types, intensity, and often, an estimated number of person-hours per sample (or site). Estimates and inferences regarding assemblage attributes (e.g., number of taxa, metrics, and IBI scores) are sensitive to sampling effort (Angermeier and Karr 1986, Angermeier and Smogor 1995, Rosenzweig 1995, Patton et al. 2000, Cao et al. 2002, Reynolds et al., 2003, Flotemersch and Blocksom 2005, Hughes and Herlihy [accepted]) because riverine habitat is heterogeneous with non-uniform distribution of organisms among habitat types (Angermeier and Smogor 1995). The number of taxa collected at a given site will, thus, increase with sampling effort, and will also vary with biogeography, sampling method and efficiency, behavior and abundance of the assemblage being sampled, and patchiness of the targeted habitat components.

Ideally, the sampling effort applied is the minimum that will allow stated objectives to be addressed as required by a study (Angermeier and Smogor 1995, Patton et al. 2000). As an example of how the question can influence the required effort, estimates of species' relative abundances have been shown to require less sampling effort for a given accuracy than estimates of the absolute number of species (Angermeier and Smogor 1995). For a bioassessment program, potential cost savings realized through the use of efficient sampling protocols translate to opportunities to enhance other aspects of a study design or program (Patton et al. 2000). This section will focus on issues related to definition of the appropriate sample unit for large river bioassessments. In other words, what is the channel length that will be sampled?

3.1.1.1 What is a Reach?

In a hierarchical context (Figure 2-4), Frissell et al. (1986) defined the word “reach” as a length of stream between breaks in channel slope, local side-slopes, valley floor width, riparian vegetation, and bank material. They further added that the reach is sometimes the least physically discrete unit in the hierarchy, but an exceedingly useful scale for describing medium- and long-term effects of human activities on streams. We use the term “sampling reach” to describe the site from which samples are collected. In linear systems, such as rivers, it is quantified as some channel length.

Many factors relevant to sampling reach length decisions in wadeable streams (e. g., Patton et al. 2000, Lyons 2002) will influence those same decisions in larger, non-wadeable rivers. Paller (1995) suggested that streams with low species richness may require greater reach length-to-width ratios (l:w) to attain precise estimates of maximum species richness (MSR). However, large Oregon rivers with low fish species richness required less sampling effort to attain MSR relative to rivers with a higher species richness (Cao et al. 2001, Hughes et al. 2002). Paller (1995) also found that the relative importance of sampling depth may depend on the behavior of individual species (e.g., substrate or open-water orientation), or upon width-to-depth (w:d) ratios. Many large rivers have an abundance of habitats supporting fish species that are difficult to efficiently sample (e. g., those associated with deep, turbid, or swift-moving waters or off-channel habitats); they can be more frequent in some regions of the country than others. For these kinds of rivers and species, Angermeier and Smogor (1995) found that greater sampling effort is necessary to attain and adequately characterize fish assemblage structure.

3.1.1.2 Approaches for Sampling Reach Length Determination

In most applications, the channel length over which data are collected is the same for physical habitat measures and biota. Exceptions to this would be measures that characterize the larger watershed of the reach, and water grab and phytoplankton samples collected at a single point in a reach. The logistical advantages to using the same reach length for multiple indicator parameters collected over the extent of the reach are clear, because the same persons can collect different data at the same place and time. However, variable reach lengths may be justified, depending on the indicator for which the sample(s) are being taken. For example, because biota move down and upriver, an argument could be made that the channel length over which physical and chemical habitat data are collected should exceed that over which assemblage information is collected.

Different approaches have been used for determining the channel length used for bioassessment of large rivers, most involving consideration of several factors including the question being addressed by the study, the level of resolution (precision and accuracy) required to address the question, and the statistical approach that will be used to analyze any resulting data. Just as critical is ensuring that sampling reach length is balanced with available resources. The following discussion is intended not as an exhaustive review of the topic, but as an overview with examples.

The reach lengths for most studies were set based on judgment, past history, or the need to match some other aspect of sampling or management activities. However, recent research has been conducted on the selection of sample reach lengths by evaluating the response of biological parameters (e.g., species accumulation curves, assemblage metrics; IBI scores) as a function of geomorphology (e.g. channel widths, meander wavelengths, riffle-pool sequences). Most of these studies have used fish assemblages (Gammon 1976, Lyons 1992, Meador et al 1993, Penczak and Mann 1993, Angermeier and Smogor 1995, Paller 1995, Yoder and Smith 1999, Patton et al. 2000, Cao et al. 2001, Lyons et al. 2001, Hughes et al. 2002, Reynolds et al. 2003, Maret and Ott 2004, Flotemersch and Blocksom 2005, Hughes and Herlihy [accepted]), although a few have used benthic macroinvertebrates (Bartsch et al. 1998, Li 2001, Poulton et al 2003, Flotemersch et al. 2006). Whether MSR of the local or regional fish assemblage, form of the final indicators (metric or index scores), or geomorphic characteristics should drive reach length determinations should depend on by programmatic considerations and the overall questions being addressed.

Biological Approach

The rationale for using biological measures for determining reach length is that in bioassessment we are, by definition, assessing the condition of biota. Therefore, the sampling effort required to produce reliable indicator results (metrics, indices) seems to be a logical determinant of reach length. In most cases, this question is addressed by over-sampling at a series of sites that cover the gradient of conditions to be included in a study and then determining the reach length for which the required data quality has been achieved. Reach length is then determined based on when a specified indicator asymptote is reached (Lyons 1992b, Angermeier and Smogor 1995, Paller 1995, Patton et al. 2000, Cao et al. 2001, Lyons 2001, Hughes et al. 2002, Reynolds et al. 2003, Maret and Ott 2004), when some level of similarity has been attained (Cao et al. 2001, 2002), or variability of that measure has been reduced to a desired level (e.g., Flotemersch et al. 2006, Hughes and Herlihy [accepted]).

