Chapter Three
Study Design

3.1. Introduction

Once the monitoring team has accepted a conceptual model and defined the objectives of the monitoring program, the next stage involves general decisions about a more detailed design that also specifies data requirements. This is a fundamental stage that ensures that the sampling and analysis programs are cost-effective. It takes place before sample collection starts, and again involves interaction with the end-users of the information.

Figure 3.1 shows a framework for undertaking monitoring program design, and Table 3.1 is a checklist. The case studies in Appendix 4 illustrate the concepts of this chapter.

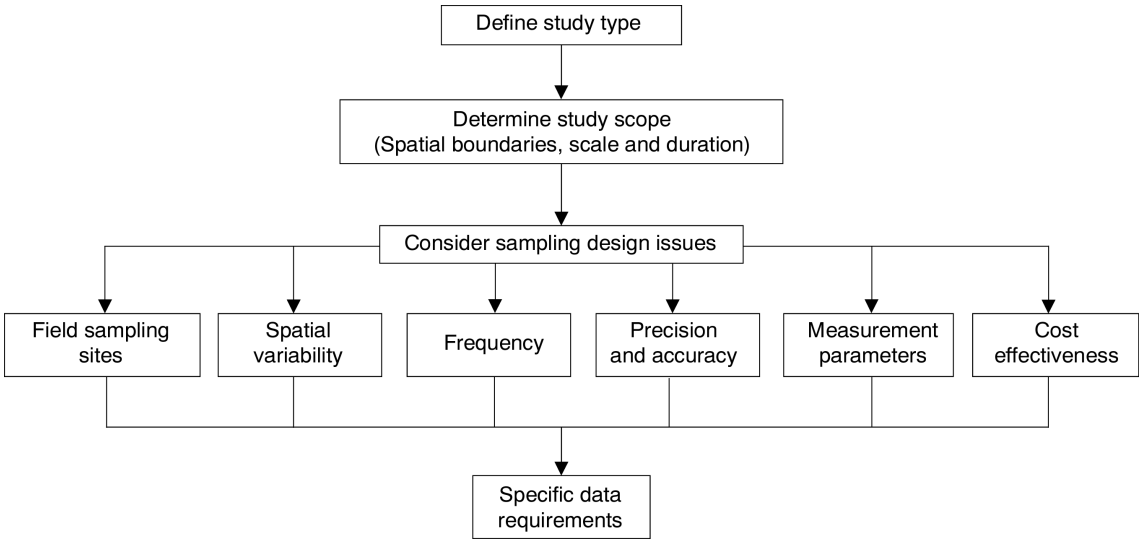

Figure 3.1. Framework for designing a monitoring study

3.2. Study Type

First the monitoring team must decide on the study type, because this will define the field sampling program and the path taken for subsequent data analyses. Three distinct study types can be identified:

- descriptive studies;
- studies that measure change;
- studies that improve system understanding (cause and effect).

Table 3.1. Checklist for designing a monitoring study

1.	Has the study type been made explicit and agreed upon?
2.	Have the spatial boundaries of the study been defined?
3.	Has the scale of the study been agreed to?
4.	Has the duration of the study been defined?
5.	Have the potential sources of variability been identified?
6.	Are there sufficient sampling stations to accommodate variability?
7.	Are the sites accessible and safe?
8.	Can sites be accurately identified?
9.	Has spatial variation in sites been considered, and have options to minimise this variation been considered?
10.	On what basis is the frequency of sampling proposed?
11.	Have decisions been made about the smallest differences or changes that need to be detected?
12.	Is replication adequate to obtain the desired level of precision in the data?
13.	Have the measurement parameters been chosen?
	(a) Are they relevant?
	(b) Do they have explanatory power?
	(c) Can they be used to detect changes and trends?
	(d) Can they be measured in a reliable, reproducible and cost-effective way?
	(e) Are the parameters appropriate for the time and spatial scales of the study?
14.	Has the cost-effectiveness of the study design been examined?
15.	Have the data requirements been summarised?

3.2.1. Descriptive Studies

Descriptive studies gather data to document the state of a system. They are the most basic of monitoring exercises. Typically they measure the spatial and sometimes temporal distributions of constituents within a water body for the purpose of (i) reconnaissance surveys, (ii) State of the Environment reporting, or (iii) assessing conformity with water quality guidelines or other agreed guidelines. They can determine background or baseline concentrations when a disturbance or development either is not expected or has not happened yet (baseline studies). In sediments, they can also identify changes that have occurred due to some earlier (historical) disturbance.

If the focus of the study has been descriptive, it is not usually possible to analyse the data subsequently to demonstrate causality. The need for this should therefore be determined in advance.

3.2.1.1. Baseline Studies

In baseline designs, no disturbance has occurred. An example of baseline designs is provided by some of the long-term water quality network programs that mainly monitor physical and chemical measurement parameters. Such programs are maintained so that they can detect or document any completely unanticipated changes in water quality. In these cases it is best to decide which measurement parameters to monitor, and the directions and sizes of changes or trends that would be important in those parameters (Green 1979; ANZECC & ARMCANZ 2000). When the monitoring team knows what changes to expect in the measurement parameter, it can refine the sampling design to avoid two very common pitfalls: either collecting insufficient data to detect the trend or change reliably, or collecting so much or such inappropriate data that ecologically trivial changes are detected. Well-designed baseline studies for which the likely nature of the disturbance can be anticipated are a prerequisite of the strong designs in section 3.2.2.

Baseline studies of sediments offer an opportunity to see the effects of historical changes in sediment contaminants, and can be used to establish the magnitude and perhaps timing of some past disturbances. The scientists and statisticians involved use their skills and insights to assemble information from appropriate studies. An example is the research program to detect increases in mercury levels in the Great Lakes (Green 1979). Because such studies are necessarily situation-specific, it is impossible to be prescriptive about the designs involved in them beyond noting that several independent lines of evidence strengthen any inferences about the effects of the disturbance.

3.2.2. Studies that Measure Change

When descriptive monitoring studies are repeated several times at the same locations, they can assess change. Such studies require relatively detailed planning so that locations can be identified and resampled. Data analyses can range from comparatively easy measurements of trends and simple correlations, to more complex evaluations that show if there has been a change of measurable significance. These are described in Chapter 6 and Appendix 5.

Monitoring is often done with the objective of evaluating the effects of a particular input or disturbance. If the timing and location of the disturbance are known, three categories of design are applicable (modified after Green 1979):

(i) *before–after, control–impact (BACI) designs*. Before the supposed effect occurs, two types of site can be identified: those that will be subjected to the disturbance and those that will not. The same parameter is monitored at both types of site before and after the disturbance to determine whether or not its *pattern of behaviour* over time at the disturbed site(s) changes relative to the control sites. After the disturbance starts, if the parameter's pattern of behaviour in the affected area(s) differs from its pattern of behaviour in the control areas, the differences are relatively unlikely to be due to chance.

(ii) *inference from change over time*. In this category of designs, no unaffected control site exists and change in a parameter can only be detected by comparison of data from one or more sites before and after the disturbance. In this design it is as likely as not that the change has occurred naturally over time, independently of the disturbance.

(iii) *inference from change over space.* In this category of designs, there are either unaffected control sites or there are sites that have been affected to varying degrees by the disturbance, but there are no valid comparable data collected before the disturbance. Sites used for the comparison may be upstream of the disturbed site, or on unaffected tributaries in river systems or estuaries, or in adjacent water bodies (e.g. wetlands, freshwater or saline lakes), or they may be distributed along some disturbance gradient (e.g. increasing distance from a point source). In such studies it is as likely as not that the values of the parameter in the disturbed and less disturbed sites differed before the disturbance.

Inferences should not be based solely on changes over time or changes over space unless there are no valid control sites or pre-disturbance data. Suitable spatial or temporal controls should always be used if they are available.

3.2.2.1. BACI Designs

The BACI designs have evolved in response to the common observation that the values of measurement parameters often differ naturally between any two ostensibly identical sites. The strongest versions of these designs base their inferences on interaction terms in a statistical analysis rather than on simple comparisons of means between sites. The logic of this procedure is best demonstrated first by discussion of Green's (1979) formulation of a BACI design, which is now regarded as the weakest of all the BACI designs, and then by an outline of subsequent improvements to the basic scheme.

Green (1979) proposed that environmental change would be detected if a measurement parameter were sampled from two separate sites, once before and once after a disturbance (Figure 3.2). One of

the sites would be the impact site (the site that would be subjected to the disturbance and potentially affected by it). The other site would be the control site, which would be similar in all relevant respects to the impact site except that it would not be subject to the disturbance. The sites would be chosen so that they were independent of each other with respect to the measurement parameter. If the impact site were affected by the disturbance, then, Green argued, this would be apparent in a significant interaction term in an analysis of variance (where the factors in the analysis would be 'time' with two levels, 'before' and 'after', and 'site' with two levels, 'control' and 'impact'). In graphical terms (Figure 3.2) the behaviour of the 'impact' site would change relative to the behaviour of the 'control' site after the disturbance. The values of the measurement parameter would not have to be identical in the two sites before the disturbance because the inference would be based on the interaction term in the analysis.

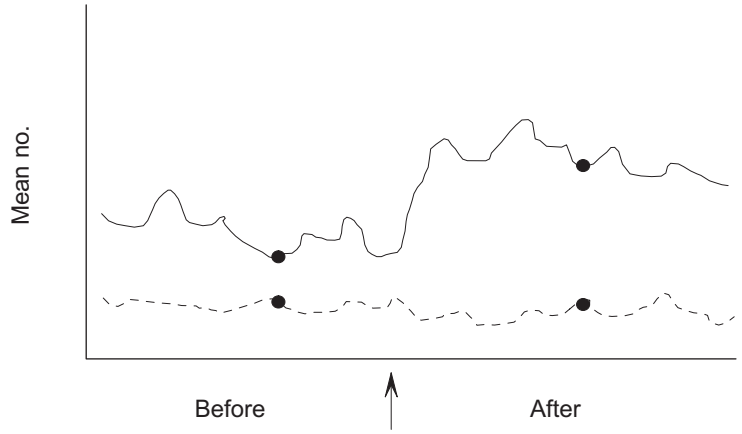

Figure 3.2. Illustration of a BACI design with single sampling events (denoted by dots) before and after a disturbance (arrow) in both a control site (dashed line) and an affected site (solid line) (modified from Underwood 1996)

Although Green's (1979) scheme was an important conceptual advance for environmental scientists, the notion of basing the inference of change on single sampling events from single sites of each type was criticised. The inference would be based exclusively on subsampling within each combination of site type and time (Hurlbert 1984; Stewart-Oaten et al. 1986); another site-specific disturbance event, unrelated to that being monitored, could confound the conclusions from such a design.