Design specifics have varied among these studies, resulting in differing conclusions. Hughes et al. (2002) sampled 100 wetted channel widths, and through data analysis, determined that 85 channel widths were needed to collect 95% of the species obtained in 75% of the reaches sampled; collection of all fish species in a reach was calculated to require 300 channel widths on average. Those findings resulted in a field sampling design specification of 100x wetted width (Peck et al. [in press]). Hughes and Herlihy (accepted) determined that 50 channel widths were needed to obtain IBI scores exceeding those obtained from 100 channel widths less than 10% of the time. In contrast, Flotemersch and Blocksom (2005) examined the effect of reach length on the variability of IBI metrics from samples covering up to 2 km, and determined that at shallow river sites 1 km total shoreline shocked was sufficient for limiting the change in metric scores to 20%. Additional recommendations were provided for deep river sites. These three studies began with different reference conditions (100 channel widths vs 1000 meters), different maximum distances (100 channel widths vs 2 km), and different values for acceptable variability (5, 10, and 20%), and thus produced different results.

Physical Approach

Fixed Length vs Multiples of the Wetted Width (MWW)

Another difference among study results is how the final reach length is framed. Some studies propose reach lengths as a function of multiples of the wetted width (MWW) of the channel (e.g., Lazorchak et al. 2000, Hughes et al. 2002, Reynolds et al. 2003, Maret and Ott 2004, Peck et al. [in press]) while others support the use of a fixed distance (Flotemersch and Blocksom 2005).

The MWW approach follows the logic that as a system gets bigger, the effort required to sample the habitat components of the system at an equivalent level should increase proportionally. In other words, a fixed length of 500 m on a river 100 m wide could potentially miss or under-represent habitat components (such as bar, glide, pool, inside bend, outside bend that recur at longer intervals). One argument against this logic is that differing amounts of sampling effort are being applied across sites, by definition. A difficulty encountered with this approach in wide or impounded rivers is long reach lengths (e.g., 5 km for a 100 m wide river if 50 channel widths are the protocol). It is possible that pre-impoundment wetted width could be used in these cases (although the information is often not readily available), or that impoundments could be sampled like lakes.

Others have set reach length as a fixed distance (Flotemersch and Blocksom 2005) rather than as MWW. Proponents of a fixed distance endorse the ease of application in the field and utility in planning field activities (Patton et al. 2000, Lyons 2001, Flotemersch and Blocksom 2005). Opponents argue that using a fixed distance results in unequal sampling effort relative to river size, and that studies of fixed lengths have had lower data quality objectives regarding reference condition, maximum level of effort, and acceptable levels of variability.

A second argument against fixed lengths is that where the reaches do not encompass a sufficient number of habitat units, the biological differences detected may be due to differences among the habitat units of the sites. This becomes a greater concern as river width increases. For example, the Ohio River Valley Water Sanitation Commission (ORSANCO) conducts biological sampling on the Ohio River using 500 m reaches (<http://www.orsanco.org/watqual/aquatic/electro.asp>). The problem of the reach not including all habitats of a meander is addressed by the development and use of habitat specific criteria for soft, hard, and mixed bottom types. But such criteria ignore the often substantial diel migrations of larger fish species.

Meander Cycles

An alternate approach to using the response of biological parameters for setting reach length is to set it independent of the biology using the geomorphology of the system. This approach has its origins in work conducted by Leopold et al. (1964) who proposed that in meandering streams, 20 times the bankfull channel width typically encompasses at least one complete meander wavelength of the system. Because fluvial characteristics are repetitive and cyclical (Dunne and Leopold 1978), this distance should theoretically include all major habitat types within a given geomorphic reach and, by default, be available to all resident biota of those habitats. Given this,

the logic behind using geomorphic meanders as a basis for setting reach length for bioassessment is clear.

However, in altered large rivers, the identity, extent, and boundaries of habitat units of a meander are often non-distinct, obscured by turbidity or impoundments, or removed by anthropogenic alteration of the channel (straightening, armoring, and dredging). These conditions can render identification of a meander an impractical option for setting reach length and highlight the value of the finding by Leopold et al. (1964) that one meander roughly equates to 20 times the wetted channel width.

Following this guidance, NAWQA uses 20x wetted width, and sets a minimum length of 500 m (to help ensure representativeness of biological data), and a maximum of 1000 m (to minimize crew fatigue) (Fitzpatrick et al. 1998). However, such inconsistent levels of effort could potentially lead to difficulties in interpretation.

Ultimately, it is the quantity and quality of information required that will dictate the level of effort that can and should be expended at each sampling location. Thus, application of the data quality objectives process, including quantification of desired indicator performance, and testing of the capacity of sampling design to meet those objectives (both site specific and area wide), should drive the appropriate reach length.

3.1.2 Regional or Area-wide Assessments

In this document, regional assessments are defined as those that assess water resource quality across a broad region for status and trends monitoring. These studies are typical of designs used to meet the 305(b) reporting requirements under the CWA, and often result in estimates of the proportion of waterbodies in a certain condition (i.e., good, fair, or poor; or attaining and not attaining).

Representativeness is a critical factor given that the objective is to estimate a parameter (e.g., mean condition) from a subsample of a larger population (e.g., all large river reaches in the region). An important note with large rivers is that it may be possible to sample the entire population in some regions. For example, in more arid regions, there may be a limited number of large river segments. If the segment can be sufficiently characterized with a reach-based sample, it is conceivable that the entire population of segments can be sampled, allowing calculation of the absolute mean and variance for the population. Most often, however, the population of segments or sample units will be large, making a census impossible. Some inference of the average condition and variance will have to be made using randomized selection of sampling reaches (Larsen 1997, Urquhart et al. 1998). To reduce bias in the final estimate (e.g., percent of river miles impaired), probability-based designs for site selection are appropriate.

The first step in this type of design is to organize continuous, linear systems like rivers into representative units. For large rivers, this could be river segments, a standard hierarchical unit defined by lengths of rivers between tributaries of a given size. The second step is creating an approach for sampling these segments randomly. This might mean creating a list of “sample

units” (or list frame), applying a code to each unit, and randomly sampling them based on specified rules. Sample units could also be selected using a grid placed over a region, selecting grid cells at random, and sampling large river segments within them. This approach can also be used hierarchically, so that large grids (tier 1) are randomly selected and then small grids (tier 2) within tier 1 grids are randomly selected for sampling (two-stage sampling). A benefit of this approach is that not all rivers would have to be digitized beforehand, which can be costly if these data do not exist. Only those segments within selected tier 1 grids would need to be digitized (Rathbun 1999). However, this is an unlikely problem for large rivers given the availability of existing digital information for rivers throughout the USA (e.g., USEPA’s river reach file [RF3] coverage or USGS national hydrography dataset [NHD]). The EMAP program used a grid selection approach as part of its probabilistic design (Overton et al. 1991, Stevens 1997, Stevens and Olsen 1991, 1999).

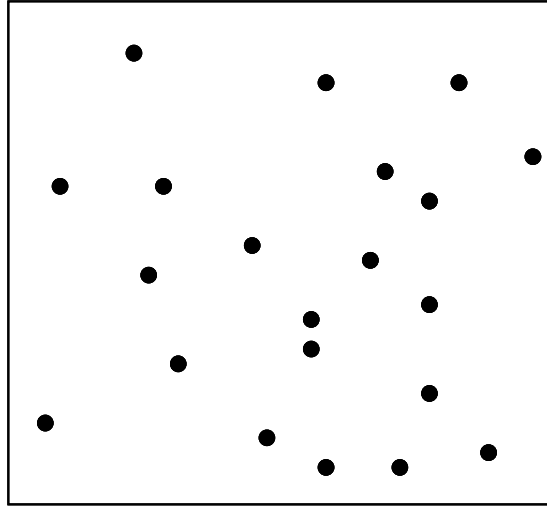
Although there are a variety of probability-based designs, only the simple random, stratified random, and systematic random approaches are discussed here. In *simple random* designs (Rathbun 1999), the entire pool of segments is the population, and sampling sites are selected randomly (Figure 3-1). This is the most basic probabilistic design. One drawback of this approach is that natural variability among sample units will increase the sample size needed to attain a given precision. Results of pilot studies should be used to determine appropriate methods, the level of precision (repeatability) a method is capable of, and, thus, how many samples are necessary to detect a desired change.

This natural variability can be partially controlled during sampling design by partitioning the region into strata based on underlying, scientifically defensible, natural classes (e.g., ecoregions or stream orders) using a *stratified random* design. The sample units are then selected randomly from these strata (Figure 3-1). The strata should be selected to maximize the differences *among* strata and minimize the differences *within* strata. By partitioning the natural variance among segments within strata, this design can achieve the same precision using a smaller sample size than a simple random sampling design, thereby reducing costs (Rathbun 1999). Sampling allocation may be made proportional to the size of the strata (e.g., if 10% of the segments are coastal plain, then 10% of the total sample effort would be randomly selected coastal plain segments) or can be apportioned based on the within-stratum variance, if known. However, at the very least, two sites are needed within any stratum to generate an average or variance estimate. Using too many strata could lead to poor variance estimates of the river overall, and thus, stratification should only be used with caution.

Systematic random is an approach for sampling site selection where the starting point (i.e., the first site) is selected at random, and those following lie at regular intervals. For example, the initial sampling location might be a 500-m segment with the midpoint at River KM 100. That point would have been randomly selected from within the 25-km distance encompassing the wadeable/non-wadeable transitional zone. Then, a reach midpoint would be located every 50 km downstream to the confluence with a channel of the same size or larger (or to tidal zone, or to estuary). Each sampling reach produces a random sample. Results from this design are used for estimating overall condition of the river system (as a mean value), or examining cumulative downstream effects. Additional information on different types of monitoring designs can be

found on the EMAP website:
(<http://www.epa.gov/nheerl/arm/designpages/design&analysis.htm>).

Simple random sampling



Stratified random sampling

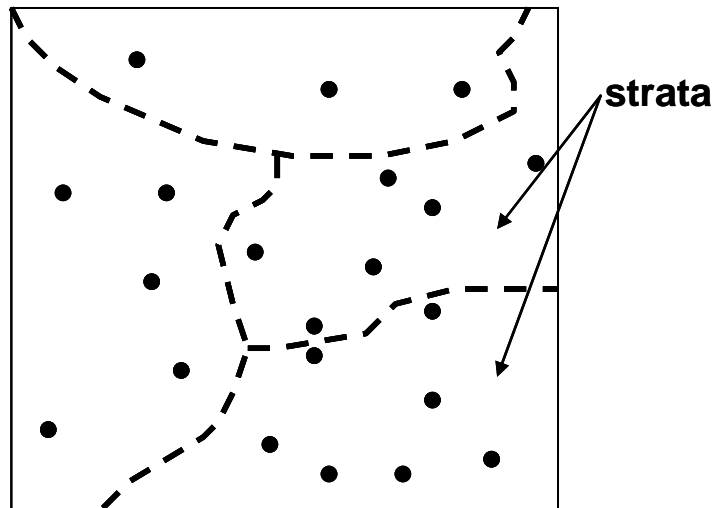


FIGURE 3-1. Examples of two-dimensional probabilistic sampling designs.