The preferred approach to circumvent this problem is to monitor more than one control site and to use multiple sampling events before and after the disturbance, as in the so-called MBACI designs (Keough and Mapstone 1995; Underwood 1996). The scheme is illustrated diagramatically in Figure 3.3. There are a number of important choices that need to be discussed when designing such programs, including the locations of sites, the number of 'before' sampling events, and the sampling effort required to model trends and dependencies through time. Although much of the literature about these designs focuses on analysis of variance, other statistical procedures (e.g. generalised linear models, see Appendix section A5.1.9) may be more appropriate and flexible for handling data that are not normally distributed. The data requirements of such procedures need to be discussed with a statistician before any data are collected.

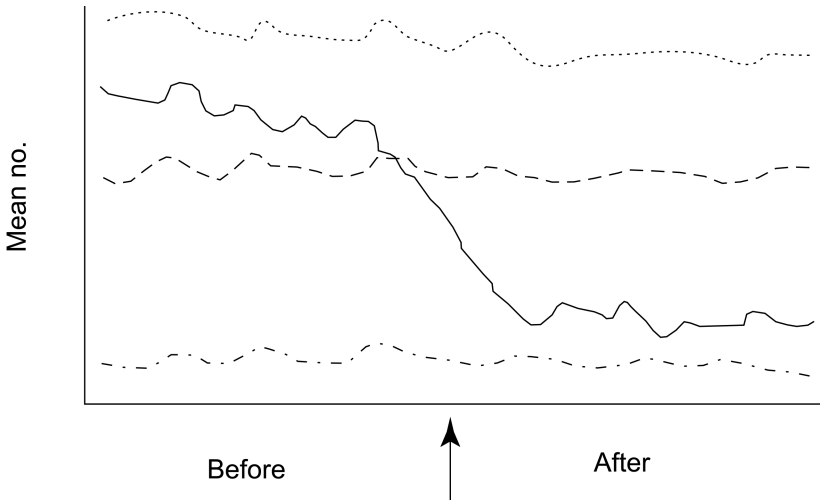

Figure 3.3. Illustration of Underwood's modified BACI design in which multiple random samples are taken before and after a disturbance (arrow) from three control sites (dashed lines) and an affected site (solid line) (modified from Underwood 1996)

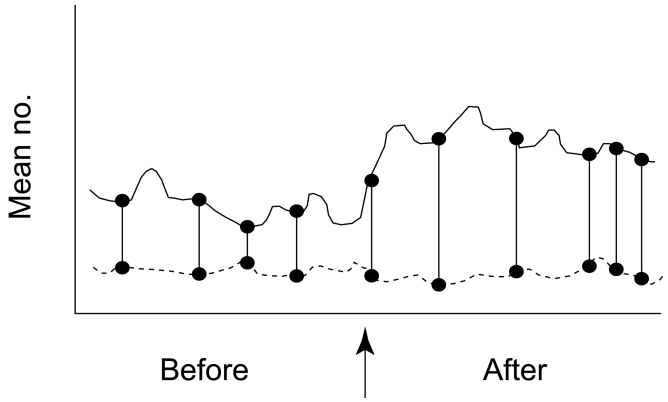

Figure 3.4. Illustration of a BACI design with multiple random samples taken before and after a disturbance (arrow) in both a control site (dashed line) and an affected site (solid line) (after Underwood 1996)

A number of variants of BACI designs have been proposed, and these are more fully discussed by their authors (Stewart-Oaten et al. 1986; Underwood 1991, 1992, 1994; Keough and Mapstone 1995, 1997). One commonly promoted variant applies to situations where there is a pair of sites: a single 'control' and a single 'impact' site sampled on many occasions before and after the disturbance (called BACIP by Stewart-Oaten et al. 1986) (Figure 3.4). For the inference to be strong from this design, the sites must be closely matched and some restrictive assumptions must be applied to the behaviour of the measurement parameter in the two sites. For example, if the measurement parameter were the abundance of fish, it would be presumptuous to assume that the population patterns and dynamics would necessarily be identical at the two sites. Thus, this approach should be used if only a single control site can be found, because localised site-specific events unrelated to the disturbance of interest can become confounded with the effect of most interest. Osenberg and Schmitt (1994) describe salutary examples of the problems inherent in these designs, for a marine system. The term randomised intervention analysis (RIA) has also been applied to this type of design (Carpenter et al. 1989).

3.2.2.2. Inference from Change Over Time

In some circumstances there are no suitable control areas, and so changes associated with a disturbance can only be inferred by comparing post-disturbance data with pre-disturbance data collected from the one site. Because there are no spatial controls, there is a chance that an unrelated disturbance may have coincided with the disturbance that is being monitored or assessed.

The main statistical procedures that can be used to analyse such data include (but are not restricted to) regression, trend, and time-series analyses. Sometimes the term 'intervention analysis' is used when time-series analysis is applied to a defined disturbance (e.g. Welsh and Stewart 1989). These procedures constitute a large and complicated area of applied statistics. Although some of the robust alternatives to the classical techniques are sketched in Appendix 5, it is beyond the scope of the Monitoring Guidelines to discuss them in detail. Expert statistical advice should be sought when planning and analysing these data, and particular attention should be paid to the modelling of interdependencies between successive sampling events and to choosing sampling intervals that are appropriate to the disturbance being monitored or assessed (e.g. Millard et al. 1985).

Often, these statistical procedures require data from a large number of sampling events and are most applicable to measurements of physical and chemical parameters (e.g. Welsh and Stewart 1989), although biological measurement parameters have been used in such designs (e.g. fish ventilation by Thompson et al. 1982). For such long-term designs, particular attention needs to be paid to coping with irregular sampling intervals and the inevitable missing data because classical statistical techniques are sensitive to both these occurrences (e.g. Galpin and Basson 1990).

3.2.2.3. Inference from Change Over Space

Often disturbances have already occurred or are alleged to have occurred, and scientists are required to judge the severity of impact or monitor the situation, either to assess whether recovery is occurring or to assess the success of remedial actions. Because such studies have no useful *pre-disturbance* data, inferences about the disturbance rely on spatial patterns. These patterns are found either in contrasts between disturbed and undisturbed sites or in sites chosen to represent a gradient of disturbance. The disadvantage of this class of design is that the observed pattern may be confounded with other environmental changes that are not related to the disturbance being monitored or assessed.

In rivers, to monitor recovery or dilution of the measurement parameter, it has been fairly common to select and sample a control site upstream of a disturbance and a series of sites downstream of the disturbance. Although this design is intuitively appealing, it has two problems. First, if sites are close together there may be inter-correlation between them that may mask changes (see also section 3.4.2). Second, there may be considerable natural variation in the measurement parameter that may not be captured in a single control site; therefore differences between the control and disturbed sites may not be due solely to the disturbance itself.

Multiple control sites, if they can be found, provide a stronger basis for inferring impacts resulting from a disturbance. A specialist statistician should be asked to clarify whether sites can be chosen to satisfy the assumptions of the analysis, or whether sufficient data are being collected to identify any spatial intercorrelations between the sites and allow valid inferences to be drawn.

Sometimes, it is not possible to find control sites that are undisturbed but resemble the disturbed site in all other important respects. Instead, *reference sites* are identified that are deemed to represent standards. Then values of the chosen measurement parameters at the disturbed site(s) are compared with values of the same parameters at the reference sites. This approach has been used for macroinvertebrate community structure in the AUSRIVAS procedure, and is explained in Box 1 (page 3-24). Alternatively, a gradient of disturbance can be identified either within the area surrounding the disturbed site (e.g. the seabed surrounding an oilrig) or across a number of sites across the landscape (e.g. a series of wetlands along a salinity gradient).

Gradients of disturbance — that is, values of measurement parameters that increase or decrease with distance from a point or boundary — are spatial patterns that are poorly described by classical

statistical techniques such as ANOVA (analysis of variance) and regression. Spatial statistical tools (e.g. Cressie 1993) should be more appropriate for describing these spatial patterns; for example, concentrations of toxicants in sediments, or abundance of species of benthic animals or plants. The classical spatial statistical tools can require very large areas to be sampled, and very large numbers of samples (Rossi et al. 1992).

Spatial analyses of data from sites that lie along a gradient of disturbance are sometimes termed *gradient analyses*. In these, some independent measure or surrogate of the disturbance (e.g. distance from source) is correlated with values of the biological measurement parameter. When an aspect of community structure is being measured, multivariate techniques such as ordination and clustering are used to relate the biological pattern to the spatial pattern (e.g. Warwick and Clarke 1993). These techniques are rapidly evolving, and novel ways are being developed to quantify the relationships between multivariate biological responses (as expressed by dissimilarity measures) and spatial patterns.

3.2.3. Studies for System Understanding

Some studies are made with the aim of finding out more about a particular system; for example, to better understand aquatic ecosystems and the physical, chemical and biological processes that operate in them. A deeper understanding may reveal relationships among the variables operating in the system, enabling predictions to be made about the behaviour of the system in situations beyond existing data and experience.

If the objective of a study is to establish cause and effect relationships, the sampling program must be designed for this purpose from the start. For this objective, the monitoring team may need to run additional experimental studies in which they can manipulate the system in a controlled manner and measure the system's response. In this case the sampling regime must be designed so that at least one of the potential outcomes is unequivocal. Manipulative experiments are routinely conducted in laboratories, but in the field they can be expensive and it may be impossible to control all the confounding variables adequately.

In studies of cause and effect, even the best experimental or survey design may be insufficient by itself. No design can completely defend against all unidentified confounding influences (Stewart-Oaten et al. 1986; Eberhardt and Thomas 1991; Underwood 1994). To establish cause and effect, therefore, the monitoring team must assemble independent lines of evidence and circumvent the potential for inferential problems akin to those faced by epidemiologists. Beyers (1998) has attempted to combine epidemiological criteria (Hill 1965) with postulates for environmental toxicology (Suter 1993) (Table 3.2). Not all of these criteria need to be met, but strength, consistency and specificity provide the strongest evidence for causation. Where a disturbance is chemical, indicators of exposure (e.g. contaminant concentrations in tissues) also provide strong evidence for causation. Whether Beyers's emphases are appropriate or not is likely to be debated as investigators try to formalise the ways in which they combine evidence in environmental studies.

It should be noted that the results from a study that measures change also contribute to system understanding by demonstrating a link between a particular human activity and a specified effect in the system under consideration. However, they do not establish cause and effect. Some other unknown cause may have resulted in the effect. To establish a cause–effect relationship some characteristic of the activity needs to be linked to the change observed.

However, studies for improving system understanding are not always done to show cause and effect. For example, the conceptual process models in Figure 2.4 and Figure 2.5 outline a system understanding with respect to nutrients and copper, respectively. A study could monitor the changing significance of processes in these models, over space and time.

3.3. Scope of a Study

When it defined the objective of the monitoring study (see Chapter 2), the monitoring team would have identified the study site in broad terms, e.g. the River Murray, or Sydney Harbour, or the Brisbane River and Moreton Bay. Now the monitoring team can set the spatial boundaries of the study and consider questions of scale and duration.

3.3.1. Spatial Boundaries

The setting of spatial boundaries is important because inappropriate boundaries might focus the study away from important driving or consequential factors. This decision must be based on the issue of concern and the ecosystem rather than on convenience and budgets. In an investigation of effects of catchment activities on rivers, lakes and estuaries, for example, the spatial boundaries would normally be those of the catchment.

As an example, for a study of the Brisbane River and Moreton Bay, the monitoring team might consider if the study should look only at these water bodies, or if it should include tributary creeks or the broader catchment, or if the study should be restricted to the major receiving waters.

The pertinent point here is that the monitoring team needs to explain the logic behind its decisions with respect to the spatial boundaries of the study.

3.3.2. Scale

Scale refers to the spatial and temporal ranges over which a system is observed, i.e. the appropriate level of resolution to answer the questions of concern. Different processes operate at different scales. For example, the movement of sediment in a river system may take tens of years at the catchment scale, toxicant effects may occur over days and may be localised, while nutrient enrichment may occur over kilometres and the response may take weeks.

The scale of the study should be chosen in relation to the study's objectives after the monitoring team has considered the measurement opportunities at the various possible scales and the likelihood of collecting reliable and valid measurements. The cost of data collection at the various scales should also be assessed.

Will the necessary measurement parameters be spatially uniform? As the spatial extent of data collection gets larger, so the distribution of the measurement parameters can become more heterogeneous and patchy, and more replicate samples can be needed to achieve the same confidence in the results. It is essential to choose an appropriate scale relative to the phenomenon under consideration and then sample at that scale.

3.3.3. Study Duration

Similar problems affect the decision about how long a study should last to address the issue of concern. Given the variability of natural rainfall and hence streamflow, what length of time might be required to achieve an appropriate understanding of the system?

The appropriate length of the study is an important decision. Few hydrologists would make definitive statements on the quantity of water resources with data from only two or three years, yet frequently conclusions from water quality studies are expected from less than three years' data. What is a reasonable duration for the study? How long will it take for a sufficient variety of rainfall events (from droughts to floods) to be experienced to allow the monitoring team to study the system under extremes?

Table 3.2. Criteria to formalise the use of independent lines of evidence in inferring causation in impact studies[a]

Name of criterion	Description of criterion	Example[b]
Strength of association	Size of the correlation between the intensity of the disturbance and the response of the measurement parameter	Sites with high concentrations of the toxicant have lower population densities of an organism than sites with low concentrations of the toxicant
Consistency of association	The association between the disturbance and the measurement parameter has been repeatedly observed in different places, circumstances, and times	The negative correlation between concentrations of the toxicant and the densities of the organism has been demonstrated in several other studies by other investigators elsewhere
Specificity of association	The observed effect is diagnostic of exposure to the disturbance	In this case, a decrease in density of the organism is not diagnostic of the disturbance because the population density of this organism may be reduced by other, natural, processes
Presence of stressor in tissues	Measurement parameters of exposure (e.g. residues, breakdown products) must be present in the tissues of affected organisms	Breakdown products of the toxicant are found in the tissues of organisms in sites with high exposure, but are below detection limits in sites where the toxicant is absent
Timing	Exposure to the disturbance must precede the effect in time	Accidental spillages of the toxicant are usually followed by sharp declines in the density of the organism
Biological gradient	A dose–response relationship exists (i.e. response of measurement parameter is a function of increases in magnitude of disturbance)	Laboratory toxicology tests have established a dose–response relationship
Biological plausibility	There is a biologically plausible explanation for causality, even if the precise mechanism is unknown	The toxicant comes from a group of chemicals known to interfere with respiration in this organism
Coherence	The causal interpretation should not seriously conflict with existing knowledge about the natural history of the organism and the behaviour of any substances associated with the disturbance	The organism is usually common in sites within the study region and is present year-round; the toxicant is readily soluble and does not breakdown readily while in solution
Experimental evidence	A valid experiment provides strong evidence of causation	A field experiment demonstrated rapid mortality in response to the addition of known concentrations of the toxicant
Analogy	Similar disturbances cause similar effects	Other chemicals related to this toxicant have shown similar dose–response curves and responses in field experiments with different but related species

[a] From Beyers (1998)

[b] A hypothetical example of the response of biological measurement parameters to a toxicant, as an illustration

3.4. Sampling Design

Environmental heterogeneity, both temporal and spatial, is probably the most significant aspect to be considered in the design of sampling programs (Eberhardt 1978; Morin et al. 1987; Kerekes and Freedman 1989). Variability will determine the number of sites, number of replicates and the frequency of sample collection. High environmental variability combined with logistic and financial constraints on sample collection and analysis often result in data that are too variable to reveal an impact, disturbance or trend.

Before a field sampling program can be planned, some idea of the expected spatial and temporal variability of the measurement parameters is needed. Often, general information about variability can be obtained from published work during the formulation of the conceptual process model of the system. For example, oxygen concentrations are known to vary diurnally and to differ between the epiliminon and hypoliminon in lakes; phosphorus, bound to sediment, is known to be transported during rainfall events. Typical types of variation are caused by:

- spatial variability because the environment is heterogeneous;
- time dependence, temporal, seasonal effects;
- disruptive processes;
- dispersion of chemical contaminants.

Normally, the design of any investigation, and particularly a monitoring program that is to be ongoing, cannot be settled without a pilot study. This short period of intensive monitoring outlines the nature of the prevailing system and particularly its temporal and spatial variability. Then the monitoring team can choose a sampling regime and frequency that should provide a representative profile of the system for each measurement parameter and piece of information required. It can decide on appropriate numbers of replicate samples to provide the precision required for the statistical analyses used in the study.

The pattern of sampling in space and time is of critical importance. Although most statistical techniques require random sampling, simple random sampling is sometimes difficult to achieve, and may not be cost- efficient. The main patterns of sampling are outlined briefly here, but they are more fully discussed in basic texts such as Cochran (1977). A major problem during sampling is representativeness. The intellectual challenge is to design a sampling approach that minimises errors. The errors in accurately representing a water body or population by a sample, and a sample by a sub-sample, can far exceed errors in analysis (Gy 1986). These and other statistical sampling issues are reviewed by Helsel and Hirsch (1992).

3.4.1. Patterns of Sampling

3.4.1.1. Simple Random Sampling

The basic requirement of most statistical procedures is that each sample unit in the population of interest has an equal probability of being selected and included in the sample. There should be no conscious or unconscious selection of units to be included in the sample. A computer-generated set of random numbers can be used, but this usually requires that a grid or co-ordinate system be established in the study site so that each potential sample member can be identified. This, in itself, can be logistically difficult. So called 'haphazard sampling', in which sample units are selected without the help of random number tables, is *not* random. Procedures such as throwing a quadrat over one's shoulder or sticking a pin in a site map while blindfolded are subject to unconscious bias, which will result in biased estimates. Random sampling is discussed in more detail by Thompson (1992).

Simple random sampling may not be the most cost-efficient sampling pattern because of variation within the site or time period of interest.

3.4.1.2. Stratified Random Sampling

Stratified random sampling can often be substantially more efficient than random sampling; it is typically used in audit monitoring or to compare water quality against a guideline value. In stratified random sampling, the system to be sampled is divided into parts (strata) in each of which the variable of interest is as uniform as possible. Strata need not be of equal size. The numbers of sample units allocated to each stratum can be either in proportion to the size (area, volume) of each stratum or in proportion to the variance within each stratum.

Strata may be spatial or temporal. For example, for water sampling to measure nutrients, chlorophyll and algae, a lake can be divided spatially into the epilimnion and hypolimnion, or an estuary can be stratified on the basis of a salinity gradient. Temporally, if nutrients are more variable in one season than another, more sampling effort can be allocated to the most variable season, particularly if estimates of the annual concentration or load of the nutrients are the focus of the program. Sometimes strata result from an interaction of spatial and temporal processes. For example, suppose fish in a lake are being collected to study the accumulation of chemical contaminants, it is important to consider fish mobility and fish age (size). Older fish often accumulate more of a contaminant. Fish ages (sizes) then become the sampling strata, instead of geographical locations or particular periods of time.

3.4.1.3. Systematic Sampling

In systematic sampling, sample units are collected at regular intervals in space or time. When properly planned and executed, systematic sampling can be as unbiased as random sampling, and can be significantly cheaper (see Cochran 1977 for a full discussion). However, care needs to be exercised to ensure that bias is not inadvertently incorporated into the sampling scheme. For instance, regular sampling schedules may coincide with periodicities in the disturbance being monitored (e.g. discharges from a factory may be consistently lower in the morning and greatest just before shut-down in the late afternoon). Similar situations can arise spatially.

There needs to be a good descriptive base of background information so that systematic sampling can be both cost-effective and unbiased, and it is essential to document the assumptions and choices made when executing such a sampling regime.

3.4.2. Selection of Sampling Sites

It is important to select sites that provide appropriate spatial information. The problem being addressed will largely determine the general locations of sampling sites. The statistical analyses that will be used to interpret results (see Chapter 6) will also guide this decision (Ward et al. 1990). When ecological impacts are being assessed, sites will normally be located relative to the likely disturbance. Only rarely will sites be located randomly, as discussed above (section 3.4.1), but when this is done the number of sites and the extent of homogeneous areas in which they may be located can be determined from the pilot study. Multivariate classification procedures can be used for grouping similar sites, to define homogenous areas (Clarke and Warwick 1994).

When selecting appropriate sampling locations, the monitoring team should consider the possibility of seasonal variations and of local variations in other parameters to be measured (e.g. sources of contaminants), by referring not only to the pilot study but also to past records. These could be records of activities in the catchment, aerial photographs, plans and maps of land use, and oral or other records of the sites and the catchments under investigation. The team may find, for instance, that the water quality should be monitored not only in major surface waters and groundwaters that might receive inputs of substances from diffuse sources, but also in small creeks hydrologically connected with those waters.