Quantifying trends in resource condition is often an important objective for regional assessments. Although there are different approaches for allocating sampling effort over time, only two are covered in this document: permanent station and serially alternating (Rathbun 1999). *Permanent station* approaches use a random sample of n sites that are all sampled during each time interval. This option provides the least spatial coverage but may provide the highest temporal resolution of trends, if temporal autocorrelation is weak. It is noteworthy that if resources allow sampling the entire population of large river segments, a permanent station temporal design is appropriate

then as well. *Serially alternating* designs (Table 3-1) or “rotating” designs partition the random sites within a stratum into sub-sets of sites that are sampled at regular intervals (e.g., every four years). This design was proposed for EMAP (Messer et al. 1991) and is the smaller scale probabilistic design used by the Maryland Biological Stream Survey (MBSS) for wadeable streams (Maryland Department of Natural Resources 1999).

TABLE 3-1. A serially alternating or rotating design for site sampling. In this example, all of the randomly selected sites are split into four sample sets. Sample sets would be serially sampled, such that each set is visited three times over 12 intervals (modified from Rathbun 1999).

		Sampling Interval (years, seasons, etc.)											
		1	2	3	4	5	6	7	8	9	10	11	12
Sample Set	1	X				X				X			
	2		X				X				X		
	3			X				X				X	
	4				X				X				X

3.1.3 Site-specific Assessments

Site-specific assessments focus on particular sites or small sets of sites, usually for the purpose of assessing the effects of a specific and known stressor source (e.g., effluent) or the effectiveness of a given intervention (e.g., restoration). Other site assessments may be performed for a unique question driven by a specific request (e.g., Is this segment of river comparable to the reference condition?). These objectives can be achieved through a variety of designs.

Traditionally, site-specific studies have been conducted using upstream vs downstream sampling, with a completely random selection of sampling locations some distance above and below the point of interest. One way of describing this is as a control-impact (CI) design. This sampling design is only able to compare the condition of the downstream reach to that of the upstream reach and use that as part of a weight-of-evidence argument for an impact. Drawing a conclusion that any effect is specifically due to the effluent is difficult because: 1) the effluent input is not replicated, and 2) since effluent pipes are not generally randomly placed, the local physical setting also likely influences the upstream and downstream conditions. As a result, it may be impossible to rule out other factors related to the upstream or downstream environment as responsible for observed differences. This effect can be reduced by comparing mean differences between the control and impacted sites to mean differences between comparable river segments without impacts. Samples through time can be used as replicates; but the impacted site would be pseudoreplicated, so there is only one true impact replicate (Hurlbert 1984). Still, some level of repeated sampling would improve weight-of-evidence arguments.

One option to reduce some of the limitations of CI analysis is to design a study to collect data prior to an impact and compare it to data collected after the impact begins. This design is referred to as a before-after (BA) design. The BA analysis requires a sufficient amount of before (pre-stressor or pre-effect) data so that the two sets of data can be analyzed as independent samples using two sample tests (t-test or analysis of variance) (Smith 2002). It is best to randomly assign the sampling dates to avoid systematic trend errors. As in the previous case, however, causal inference is problematic because observed trends may be due to climatic

differences or other natural events before and after the impact. Thus, this design does not have the controls that would account for natural, widespread changes. In addition, the impact may affect the variance structure rather than the mean, making detection difficult. Lastly, sufficient *before* data are often not available or bias exists because of *when* the sample was collected (sample timing), either of which would affect statistical power and inference. However, the BA approach could be used in building a weight-of-evidence assessment.

Incorporating a control site into the design of a BA approach provides some control of natural variability associated with time. In this design, data are also collected from the control site before and after the impact. It is best to randomly assign the sampling dates. The data are analyzed using a two-way ANOVA (BA and CI) with interaction (BA x CI), and the design is known as a before-after control-impact (BACI) design (Smith 2002). Such designs have been criticized because the sites are not randomly assigned and there is only one treatment area (Hurlbert 1984). One way around this statistical hurdle is to pair sampling at the control and impact areas and sample several times, resulting in a before-after control-impact paired design (BACIP) (Eberhardt 1976, Stewart-Oaten et al. 1986). The BACIP designs are treated much like a repeated-measures design with multiple times on one site instead of multiple treatment replicates. Each site-pair-time combination is treated as a unit. The ANOVA models in this analysis have BA, CI, sample time, and interaction (BA x CI) terms (Smith 2002). However, a simpler analysis of this design calculates differences for values collected at each site-pair-time unit and compares mean differences before and after the impact (Stewart-Oaten et al. 1986). This has also been called the paired BACI or BACI paired series approach (Smith 2002).

Variations on the BACI models have included increasing the number of randomly selected control sites (asymmetrical BACI design) (Underwood 1991, but see criticism from Stewart-Oaten and Bence 2001), including additional impact sites (Ellis and Schneider 1997), and using multivariate extensions of BACI (Faith et al. 1991, Kedwards et al. 1999).

3.1.4 Gradient Studies

The last class of designs discussed here are those that investigate the nature of the response to specific stressors. Rather than attempting to answer a yes-or-no question, these approaches investigate ecological response to gradients in stressor levels. The objective is to provide information to improve future management actions. An example would be to ask how biological condition changes in response to increasing urbanization density. Is the response linear or non-linear? Are there thresholds in the response? Such information can help land use planners manage future development differently. Another objective might be to define the response of a particular taxon to a known stressor. This information could be helpful in developing stressor tolerance values for taxa.

The main design approach in these studies is regression, where samples are collected along the entire gradient of the factor of interest (e.g., conductivity) and ecological response is measured. If pure hypothesis testing were the desired goal (e.g., do benthic IBIs respond to urbanization density?), then the levels of the independent variable should be controlled by the experimenter and all else left equal. This is not really possible for most assessment designs because there is rarely the opportunity to control land use intensity, but randomization schemes could be used to

reduce site selection biases. One factor that must be considered is that as samples are taken further downstream in a large river basin, the influence of small, degraded streams on downstream water resource quality is often masked due to the overwhelming differences in flows. However, if there is less interest in testing the hypothesis and more in defining or modeling the relationship, then the level of control on the independent variable is less important. It is easier to use existing gradients and to define the response with regression models.