It is also important that sites be selected to minimise any artefacts from human interventions that are not part of the monitoring program. For example, flow may be modified around jetties or bridges and that may affect some benthic measurement parameters, resulting in spurious data if the effect of the jetty or bridge is not the focus of the monitoring program. Similarly, weirs and similar structures in

rivers often alter both the flow and the chemical conditions, and sampling sites need to be located far enough up- or downstream of such structures if the water quality of the free-flowing water is the major focus of the monitoring program. In the field, the actual sampling sites will usually be selected by personal judgment using pragmatic considerations such as accessibility and safety.

When control or reference sites are included in the design, care needs to be taken to ensure that they are closely matched with the site being assessed. Sometimes information on covariates can be collected at all the sites and used to adjust the values of measurement parameters for inherent differences between the sites; the assumptions of the statistical analyses of such data need to be met (see Chapter 6). For example, in studies of metals in sediments, sediments from the reference site should have grain size and organic content similar to those from the test sites.

If sampling sites are too close together, or samples are collected at too close a time interval, autocorrelation or serial correlation between sites can invalidate the assumptions of independence made in some classical statistical designs. What constitutes *too close* (spatially or temporally) depends both on the nature of the measurement parameter and the dispersion of the contaminant. The monitoring team should consider whether to select alternative sites, or if sufficient data can be collected to implement designs that can model these spatial patterns properly.

Some water quality programs (usually based on chemical and physical measurement parameters) rely on networks of sites. There are high costs associated with monitoring, so monitoring programs should be optimised with regard to networks and sampling. A number of 'spatially optimum water sampling plans' exist, and their merits have been reviewed (Dixon and Chiswell 1996).

Finally, some pragmatic considerations for selecting sites.

- Safe access must be ensured under all conditions. If the sites are inaccessible during the wet season, for example, then the monitoring program cannot address questions about water quality during wet seasons.
- Sites also need to be accurately identifiable so that they can be sampled repeatedly. Global positioning systems greatly ease this task in areas of low relief and in off-shore marine environments.
- Groundwater quality monitoring programs often require a carefully staged approach, taking account of local geology, the vulnerability of the aquifers to contaminants and land use patterns, and any changes in pattern.

3.4.2.1. Spatial Variation Within a Sampling Site

There may be spatial variation *within* a site that needs to be quantified in the monitoring program, because otherwise the estimates of the chosen measurement parameter may be imprecise or even inaccurate. For example, in thermally stratified waters the depth of sampling is important because the concentrations of many measurement parameters (e.g. hydrogen ions, dissolved oxygen, nitrate, hydrogen sulfide, plankton) can vary greatly between the top and bottom layers. In rivers, samples taken from the edge rather than from mid-stream are likely to contain quite different amounts of suspended material and therefore different amounts of various compounds bound to the particulate matter. In benthic sampling for biological parameters (e.g. invertebrates, algae) or for sediments, the habitats or sediment types may vary at a site. In formal terms, these different habitats or water types within a site are called *strata*.

It is important that the monitoring team recognises that stratification in the measurement parameter will affect the data being obtained. There are three options for dealing with such strata:

- restrict the scope of the inference to a particular stratum. For example, if sandy sediments dominate the substrate at all the study sites, it may be sensible to confine sampling to sandy substrates. The stakeholders must be made aware that the inferences drawn are applicable only to 'sandy substrates within the sites' and cannot be generalised to strata that were not sampled within the sites.

- divide the sampling effort among the strata. Here the goal is to estimate the value of the measurement parameter for each site as a whole rather than for a stratum within the site. Stratified random sampling (see section 3.4.1.2) is an example of this procedure that is fully explained in basic texts (e.g. Cochran 1977; Elliott 1977). The number of sample units allocated to each stratum can be determined by the relative sizes (e.g. area or volume) of each stratum, or by the within-stratum variation of the measurement parameter(s).
- make separate estimates for each stratum (if this is consistent with the study objectives). Here the monitoring team may want to identify the nutrients in each stratum. For example, at each site in a reservoir separate nutrient samples can be taken from the epilimnion and the hypolimnion (i.e. two strata). They are then kept separate throughout the analyses.

When measurement parameters are being sampled in the water column, it is sometimes assumed that the water is well mixed and that a mid-water or mid-stream sample will be sufficiently representative. This may not be the case. Even in fast-flowing mountain streams, water can be observed flowing upstream in eddies. In larger rivers, tributary water may not mix fully with the mainstream for many hundreds of metres or even kilometres. In estuarine waters, salinity may be significantly stratified, and all water bodies can have gradients of redox potential and temperature. Even if the monitoring goal is just to measure the average concentration of a chemical in the water at a site, the sampling process must be planned so that the within-site variation is included in the estimate.

The same situation apples to the monitoring of aquifers, where groundwater quality is almost always stratified vertically, and where there can also be significant lateral variation in quality (e.g. in areas where there are multiple point sources or variable diffuse sources of contamination). There is much less dispersion of contaminants in groundwater than in surface waters, and so natural spatial variability is potentially much greater than in surface waters.

3.4.3. Sampling Frequency

The objectives of the monitoring program dictate the basis for determining sampling frequency. Thus a program to detect conformity with a guideline might be based on daily, weekly or quarterly sampling. The monitoring team must decide whether seasons are important, e.g. wet vs. dry season in the tropics, winter vs. summer in temperate regions where snow melts might be important.

Patterns in time include natural systematic changes, ranging from tidal cycles to larger scale events such as the El Niño–Southern Oscillation. These may be periodic and predictable (e.g. tides, seasonal filling of a wetland) or non-periodic (e.g. storms or floods in streams). Within these events, there are unpredictable variations (e.g. changes in the recruitment of a species after minor natural disturbances; changes in the concentrations of chemical measurement parameters after rainstorms).

Non-periodic events such as storms and associated runoff can have a dramatic impact on water quality that might be missed by sampling on a fixed time interval. If the monitoring team identifies this possibility during its preliminary assessments, it must design a sampling program that includes these events. Rapid changes in flow can profoundly affect the concentrations of measurement parameters and therefore the representativeness of sampling. Even under relatively stable flow conditions, the team must measure hydrological parameters when water sampling if the program is measuring the loads of a measurement parameter rather than its concentration.

The monitoring team must understand the system and the problem and issue being investigated, as illustrated by its conceptual process model, before it can select appropriate time intervals for sampling. The program objective and the expected statistical analyses can both influence the time interval chosen between samples. For example, the team may want to be 95% certain of detecting a 5% increase in nutrient levels. Once this objective is set, it will be a relatively straightforward statistical matter to determine the frequency of sampling.

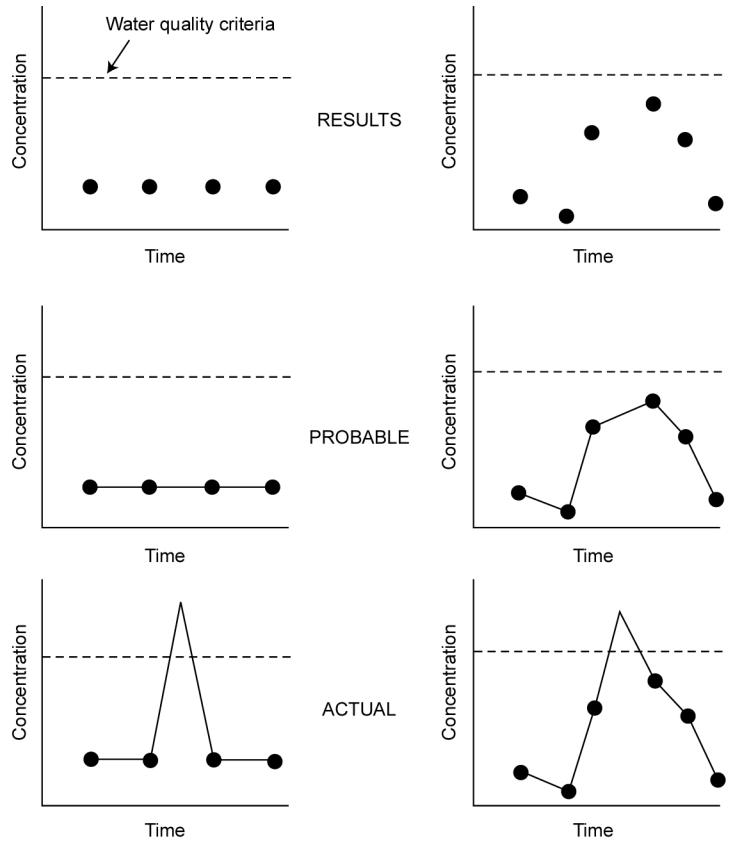

Figure 3.5. Frequency of sampling: interpretations of sampling data (modified from Maher et al. 1994)

The values of a particular measurement parameter may not vary at all time scales. If a measurement parameter has a predictable temporal pattern (e.g. recruitment with onset of the wet season, or deoxygenation during thermal stratification), the monitoring program must sample this measurement parameter at a frequency that suits this periodicity. Then trends can be estimated, especially in chronic or 'press' impacts. If a disturbance is only likely to take place at a certain time of year, for example mine wastewater discharge during the wet season in Kakadu (Humphrey et al. 1995), then sampling can be targeted to such predictable 'pulse' disturbances. At the other extreme, to measure the effects of highly variable and unpredictable disturbances (e.g. stormwater discharges), the monitoring program must sample at several time scales. Some measurement parameters give snapshots of immediate condition; some are integrating measures that reflect conditions over the past (x) months. These time-scale decisions need to be based on:

- the characteristics of the parameter being measured;
- the purpose of the data collection;
- the statistical or other tools that will be used to interpret the data; for instance, for time series analysis the monitoring team may have to decide on and set a definite sampling interval;
- the characteristics of the response of interest; for example, weekly measurements might be appropriate for measuring the development of an algal bloom but not for investigating fish. The generation time of the organism might be the critical determinant of time scales.
- recognition that a process cannot be measured if it takes longer to happen than the period over which measurements are made.

The case study of the Great Barrier Reef (see Appendix section A4.4.2) illustrates decisions about sampling frequency.

The frequency of sampling is especially important when the objective of the monitoring program is to ensure that particular guidelines or standards are not exceeded. Figure 3.5 shows some possible mis-interpretations that arise from sampling at inadequate frequencies. The data values obtained at the selected sampling frequency are all below the water quality guideline. Actually, there were excursions above the guideline between samplings that were not evident at the selected sampling frequency. Mathematical formulae are available to calculate the sampling frequency required for a particular study (Sharp 1971; Montgomery and Hart 1974), but these are not in widespread use.

3.4.3.1. Specific Concerns for Biological Measurement Parameters

Biological sampling should also take into account the time-dependence of an organism's behaviour. For example, Magmann (1991) re-examined a published study on the Northern Red Belly (*Phoxinus eos* and *Phoxinus neogaeus*) in which the densities of both fish species were reported to be at their highest at or near the shore. The original conclusion, based on a 16–18 hour trapping period beginning at 1600–1900 hours, was that both species exploited the same microhabitat. However, the initial study failed to recognise that *Phoxinus eos* has a diurnal pattern of inshore–offshore migratory behaviour. These fish swim in shoals in the inshore zone (<0.5 m) depth during the day and migrate to the offshore zone (>2 m depth) at sunset when shoals break up into single fish, then go back to inshore zone at sunrise. A shorter sampling interval (3–4 hours) was required to observe this movement. The density of fish offshore seemed to be lower because the fish shoals had broken up. Unfortunately there are few behavioural data for Australian species, so guidance is difficult to obtain.

Biological parameters may have the problem of serial correlation because of the long life-span of the organisms involved. The size of fish populations, for example, may depend on year-to-year variations in recruitment that may not be consistent across all the sites included in a study. Auxiliary data on the age-structure of the populations would be necessary to unravel these effects. Serial correlation will cause problems if the statistical methods being used assume that the measurements at different times are independent. Chapter 6 gives further discussion of these concerns.

3.4.3.2. Specific Concerns for Chemical and Physical Measurement Parameters

Special care should be taken when measuring some chemical and physical parameters, e.g. dissolved oxygen and pH levels in still or slow-moving surface waters. The values can change dramatically in these waters during the day, through photosynthesis and respiration. For example, dissolved oxygen must be measured before sunrise to obtain the diurnal minimum; diurnal pH fluctuations occur when carbon dioxide concentrations vary, and the pH decreases at night when dissolved carbon dioxide and carbonate accumulate in the absence of photosynthesis. The practice of sampling at a certain time of day, without regard to the changes that occur between daylight and darkness, can therefore result in misleading data.

If concentration measurements are being used to calculate loads, it will be important to decide how to relate flow and concentrations, and on what time basis. There are four common types of system with differing dominant processes that must be recognised when considering the frequency of sampling:

- base flow and point source discharges are the major determinants of water quality; or
- runoff (volume of storm event) and non-point sources are the major determinants of water quality; or
- remobilisation is the major determinant of water quality; or
- diurnal cycles (tidal cycles or biological activity) are important.

Australian freshwaters are characterised by highly variable flow, so flow is a major issue for them. It affects both water quality and biology. Changes in flow can alter water quality parameters rapidly and sometimes unpredictably for several reasons including these:

- hydrological changes alter the relative proportions of discharge originating from runoff, baseflow and groundwater. Runoff water may be of better quality than groundwater and baseflow (Hart et al. 1987) but is not invariably so. This can often be observed as a period of decreasing electrical

conductivity during the rising limb of the hydrograph, followed by rising electrical conductivity during the falling limb of the hydrograph. Alternatively, runoff may contain increased concentrations of nutrients (from fertilised fields, urban areas or sewage treatment plants) or heavy metals and organic compounds (from contaminated sites).

- rainfall events within a catchment may give different patterns of water quality depending on their locations;
- deliberate releases of contaminants may coincide with extreme hydrological conditions to take advantage of the large dilution factors available then;
- extreme rainfall conditions may breach bunds and other containment devices used for the retention of contaminants;
- in the case of temporary water courses, very large changes in water quality may occur during extreme recessional flow and during the 'first flush' of new flow. In the latter case, chemical species that have accumulated in catchment soils and near-surface groundwater (sometimes acidic from organic degradation or sulfide oxidation) may dramatically alter the concentrations of some measurement parameters.

It is well known that in many water bodies the total heavy metal and phosphorus concentrations are correlated with flow discharge, particularly during the early part of a stormwater runoff or flood event. At this time there are more suspended solids and correspondingly higher concentrations of associated heavy metals and phosphorus. Therefore if a river is only sampled at base flow for chemical and physical measurement parameters, the resulting data will not truly represent the natural range of heavy metal and phosphorus concentrations in the water body. Similarly, in the initial stages of an algal bloom, the numbers of algal cells may double every 2–3 days. If the monitoring program is measuring some aspect of nutrient fluxes, the sampling needs to reflect flow events that transport materials into and through the aquatic system. Catchment exports to rivers and estuaries are also assessed by intensive sampling during events.

Traditionally, the designers of monitoring programs have not considered sampling under different flow regimes. However, in the majority of Australian rivers, most (70–90%) of the annual flow and constituents are discharged under high flow or event conditions even though these may prevail for only 1–10% of the time. Under these conditions, the dominant water quality processes are the transport and deposition of discharged material during the flow event, followed by in-stream remobilisation of deposited material in the 10–30 days following the event.

Where the issues underlying the monitoring program relate to flow, the monitoring team must consider the following:

- the importance of flow-based monitoring and of capturing first flush and peak events;
- the need to measure and record flow data in conjunction with analyte concentration data obtained at the same time;
- the need to sample and obtain information at all flow regimes, including low flow, so that water quality can be described for all conditions of the water body.

To solve the difficulty of sampling at all flow regimes, a range of robust and reliable automatic sampling devices can now be obtained. These are capable of being automatically triggered by rising flow, and can automatically take samples at predetermined times or stream heights, to provide a comprehensive picture of changing constituent levels throughout an event. These data can then be used in association with flow data to calculate the contaminant export (or load) for a storm or flood event. Monitoring effort can be reduced over the long periods between events because there is little water quality variation during these conditions. An exception to this condition is the case of base flow with point source discharges.

3.4.4. Sample Numbers and Precision

An important aspect of the sample design is the number of samples to be collected to address the monitoring program's objective. This will largely depend on the nature of the investigation. In descriptive studies or in studies to determine cause and effect, the number of samples will determine the power of the data to assess differences. In studies that measure change, there must be enough samples to detect the minimum effect, or smallest differences or changes, that will cause management action — the 'effect size' (Keough and Mapstone 1995 p.102).

The monitoring team must decide on the required precision and accuracy. How many samples are needed for measuring each parameter at each site precisely on each sampling occasion; how many samples can the monitoring program afford to take? The team will base its decision on the results of the pilot study or on other reliable estimates of the variance and the costs of sampling (Keough and Mapstone 1995). The appropriate level of replication is not a simple decision (Segar et al. 1987; Mapstone 1995) because it must:

- be scientifically attainable;
- be attainable through a sampling and analysis program which can be accomplished in a cost-effective manner;
- minimise the risks of falsely detecting a disturbance or environmental impact when one has not occurred (giving a false alarm), or alternatively missing an environmental impact if it has occurred (giving a false sense of security);
- detect differences or changes that are environmentally important — that is, the change must have ecological meaning to the system of concern.

The smallest differences or changes that must be detected determine the number of spatial and temporal replicates needed (Norris and Georges 1986) and the precision needed. If a copper guideline concentration is 5 µg/L, is it important, environmentally, to be able to detect 5.01, 5.1 or 5.5 µg/L? This not the same as statistical significance (see Appendix 5 for a fuller explanation; see Mapstone 1995 for an explanation in a hypothesis-testing framework).

Once the difficult scientific and socioeconomic questions have been answered about the size of the differences or magnitude of the trends that must be detected, the number of replicates required can be calculated (see Appendix section A5.1.10). This is effectively an application of statistical power analysis, and more detailed explanations of the basics are found in many introductory texts (e.g. Cohen 1988; Sokal and Rohlf 1995). Various formulae are available for calculating the required numbers of replicates (e.g. Norris et al. 1992; Keough and Mapstone 1995), although investigators should be aware of the distributional assumptions behind such formulae (see section 6.3.5). On the other hand, decisions about optimum sample sizes for complex designs may not be easy (Green 1989, 1994; Norris et al. 1992; Keough and Mapstone 1995, 1997), and professional statistical assistance may be required.

Two related issues need to be borne in mind during this process of determining the number of replicates required for the program. First, investigators and their statistical consultants need to be clear about what constitutes a true replicate for the question being addressed by the monitoring program; 'pseudoreplication', where there is autocorrelation between apparent replicates, has been rife in many environmental programs (Hurlbert 1984; Eberhardt and Thomas 1991). Second, many programs will require sub-sampling within sites and within time periods to improve the precision of estimates.

There are trade-offs with costs, but unless the sampling is done in a way that enables the required data to be collected it cannot hope to answer the study objectives. If the monitoring team finds that resources are limiting, they may need to reconsider the sampling objectives. The monitoring team will need to decide on the sampling effort that is required to test critical hypotheses, if these are being used. If precision will be below that at which the critical hypotheses can be tested, the proposed sampling design is a waste of time and money. If the information generated by the monitoring program is to be used to make decisions, priorities will often be based on the risks associated with making wrong decisions. Risk is often viewed not in environmental terms but as political or social costs.

3.5. Selection of Measurement Parameters

The selection of measurement parameters is a vital element of the monitoring program design. A wide range of physical, chemical, ecotoxicological and ecological measurement parameters can be used to provide information on water quality. There is no simple or single physical or chemical measurement parameter that defines the quality of water. The choice of measurement parameters depends on the values ('environmental values') assigned to the water body (ecosystems, drinking water, recreation, industry, agriculture, aquaculture), and therefore on the objectives of the study. Furthermore, the guideline values acceptable for a measurement parameter for a particular use can differ geographically and temporally (ANZECC & ARMCANZ 2000).

The Water Quality Guidelines promotes the idea of integrated assessment. This approach merges biological (effects) and chemical (causes) approaches, and combines holistic field evaluations of impacts at the community and population level with laboratory toxicity tests. While the order of importance of biological versus chemical and physical monitoring can be debated, all provide important information as part of the integrated assessment of ecosystem health. Three-pronged studies (the triad approach), using chemistry, ecotoxicology and ecology have been promoted for sediments, and apply equally to waters (Chapman 1990).

Chemical measurements provide concentrations of specific contaminants that might be the cause of specific effects or modifiers of them. Biological assessment can be broadly subdivided into laboratory studies of chronic and acute impacts on individual species (ecotoxicology), and field measurements of structure, populations of species and their diversity, and function (aquatic ecology). Ecotoxicological and ecological measurements are non-specific, responding to the sum of the contaminants in the system. They integrate the effects of these contaminants over time and provide a more direct measure of the health of an aquatic ecosystem. Some taxa appear to be extremely susceptible to certain chemical contaminants and so provide a sensitive tool and early warning system for detecting slight contamination.