Simple linear regression is used to define the response of one dependent variable (y-axis) to an independent variable (x-axis). Further, multiple linear regression is used to explore the response of a dependent variable to several independent variables (individual, transformed, or combinations of independent variables). With multiple linear regression, the relative effects of several potential explanatory variables can be examined simultaneously using a variety of approaches, setting a fixed multi-variable model, adding one predictor at a time, or starting with all of the independent variables and removing one at a time. In any case, the effectiveness of gradient designs depends on bracketing the gradient as well as possible. It is important to realize that certainty about responses is highest in the region where there are the most data (usually along the middle of a gradient), and lowest where data are least (usually at the extremes). This information must be extrapolated if there is interest in responses beyond the range of the gradient used to develop the models. Extrapolation is risky and any model should only be applied with great caution beyond the range of the independent variables used. With multiple regression using a number of transformed variables, this range is often difficult to identify.

3.2 Coordinating Sampling Design with Management Objectives

As discussed in Chapter 2, it is important to understand and present the specific questions, general goals, and potential uses for the assessment results; the DQOs that correspond to these goals (for the ultimate data user), and the quality of the measurement data that are necessary for the DQOs to be met (MQOs) (Figure 3-2). For biological monitoring and assessment practitioners, the following questions are common: *How healthy is the river? Is the river getting better? What is the condition of our watershed?* If there is general agreement that ecological indicators (in particular, multimetric indices of biological integrity [IBI, Karr et al. 1986]) and the ratio of observed to expected (O/E) taxonomic diversity (Wright 2000) provide the most appropriate information about overall water resource health or condition, then important decisions concerning the spatial placement of sampling sites and frequency of sampling.

Next, the study design process should allow specification of the spatial scale needed to address the objective: *Is the assessment intended to be for a single, particular river reach (e.g., 1 km, 10 km, the entire 2-km reach between two cities) or for all non-wadeable reaches within an entire watershed (at whatever scale the watershed might be defined)?* That is to say, is the objective to make defensible statements of condition for individual sites, for area-wide scales, or both?

Answering questions at area-wide scales requires aggregating multiple site-specific assessments to the scale of interest. However, if there is a probability-based component to the site-selection process (Stevens and Olsen 1991, Larsen 1997, Urquhart et al. 1998), then data can be used at

multiple spatial scales; from site-specific, to watershed-wide, to region-wide scales. Answers can be expressed in the following forms:

- The overall biological condition of River X at River Mile 27.14 is “fair” (IBI, 42 ± 9.4).
- The mean biological condition of non-wadeable river reaches in Watershed Y is “good” (IBI, $\bar{x} = 74$, $n = 12$, 90% CI = ± 7).

These questions can be answered in a credible and defensible manner, only if data of sufficient quality and quantity are collected. Once the data user settles on the types of questions s/he is asking (or is being asked) and the kinds of answers that would be satisfactory (e.g., with known and acceptable confidence), then data of the required power and sensitivity should be specified in the DQOs.

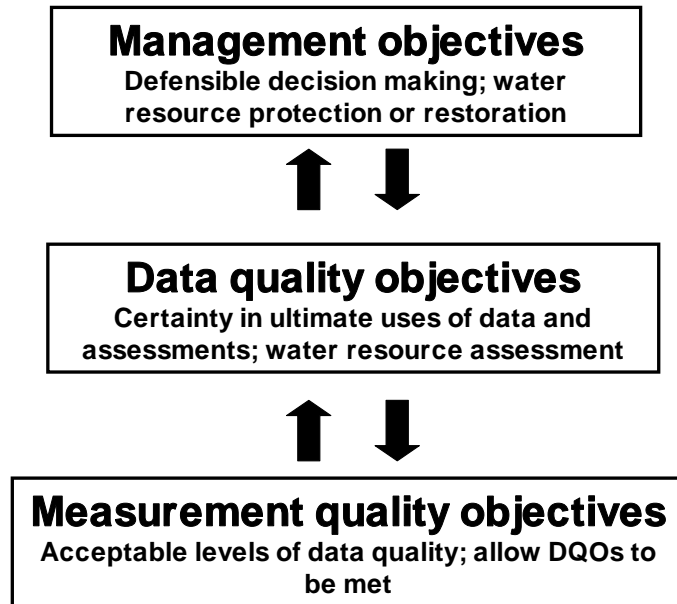


FIGURE 3-2. The relationship among management, data quality, and measurement quality objectives.

3.3 Data Quality Objectives

DQOs are statements of the level of uncertainty that a decision maker is willing to accept in decisions made on the basis of the measurement data (Smith et al. 1988, USEPA 2000b). An example DQO statement by a data user is:

This monitoring program should be able to detect a 20% change in mean biological index score (sensitivity), 80% of the time (power), with 95% confidence (certainty).

From this, or a similar statement, if there is a known or estimated precision value, a power analysis can be performed to help determine how many samples or sites are necessary to be able

to meet the stated DQOs (Osenberg et al. 1994, Urquhart et al. 1998). The greater the variance associated with an indicator, the larger will be the number of samples necessary to detect true change. Figure 3-3 presents the results of a power analysis, which show that 10 samples are necessary to be able to detect a 20% decrease in mean (\bar{x}) indicator value, with 95% confidence. How those 10 sites are arrayed throughout the landscape (or watershed) is dependent on the spatial scale of the question to be answered. For example, if one wants to have this level of data quality for three watersheds of different sizes, each watershed would need to be samples at 10 randomly-selected sites, regardless of its size. The key is to ensure that the locations are selected without bias.