Table 3.3. Checklist for selection of measurement parameters[a]

Relevance	Does the measurement parameter reflect directly on the issue of concern?
Validity	Does the measurement parameter respond to changes in the environment and have some explanatory power?
Diagnostic value	The measurement parameter must be able to detect changes and trends in conditions for the specified period. Can the amount of change be assessed quantitatively or qualitatively?
Responsiveness	Does the measurement parameter detect changes early enough to permit a management response, and will it reflect changes due to the manipulation by management?
Reliability	The measurement parameter should be measurable in a reliable, reproducible and cost-effective way.
Appropriateness	Is the measurement parameter appropriate for the time and spatial scales of the study?

[a] Adapted from Maher and Cullen (1997)

The monitoring team's conceptual process model has already defined the water body and the problem and underlying issues that it is monitoring. Now the team must decide whether to measure the driving or causal factors (e.g. contaminant concentrations), or the consequential or resultant factors (e.g. effects such as toxicity, or algal biomass), or both — and why. The fundamental questions are these: how will the two sets of data be used; will the chosen measurement parameters have relevance to the problem or issue; will they change (or react to change) within the time-frame of the monitoring program; are they readily measurable? Table 3.3 is a checklist of these and other characteristics for selecting appropriate measurement parameters. These considerations have led to an upsurge in the

monitoring of biological outcomes rather than chemical inputs to a system, sometimes in community-based programs such as the algal alert programs and the biological components of Waterwatch.

A trade-off may be required between the exactitude of some measure and its cost or difficulty of measurement. The monitoring team might wish to know the concentrations of dissolved or bioavailable contaminants (e.g. phosphorus, metals) but may settle for total concentrations because they are easier to measure and more reliable (Lambert et al. 1992).

In many studies, parameters are measured that are not related to the conceptual process model of the system on which the study is based and for which no predictive power has been assumed. The reasons for including these measurements need to be justified.

The Monitoring Guidelines does not recommend any particular biological or physical or chemical measurements. For detailed description of physical, chemical and biological monitoring approaches, including ecotoxicology, see the Water Quality Guidelines (ANZECC & ARMCANZ 2000).

3.5.1. Physical and Chemical Measurement Parameters

The physical and chemical measurement parameters used for assessing water quality are discussed in detail in the Water Quality Guidelines (ANZECC & ARMCANZ 2000). Physical measurement parameters include flow, temperature, conductivity, suspended solids, turbidity and colour. These are important parameters themselves, and modify the impacts of chemical stressors. Chemical measurement parameters include pH, alkalinity, hardness, salinity, biochemical oxygen demand, dissolved oxygen and total organic carbon. In addition, other major controls on water chemistry include specific major anions and cations, and nutrient species (phosphate, nitrate, nitrite, ammonia, silica). These controls together with the physical measurement parameters determine the stability, chemical forms and bioavailability of a range of minor and trace contaminants such as metals, metalloids and specific organic compounds. Chapter 5 discusses laboratory approaches to measuring these parameters.

Table 3.4. General measurement parameters used for assessing aquatic system health

Measurement parameter	Input	Potential effects
Electrical conductivity	Salt	Loss of sensitive biota
Total phosphorus	Phosphorus	Eutrophication (nuisance algae)
Ratio of total phosphorus to total nitrogen	Phosphorus and nitrogen	Cyanobacterial blooms
Biochemical oxygen demand	Carbon in organic material	Asphyxiation of respiring organisms, e.g. fish kills
Turbidity	Sediment	Changes in ecosystem habitat
		Loss of sensitive species
		Altered light climate that affects productivity and predator–prey relationships
Suspended solids	Sediment	Changes in ecosystem habitat, loss of sensitive species
Chlorophyll	Nutrients	Eutrophication
pH	Acid drainage	Loss of sensitive biota
Metals, organic compounds	Toxicants	Loss of sensitive species

The protection of aquatic ecosystems is a specific issue discussed in the Water Quality Guidelines that illustrates the selection of physical and chemical measurement parameters. Table 3.4 lists physical and chemical measurement parameters that are frequently used for assessing the general health of aquatic environments. (The term 'health' refers to the condition and functioning of an ecosystem in comparison to conditions and function that are thought to be natural.) Other physical measurement parameters, such as rainfall, catchment morphology, geology, water colour, inflow rates and temperature, may also be crucial causal factors underlying the general measurement parameters.

There have been attempts to integrate various water quality indicators into single indices of water quality, to simplify the presentation and communication of results (Ellis 1989 section 10.2; Ward et al. 1990 pp.73–74; and references therein). Suitable integration of physical and chemical indices can result from the deliberations of water quality experts supported by use of appropriate statistical methods. Another approach is to use multivariate statistical methods (Ellis 1989). However, such integration of indicators reduces the information presented and is not a substitute for detailed presentation of data. There is also the risk that deterioration in one indicator can be masked by improvement in another. Ellis (1989) recommends building up practical experience of the performance of an index over a trial period. Use of indices seems to have been limited in Australia.

3.5.2. Ecotoxicological Assessment

Ecotoxicological studies assess the chronic and acute toxic effects of contaminants on biota in waters and sediments. The studies include the application of laboratory bioassays, and the measurement of biomarkers, focusing on effects at the species level. The ways in which organisms deal with contaminants in terms of bioaccumulation, bioconcentration and regulation are important in determining the ultimate toxic impacts. Ecotoxicological assessment techniques are summarised in Table 3.5.

3.5.2.1. Toxicity Testing Using Sensitive Bioassays

For a limited suite of individual chemicals, it is well recognised that the application of water quality and sediment quality criteria does not necessarily provide adequate protection for aquatic life. Discharges from point sources, such as sewage effluents, and diffuse sources, such as urban runoff, are complex mixtures that contain many unknown compounds which may act together to increase or ameliorate the toxic effect.

Rather than attempting to identify all the chemicals in a sample, toxicity tests (bioassays) using living organisms, are useful as measurement parameters of water quality — in particular the potential toxicity of contaminants in aquatic systems. Bioassays with bacteria, algae, invertebrates and fish are widely used to assess the environmental impact of chemicals in marine and freshwaters, and sediments. Laboratory toxicity data have been used in ecological risk assessments, to derive water and sediment quality guidelines, to investigate the bioavailability of contaminants and to establish cause–effect relationships for particular chemicals.

Acute (short-term) toxicity tests typically measure organism survival over 96 hours or a sub-lethal effect such as light inhibition in light-producing bacteria. Chronic tests determine toxicity over a significant portion of an organism's life span (e.g. weeks, months or years), or use sensitive early life stages such as larvae. Such tests measure, for example, the inhibition of growth rate in microalgae, the inhibition of fertilisation in macroalgae, and developmental abnormalities in scallop larvae. Because different organisms have different sensitivities to the same chemical, batteries of toxicity tests on sensitive species from different trophic levels are currently used. Appropriate test species and ecologically relevant endpoints are selected to suit the aims of the particular study. As toxicity testing is a particularly specialised field of work, advice from experts in this area should generally be sought before opting for particular toxicity tests.

Table 3.5. Summary of ecotoxicological approaches and measurement parameters

Approach	Measurement parameters	Advantages	Disadvantages	Overall value
Single species aquatic bioassays	Algae, bacteria, invertebrates and fish	Ultimate measure of chronic and acute biological impacts which, in combination with Toxicity Identification and Evaluation protocols, can identify and target the source of toxicity	Difficult to extrapolate from laboratory bioassays and mesocosms to the ecosystem as a whole Tend to concentrate on acute not chronic tests; short life-span chronic test species may not be representative	Ultimate measure of water quality with respect to toxicants
Whole sediment or sediment pore water bioassays	Algae, invertebrates	Pore water tests use water organisms; whole sediment tests are better	Difficulty in maintaining field chemistry (redox conditions) in laboratory studies; pore water tests not always on ecologically relevant species	Ultimate measure of sediment toxicity
Biomarker studies	Algae, bacteria, invertebrates and fish (e.g enzyme changes, scope for growth, deformities)	Indicators of chronic stress or exposure	Difficult to relate some changes to specific chemical exposure or to extrapolate from biomarker changes to whole organism or ecosystem effects	Currently more indicative of exposure than effects
Bioaccumulation and biomagnification of toxicants	Principally macro-invertebrates and fish	Some techniques can target particular toxicants, others are non-specific; diagnostic potential good; indication of accumulation of bioavailable chemical contaminants	Require sophisticated equipment for analysis of toxicants, and high level of expertise Need to establish factors affecting bioaccumulation, e.g. size, sex, age, exposure history Difficult to interpret ecological significance	Greatest potential is for detecting known toxicants Can assess exposure to chemical contaminants, but need a very good understanding of intrinsic and extrinsic factors affecting accumulation

Test species can be cultured in the laboratory (phytoplankton, cladocerans) or collected from the field immediately prior to testing, e.g. sea urchins. Tests using readily cultured species have the advantage that they are highly reproducible and are not subject to seasonal availability. Microbial tests, in particular, are rapid, relatively inexpensive, sensitive to some contaminants, and do not have animal ethics constraints on their use. It is important, however, to use a suite of toxicity tests from different trophic levels (OECD 1984), if attempting to relate the results to potential effects in the environment.

Standard toxicity test protocols have been published by OECD (1984), ISO (1989), USEPA (1993, 1994a,b) and Environment Canada (1990a, 1992). The protocols have been adapted to local species of marine and freshwater algae (Stauber et al. 1994; Gunthorpe et al. 1997), invertebrates (Julli et al. 1990; Krassoi et al. 1996; Simon and Laginestra 1997) and fish (Munday et al. 1991; Hyne and Wilson 1997). Choices of test species (marine, freshwater, tropical, temperate) and test endpoints (e.g. survival, growth inhibition, reproduction, developmental abnormalities) depend on site-specific requirements.

For freshwater testing, growth and enzyme inhibition bioassays have been developed with Australian isolates of freshwater green algae such as *Chlorella* species (Stauber 1995; Franklin et al. 2000). Protocols have also been devised with a number of freshwater cladoceran species including *Daphnia carinata*, *Ceriodaphnia dubia* and *Moinodaphnia macleayi* (Julli et al. 1990; Hyne et al. 1996) and the freshwater shrimp *Paratya australiensis* (Abdullah et al. 1994). Acute toxicity tests have also been developed with eastern rainbow fish (Kumar and Chapman 1998) and the tropical purple spotted gudgeon (Markich and Camilleri 1997).