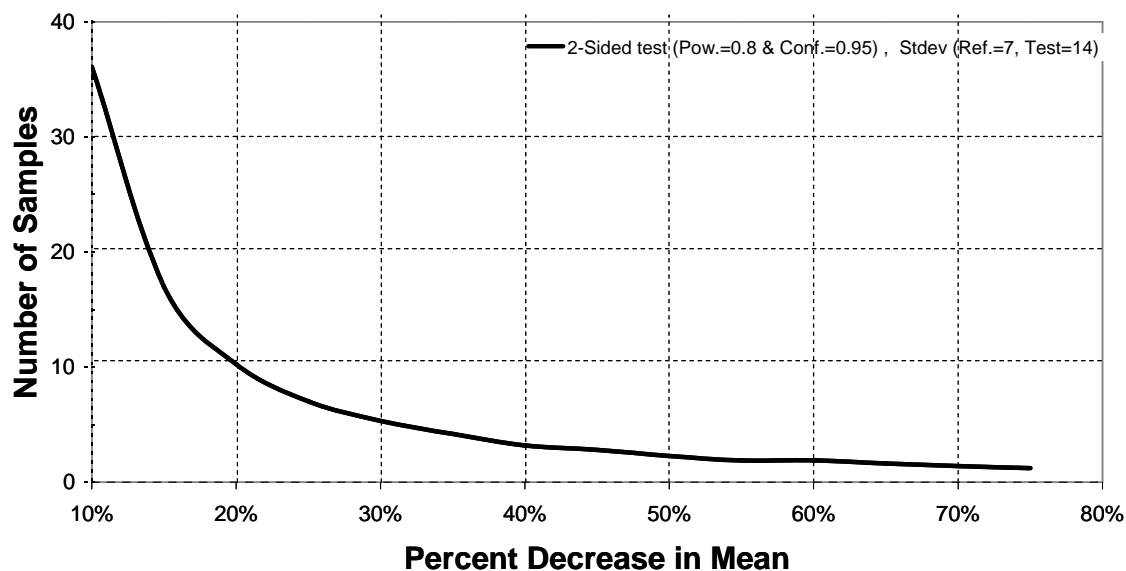


FIGURE 3-3. Results of power analysis showing the relationship between number of samples and the ability to detect differences (or changes) in mean index score (Stribling and Davie 2005). The index tested is multimetric and calibrated for Level 4 ecoregions in the Georgia Piedmont; benthic macroinvertebrate sampling methods are those of the Georgia Environmental Protection Division for wadeable streams.

If a certain level of data quality cannot be assured, then it is possible that the required DQOs cannot be met, with a resulting increased uncertainty (diminished defensibility) in addressing management objectives.

3.4 Measurement Quality Objectives and Performance Characteristics

Data quality is “the magnitude of error associated with a particular dataset” (Taylor, in Keith 1988). Overall error can be segregated into two types: random and systematic. Controlling error in datasets is necessary to ensure that reliable information is available to ecosystem managers and other decision makers. *Random error*, or variability, is error associated with natural variability; efforts to manage this kind of error are focused largely on sampling design such as by definition of temporal strata (e.g., seasonal index periods), stratification of sampling locations

(site classes), and randomized site selection (Figure 3-4). *Systematic*, or method, error results from how samples are taken and processed, and its control is largely through effective QA/QC. Although random and systematic error are often not completely independent (i.e., there is interaction between them in particular measurement systems), they do in some manner individually contribute to the overall variability of the final result. In fact, if some aspect of a sampling design is incorrect and gets implemented, data produced can exhibit substantial systematic error. However, it is possible to partition the potential error sources and use various control techniques to manage the error.

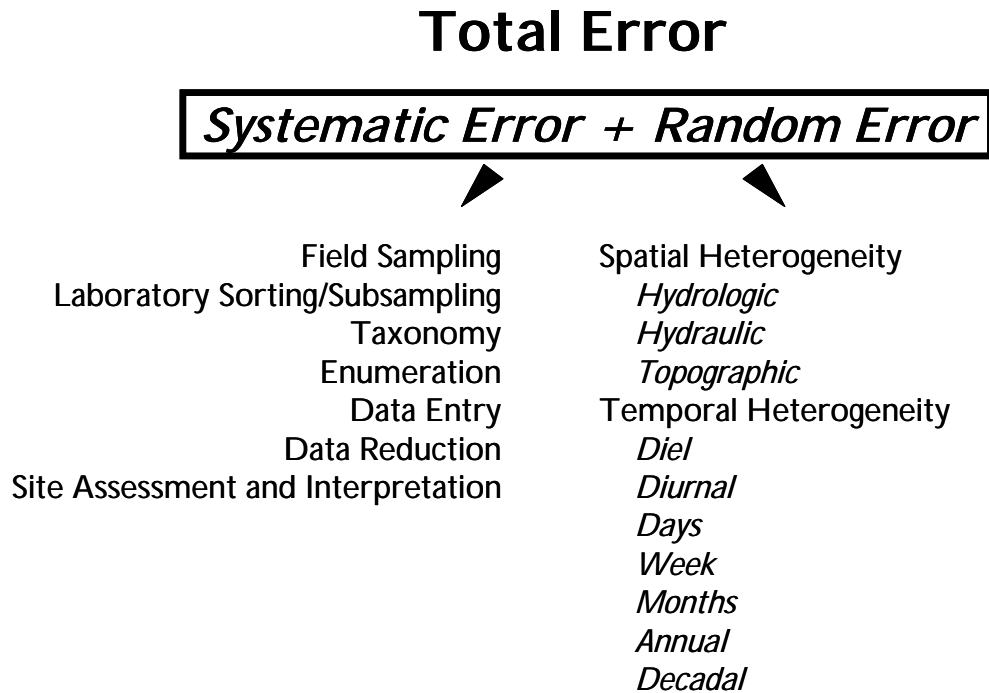


FIGURE 3-4. The overall variability of any measurement system results from both systematic error and random error. In biological assessment protocols, variability results from each step of the process and the spatial and temporal distribution of the samples.

An approach for ensuring that only data of known and acceptable quality are used is to establish and apply measurement quality objectives (MQOs). They can be established for any aspect of the biological assessment process, and MQOs may be quantitative or qualitative. Because a biological assessment protocol is a series of methods (Stribling et al. 2003), it is necessary to either describe the quality of data produced by each method or to assume sufficiency and acceptability. Different indicators require different activities to arrive at the endpoint. For example, the assessment process using the benthic macroinvertebrate assemblage is made up of at least seven methods or activities (see Chapter 6), and the quality of data and information produced by each can affect subsequent activities. Estimates of field sampling precision are directly affected by how the samples are processed (i.e., laboratory sorting, subsampling, and taxonomy). If laboratory activities are not performed at an acceptable level, any discussion of field precision may be meaningless. The magnitude of error that adversely affects a data user's

ultimate interpretation of an endpoint is likely unknown, or at least poorly understood. Routine documentation of data quality at each step of the bioassessment process improves defensibility of the end result. Acceptability of different rates and magnitudes of error is dependent on the needs of the data user. In the respective chapters on assemblage, components of the assessment process are segregated for purposes of defining performance characteristics.

PERFORMANCE CHARACTERISTICS DEFINITIONS

Precision – the nearness of two different measures of the same property (Taylor 1988, Taylor and Kuyatt 1994).

Accuracy – the nearness of a measurement to its true value, or analytical truth (Taylor 1988, Taylor and Kuyatt 1994, Clark and Whitfield 1994); the inverse of bias.

Bias – distance from a known value caused by systematically favoring some outcomes over others (Smith et al. 1988, Clark and Whitfield 1994); the inverse of accuracy.

Representativeness – that a value or entity depicts the property it is intended to depict.

Completeness – a measure of the number of valid data points relative to the planned number of data points (Smith et al. 1988).

Sensitivity – amount of change an indicator can detect relative to an independent variable (such as a disturbance gradient).

Although the importance of different performance characteristics should be determined by the ultimate data user, those data users should understand the potential error source interactions. The performance characteristics most commonly discussed are precision, accuracy, bias, representativeness, and completeness. Others which may be of importance and concern include selectivity and interferences, though they are often thought of as components of bias.

Individual performance characteristics are relevant to some components of the assessment process, but not to others because they may not be applicable. Further, some can be described quantitatively (QN) and others qualitatively (QL). Although there is differential rigor in how these aspects of data quality are communicated, and use of “na” may seem particularly trivial, it may be important. For example, it is important for non-specialists reviewing biological assessments to know that the concept of accuracy is not relevant to field sampling, while it is highly relevant to the final assessment of conditions. The analytical truth for benthic macroinvertebrate field sampling would be all organisms, in totality, present at a site. This value would be impossible to document, even with an enormous sampling effort. Table 3-2 presents formulas and explanations for quantitative performance characteristics. Documenting performance characteristics for a protocol or a program demonstrates the level of data quality that is achievable, and the quality of data associated with a program, project, or dataset.

TABLE 3-2. Formulas and explanations for quantitative performance characteristics.

Relative percent difference (RPD) – field sampling precision

This statistic represents the proportional difference between two measures and is calculated using the equation:

$$RPD = \left(\frac{|A - B|}{A + B} \times 2 \right) \times 100,$$

where A is the metric or index value of the first sample and B is the metric or index value of the second sample (Berger et al. 1996).

Root mean square error (RMSE) – field sampling precision

A kind of generalized standard deviation, this precision statistic is a pooled standard error for a set of k group means, usually associated with a one-way ANOVA, and is calculated by:

$$RMSE = \sqrt{\frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{\sum df_{1...k}}},$$

where y_{ij} is the i^{th} individual observation in group j , $j = 1 \dots k$ (Zar 1999).

Coefficient of variability (CV) – field sampling precision

This statistic is a unitless measure of precision calculated from the RMSE by:

$$CV = \frac{RMSE}{\bar{Y}} \times 100,$$

where \bar{Y} is the mean of the dependent variable (e.g., metric, index; Zar 1999).

Detectable difference (DD) – sensitivity of biological metrics, index, or O/E score

The detectable difference of the indicator defines the bracket around the observed mean (of metric, index, or O/E score) within which the true mean will be found with specified confidence, and thus, of the smallest difference between values that is significant. The implicit assumption here is that the frequency of repeat sampling is adequate to provide precision estimates representative of natural variability in the context of the method or protocol being used. Also, since the distribution is unknown, degrees of freedom (df) is set for an unlimited number of samples, or ∞ (Zar 1999). For a 90% detectable difference of a single observation (i.e., $\underline{p} = 0.10$), the RMSE value is multiplied by 1.64 (from a standard t -table, e.g., Zar [1999]):

$$DD_{90} = RMSE \times 1.64$$

for 95% detectable difference ($\underline{p} = 0.05$), the t -value multiplier is 1.96; and so on. With additional replicate samples, the detectable difference is divided by the square root of the number of replicates:

$$DD_{90}(2-tailed) = (RMSE \times 1.64) / \sqrt{n}$$

TABLE 3-2. (Continued)

Percent completeness (%C) – field sampling, taxonomy, site assessment and interpretation

Percent completeness is a measure of the number of valid samples that were obtained as a proportion of what was planned, and is calculated as:

$$\%C = \frac{v}{T} \times 100,$$

where v is the number of valid samples and T is the total number of planned samples. For percent taxonomic completeness, v is the number of specimens in a sample that were identified to the target taxonomic level and T is the total number of specimens in the sample.