Marine toxicity testing protocols with Australian algal species include growth inhibition and enzyme inhibition tests with diatoms (*Nitzschia closterium*) and green microalgae (Stauber et al. 1994), and fertilisation, germination and growth inhibition tests using macroalgae such as *Hormosira banksii* (Gunthorpe et al. 1995; Kevekordes and Clayton 1996), *Phyllospora comosa* and *Macrocystis angustifolia* (Burridge et al. 1995, 1996). Marine invertebrate tests include inhibition of scallop larval development (Krassoi et al. 1996), fertilisation inhibition in sea urchins (Simon and Laginestra 1997), and amphipod survival (Burridge et al. 1995). Marine fish tests using larval survival of Australian bass and estuarine perch (Hyne and Wilson 1997) have also been developed; however, their application is currently limited by animal ethics restrictions on fish testing, particularly in NSW.

For sediment toxicity testing, an acute 10-day sediment toxicity test and a 14-day chronic test with the estuarine amphipod *Corophium* sp. has been devised (Hyne and Everett 1998), as well as a test using microphytobenthos such as the diatom *Entomoneis* cf. *punctulata* (Adams 2000). An acute 10-day freshwater sediment test has also been developed using nymphs of the mayfly *Jappa kutera* (Leonard et al. in press). While toxicity tests are useful measurements of water quality, they cannot identify specific toxicants causing an observed effect. Toxicity identification and evaluation (TIE) combines chemical manipulation, analytical chemical techniques and concurrent toxicity testing to determine the components of the effluent, water or sediment that are causing the observed toxicity. Protocols for freshwater and marine toxicity identification evaluation have been developed by the USEPA (1991a, 1996a) and modified for use with Australian species in freshwaters (Pablo et al. 1997) and marine waters (Hogan 1998; Doyle 1998).

3.5.2.2. Measurement of Biomarkers

A biomarker is a 'variation in cellular or biochemical components or processes, structure or functions that is measurable in a biological system or sample' (National Research Council 1987). Biomarkers include change in enzyme activity, biochemical changes, physiological changes, histopathological changes and physical deformities. The most intensively studied group of organisms is fish, particularly marine fish (Holdway et al. 1995). A number of studies have examined the abilities of sub-cellular biomarkers to respond effectively to, and generally indicate, the effects of a number of chemical contaminants. The most promising use of biomarkers is as a screening tool for the detection of exposure to contaminants, e.g. with mixed function oxidases. However, biomarkers give little information about the effects of chemical contaminants and it is difficult to relate biomarker changes to effects at the individual organism, population, community or ecosystem level.

Biomarkers have been applied worldwide in both freshwater and marine ecosystems. Applications include investigation of deformities in chironomid mouthparts (Warwick 1989), and cellular and enzymatic changes (Gunther et al. 1997; Viarengo et al. 1997). At this stage in Australia, biomarkers have been developed only for estuarine and marine systems, e.g. biomarkers in flathead to detect contaminant exposure (Holdway et al. 1994, 1995), histopathological changes in flounder (Stauber et

al. 1996) and other organisms (Twining and Nowak 1996) and enzymatic changes in algae (Peterson and Stauber 1996).

3.5.2.3. Measurement of Bioaccumulation

Two major difficulties in identifying toxicant impacts on ecosystems are that contaminant events may be episodic and that toxicants may have effects at very low concentrations. Where a particular toxicant is known or suspected to occur, monitoring of toxicant levels in biota can be a useful technique, particularly when concentrations of toxicants in water are too low to be measured chemically.

There are two biological phenomena that can assist with this approach: bioaccumulation and biomagnification (Connell 1981). Bioaccumulation refers to the continued accumulation of particular contaminants by some taxa throughout their lifetime, e.g. accumulation of trace metals by molluscs. The existence of this phenomenon means that episodic contamination events will be integrated over time in an organism's tissues. Biomagnification refers to the increase in concentration of a contaminant up the food chain, such as occurs with some organochlorine pesticides. This phenomenon makes it easier to detect low concentrations of chemical contaminants in an ecosystem. A disadvantage of this approach is that the toxicant must be known. With the increasing complexity of industrial and other effluents, it is not always possible to identify the key toxicants.

3.5.2.4. Early Detection of Change

Sub-lethal tests can be part of programs aiming for early detection of change. If potential adverse effects of a disturbance to an ecosystem can be predicted or detected quickly, more substantial and damaging effects may be avoidable through management action. Specific and sensitive early detection programs may be set up to provide, firstly, predictive information from laboratory-based direct toxicity assessment, and secondly, early detection in the field. Each involves measurement of sub-lethal responses by organisms.

Measurements of bioaccumulation, or biomarkers, or bioassays using sensitive species, as discussed above, all are techniques that might be used. They provide different information. Bioaccumulation is an organism-specific integration of water (or sediment) contaminants, so particular compounds must be measured. Biomarkers and bioassays provide more generic responses to acute or chronic impacts; chemistry then identifies the specific stressor that requires management. For more information on bioassays refer to the Water Quality Guidelines section 8.3.6, on direct toxicity assessment (ANZECC & ARMCANZ 2000 Volume 2).

3.5.3. Ecological Assessment

Ecological assessment aims primarily to measure the structure and function of biological communities. It principally involves field-based measurements that examine effects on the relative abundance and diversity of species, community structure and composition, and how these are altered as a consequence of known or unknown stressors and their modifiers in both waters and sediments. A summary of techniques for ecological assessment is given in Table 3.6, which also shows the variety of taxonomic groups that have been tried out for ecological assessment of ecosystem health.

Macroinvertebrates have been selected as the key indicator group for bioassessment of the health of Australia's streams and rivers under the National River Health Program (Schofield and Davies 1996). Active monitoring programs by state and territory agencies using the AUSRIVAS–RIVPACS-type approach (see Box 1, page 3-24), and underpinned by extensive R&D, are well established across the country. A large number of studies have been published on the use of benthic invertebrates in the assessment of both freshwater and estuarine water quality in Australia (e.g. *Australian Journal of Ecology* volume 20, number 1, 1995; Deeley and Paling 1999). The responses of various invertebrate taxa to specific types of chemical contamination are reasonably well documented for Northern Hemisphere waters (e.g. Hellawell 1986).

Box 1. AUSRIVAS — what it is, how it works and its origins

The Australian River Assessment System (AUSRIVAS), and its basis, the British RIVPACS system (River Invertebrate Prediction and Classification Scheme), are rapid standard methods for rating the ecological health of freshwaters by biological monitoring and habitat assessment. Both AUSRIVAS and RIVPACS consist of models that predict the fauna (usually macroinvertebrates) expected to occur at a test site on the basis of its environmental attributes — its geographic, physical and chemical features. When the test site is sampled, the fauna observed are compared to the models' expectations for that sort of habitat, and the resulting observed/expected (O/E) score is an integrated indicator of river health.

AUSRIVAS was developed over the period 1993–1997, and has established protocols for invertebrate sampling, habitat assessment and model development. It is now being used in a national assessment of river health involving some 6000 sites to be sampled over three years. In the development of AUSRIVAS, aquatic invertebrates at more than 1500 minimally disturbed sites (reference sites) were sampled across Australia to establish a reference site database from which to build the predictive models. Sites were sampled in two seasons: autumn and spring for temperate regions, and the wet and dry seasons in the tropics. Five types of habitat were sampled and the two most common aquatic habitats from riffle, edge, main channel, macrophytes and pool rocks were selected for use in constructing a model in each state or territory. Macroinvertebrates were chosen because they have these valuable measurement parameter characteristics:

- the fauna is well known taxonomically;
- the fauna is diverse with known differences in response to different contaminants;
- there are enough individuals within a sample to provide abundance data to be used in analysis, and yet numbers are not unmanageable;
- generation times of taxa are such that they integrate ecological impacts over a satisfactory period.

The AUSRIVAS O/E score is responsive to a variety of environmental effects, including water quality, habitat condition, and changes in flow regime. Various O/E score categories (bands) are used to provide a 'biological thermometer' of the overall condition and severity of disturbance for various sites. This allows the general health of the waters at the survey sites to be characterised in a nation-wide context. The AUSRIVAS scores do not provide a clear indication of the cause of a disturbance or contamination in waterbodies, but they can be used be used to identify those waterbodies that are 'stressed' or those that need further investigation and management action. The method depends on the existence of reference sites that are close environmental equivalents to the test site. If a test site's environmental attributes do not match any in the reference set, its health cannot be assessed unless suitable new reference sites are added to the database.

For more information, see http://ausrivas.canberra.edu.au/ (Coysh et al. 2000), Schofield and Davies (1996), Kay et al. (1999), Marchant et al. (1999), Turak et al. (1999).

Table 3.6. Summary of ecological assessment approaches and measurement parameters

Approach	Measurement parameters	Advantages	Disadvantages	Overall value
Diversity indices	Various	Provide summary of complex data; easy to understand, allow comparisons between sites or times	Ecological significance of indices is unclear; can be affected by sampling and analytical factors	Attractive for their simplicity, but their ecological value is questionable
Biotic indices	Principally macro-invertebrates and algae	Simple, easy to interpret summaries of complex data; can provide contaminant-specific response	Detailed knowledge of contaminant tolerance required for diagnostic use	Usefulness limited by baseline; site-specific and contaminant tolerance information needed
Stream community metabolism	Benthic flora and fauna	Integrates impact across the entire benthic biota; relatively rapid; provides a simple output	Technique not proved; may be less useful in disturbed catchments; diagnostic capability unclear	Technique has potential, but its sensitivity and diagnostic capacity have not been demonstrated
Macro-invertebrate community structure (e.g. AUSRIVAS) for rapid biological assessment; quantitative methods for site-specific studies	Macro-invertebrates	Integrates over appropriate temporal and spatial scales; much background information available; good diagnostic capability	Relies on complex modelling approach; output not as readily understood as other techniques	Great potential for identification of impacts; reasonable potential for establishing causes of impacts
Macrophyte community structure	Macrophytes	Easily sampled, respond to a range of impacts	Gives poor under-standing of factors affecting community structure; insensitive to some chemical contaminants	Limited use
Fish community structure, biomarkers (biochemical, physiological, immunological or histo-pathological)	Fish	Readily sampled, taxonomically well known	Gives poor knowledge of population dynamics and water quality factors; temperate fauna are impoverished; biomarker techniques require sophisticated equipment and high level of expertise	Community structure uses more applicable in tropical than temperate waters
Algae: biomass and community structure	Algae	Sensitive, taxonomically well known, has diagnostic potential; community structure (AUSRIVAS-type) approach most promising	Identification requires high level of expertise; community structure approach not well tested	Community structure approach has good potential
Bacteria, protozoa and fungi: community structures	Bacteria, protozoa and fungi	Organisms occupy key ecological role so community change can provide valuable key to impacts	May recover too rapidly from impact for monitoring purposes; taxonomy and response to chemical contaminants poorly known	Limited use at present; would require extensive taxonomic and diagnostic work before they could be useful