Percent sorting efficiency (PSE) – sorting/subsampling bias

Percent sorting efficiency is calculated as:

$$PSE = \frac{A}{A + B} \times 100,$$

where A is the number of organisms found by the original sorter, and B is the number of missed organisms recovered (sort residue recoveries) by the QC laboratory sort checker.

Percent difference in enumeration (PDE) – taxonomic precision

Precision of sample counts is determined by calculating percent difference in enumeration by comparing results from two independent laboratories or taxonomists using the formula:

$$PDE = \frac{|n_1 - n_2|}{n_1 + n_2} \times 100,$$

Percent taxonomic disagreement (PTD) – taxonomic precision

Precision of taxonomic identifications is determined by calculating percent taxonomic disagreement by comparing genus-level taxonomic results from two independent taxonomists, using the formula:

$$PTD = \left[1 - \left(\frac{comp_{pos}}{N} \right) \right] \times 100,$$

where $comp_{pos}$ is the number of agreements and N is the total number of organisms in the larger of the two counts (Stribling et al. 2003).

TABLE 3-2. (Continued)

Discrimination efficiency (DE) – accuracy of site assessment and interpretation

The accuracy of the Index of Biological Integrity (IBI) and individual metrics is characterized as their capacity to correctly identify stressor conditions (physical, chemical, hydrologic, and land use/land cover) and is quantified as discrimination efficiency using the formula:

$$DE = \frac{a}{b} \times 100,$$

where a is the number of stressor sites identified as below some specified acceptance threshold, and b is the total number of stressor sites.

3.5 Performance-based Methods Systems

Performance-based methods systems (PBMS) require that acceptable data quality be defined relative to MQOs. Once MQOs are established, any protocol or program producing data meeting those acceptance criteria are acceptable for use. Using a PBMS enhances monitoring programs in that it:

- Provides the means to objectively screen data quality and quantify acceptable measurement error,
- Improves credibility and defensibility of biological assessments,
- Allows for communication of the data quality to secondary user(s), and
- Provides the necessary information for determining comparability among programs, protocols, methods, and data.

The PBMS (Figure 3-5) integrates decisions on the acceptability of data quality with their utility for management decisions (see the website for the Methods and Data Comparability Board of the National Water Quality Monitoring Council (NWQMC) (<http://acwi.gov/methods/>) for more information on PBMS). If performance characteristics are documented for one program or dataset, it looks similar to what should be routine QA/QC. If documented for two, determination of comparability between the two programs is relatively straightforward (Figure 3-6).

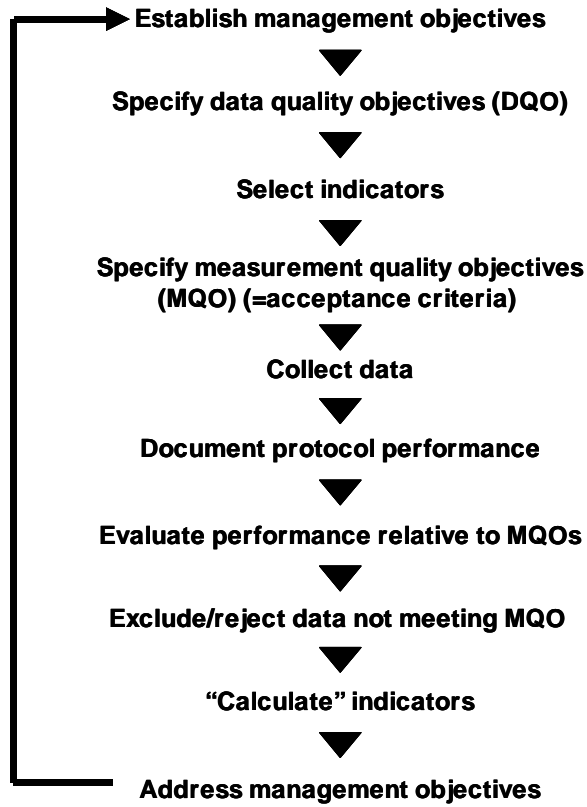


FIGURE 3-5. Use of MQOs and performance characteristics to ensure defensibility of management decisions (USEPA in preparation).

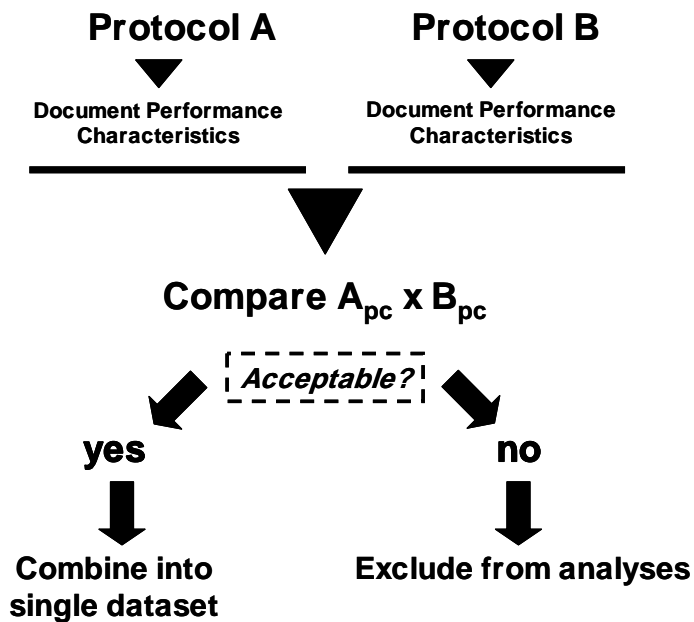


FIGURE 3-6. Framework for analyzing the comparability of multiple biological assessment protocols.