Fish have considerable potential for use in the bioassessment of water quality in some locations (Harris 1995). Australia has a freshwater fish fauna that is highly diverse in the northern part of the continent but of low diversity in southern and inland regions. Fish populations and communities respond to changes in water quality but are also strongly influenced by changes in hydrology (which affect recruitment, habitat and food availability) and physical habitat structure (such as organic debris, bottom substrate and pool dimensions, and barriers to migration). Fish have been little used for assessing water quality or human-induced effects on water quality except in the bioassessment program at Ranger uranium mine (Humphrey et al. 1990) and in a series of studies in south-east Queensland (summarised in Arthington et al. 1998). Current attempts to develop standardised bioassessment approaches using fish are in their infancy in Australia. None of these methods is applicable at a broad scale and none has been extensively tested to date. Approaches based on comparative measures of community composition are compromised in most of southern and inland Australia where species diversity is low, fluctuations in species abundance and occurrence are extreme (driven by unpredictable flow events), and the relative dominance of exotic species is high. Together, these factors mean that fish diversity can be strongly affected by the chance appearance or disappearance of a single species, and so it is often a poor measure of ecological integrity, requiring further work. Fish populations and communities change and respond to environmental factors at much longer time scales than most other aquatic biota (e.g. algae, macroinvertebrates).

Periphytic diatoms have been used in streams and rivers in the United Kingdom (Whitton and Kelly 1995) and, more recently, in Western Australia, South Australia, Victoria and New South Wales as part of the National River Health Program (John 1998). The Western Australian and New South Wales–Victorian work has demonstrated that ecologically disturbed and undisturbed sites consistently have different assemblages of diatoms, suggesting that this approach could be used routinely to identify disturbance. This work has also found that certain groups of diatoms are well correlated with water contaminants; i.e. the technique may have diagnostic potential.

Bacteria, protozoa and fungi have not been widely used in ecosystem health studies, but bacteria and protozoa have been used extensively to test that waters are safe for human use.

The potential for the development of macrophyte bioassessment procedures for Australian streams is currently being evaluated under the National River Health Program, but at the moment there are few universally applicable protocols available.

Before choosing a particular taxonomic group as a measurement parameter of water quality or ecosystem health, the monitoring team should check that the taxonomic group fulfils this set of criteria:

- the measured response reflects the ecological condition or integrity of the site, catchment or region to be monitored;
- approaches to sampling and data analysis can be highly standardised;
- the response can be measured rapidly, cheaply and reliably;
- the response has some diagnostic value.

3.5.3.1. Measures of Macroinvertebrate Community Structure

Macroinvertebrate communities provide the most developed indication of ecological health. Invertebrate data are analysed by aggregating them into measures or indices. For example, the Macroinvertebrate Community Index (MCI) (Stark 1985, 1993) has been developed in New Zealand and is now widely used there by regional councils to detect and monitor water quality degradation. Similarly, Chessman (1995) has developed the SIGNAL index (Stream Invertebrate Grade Number — Average Level) for invertebrates identified to family level in south-eastern Australia. These measures are forms of biotic indices. They are based on the premise that contaminant tolerance varies between species or higher taxa; they produce contaminant tolerance scores.

Two problems arise in developing and applying contaminant tolerance scores in Australia and New Zealand. First, most tolerance information relates to organic contaminants, although knowledge of

tolerances to acidification and heavy metals is growing; also, in Australia, most information on contaminant tolerances comes from the wetter and better studied temperate south-east and south-west of the country. Second, many tolerance indices are developed for family or coarser-level identifications (many groups of invertebrates are hardly known, taxonomically, beyond this level), but it is acknowledged that for some groups the constituent taxa may vary widely in their tolerances.

Four generic biodiversity-type protocols and four early-detection-type protocols have been developed for streams and wetlands using macroinvertebrate species or communities (see the Water Quality Guidelines Table 3.2.2 and Section 8.1.3).

Another biotic index, the ETP index, is based on Ephemeroptera, Trichoptera and Plecoptera. These three families of macroinvertebrates are sensitive to most types of contaminant, and so the numbers of individuals in these orders should decrease with a decrease in water quality. The numbers of some Diptera and tubificid worms may increase in response to contaminants, especially organic contaminants. This response has been used in constructing indices by examining the ratios between tubificids and other organisms, or by just counting the numbers of taxa belonging to the sensitive groups (e.g. Plafkin et al. 1989). Virtually all the indices or other measurements using these assumptions have been developed in relation to organic contaminants in rivers. However, some species of Trichoptera and Ephemeroptera are highly tolerant of trace metal contaminants (Norris 1986), so caution is advised in the general application of indices based on the assumptions just discussed. Other difficulties include the large amount of initial work that may be needed to define contaminant tolerances and to define 'clean' freshwater communities, and the limited number of taxonomic keys to many species (Resh and Jackson 1993).

Waterwatch programs across Australia collect macroinvertebrate samples as part of regular spring and autumn surveys. Using a national protocol, Waterwatch collects data on the abundance and diversity of macroinvertebrates identified to the order level only, at regular sites in local waterways. At this stage no adequate diagnostic tool is available to convert these data into a biotic index, though various trials have been conducted. Waterwatch is working with scientists to develop a modified SIGNAL-type scoring system for this purpose, and the data are being collected and compiled throughout Australia. Aside from the scientific purpose of collecting data, the collection and processing of macroinvertebrate samples by the general community has enormous educational value. Further information can be found on the Waterwatch web site (www.waterwatch.org.au).

The SIGNAL score is widely used and has been shown repeatedly to have strong relationships with water quality variables (Growns et al. 1995, 1997; Chessman et al. 1997; Chessman and McEvoy 1998). Research is currently being conducted through the National River Health Program to develop and test a series of other indices for macroinvertebrate community structure to assess river health and the impact of water quality changes.

The most widely used protocol is the Australian River Assessment System, AUSRIVAS, based on the RIVPACS model developed in the UK (see Box 1, page 3-24). Through the National River Health Program, the first Australia-wide assessment of the health of Australia's diverse and unique aquatic systems has been undertaken using macroinvertebrates and the AUSRIVAS method.

3.5.3.2. Rapid Biological Assessment

The Water Quality Guidelines make prominent reference to rapid biological assessment (RBA) in the context of rapid and cost-effective techniques for obtaining first-pass, not necessarily quantitative, data over broad geographical areas. Rapid techniques are suitable for determining the extent of a problem such as river health (see the Water Quality Guidelines section 3.2.1.3 and section 8.1.1.1). Further discussion on RBA approaches using stream macroinvertebrate communities is provided in Resh and Jackson (1993), Lenat and Barbour (1994) and Resh et al. (1995).

The most commonly used RBA method in Australia is AUSRIVAS. Rapid bioassessment protocols are also being developed for riverine benthic algae (diatoms) and fish, as well as for

macroinvertebrate communities in wetlands and estuarine sediments (see the Water Quality Guidelines section 3.2.1.3/1).

The data obtained from RBA may be suitable for broad-scale auditing or screening purposes and for broad-scale management and for use in an early warning system. In most cases RBA should be followed by detailed studies using quantitative methods for site-specific assessments.

The Water Quality Guidelines advises against the use of RBA methods alone for detailed site specific assessments and comparisons in time and space, though these methods may usefully support quantitative methods in this role (see Water Quality Guidelines Table 3.2.1, Table 8.1.2 and section 7.2.1.1/1). Rapid biological assessment methods have been suggested for monitoring the outcome of environmental flow studies or river rehabilitation or river restoration projects and for setting baselines for environmental impact assessment. These issues are complex because the RBA approach compromises various aspects of sampling design and its implementation, and that can have cumulative effects on the results obtained. For example RBA methods use sample collection methods that are simple in comparison to those required for quantitative assessment. Likewise, RBA may sample only single test sites rather than the set of sites required for statistical purposes to account for between-site variation. Also only a sub-sample of the animals collected may be identified, and only to the family or genus level rather than to species level. Abundance data for the taxon identified are not used by the models. All these elements when combined tend to make it difficult to 'gear-up' the RBA methods for quantitative analysis. For example, working out how many AUSRIVAS samples would need to be collected upstream and downstream of a point source of contaminants to monitor their impact and the selection of appropriate control sites may be irrelevant. The limitations of the other aspects of the protocol such as the collection methods, sample processing and analyses methods may not provide the level of detail required for coping with the variation between samples. The cumulative effect of these limitations restricts RBA methods, making it difficult to apply them or modify them for uses beyond those for which they were designed. The suggested appropriate application of RBA and quantitative methods is summarised in the Water Quality Guidelines Table 8.1.2 and the differences in the protocols for these methods are given in the Water Quality Guidelines Appendix 3 Methods 3A(i) and 3A(ii).

3.5.3.3. Whole Ecosystem Ecological Assessments

The conservation, maintenance, rehabilitation and restoration of healthy aquatic ecosystems and biotic integrity have become important objectives of water management worldwide (Gore 1985; Karr 1991; Rapport 1991) and also in Australia (Norris and Thoms 1999). The focus on healthy ecosystems also applies to lakes, wetlands, estuaries and other water bodies including groundwater ecosystems.

The term 'health' is usually defined in terms of ecological integrity (Schofield and Davies 1996; Karr and Dudley 1981) as:

> the ability of the aquatic ecosystem to support and maintain key ecological processes and a community of organisms with a species composition, diversity, and functional organisation as comparable as possible to that of natural habitats within a region.

Several single integrated measures of the integrity or health of aquatic ecosystems have been developed. Principally these are the *diversity indices* and *biotic indices* and the application of various biological measurement parameters, either on their own or combined with other measures for monitoring water quality. It is generally agreed that an integrated approach is desirable, but in practice several agencies are monitoring single measures as a priority.

Another approach makes integrated measurements of the metabolism of an aquatic community or some other ecosystem process involving nutrients, production or carbon metabolism. Ecotoxicological methods have also been used. Gower (1987) and Legendre and Legendre (1998) give guidance on the use of integrated techniques in freshwater. The Land Ocean Interactions in the

Coastal Zone Program (LOICZ) has considered community metabolism and modelling approaches to marine biodiversity; see http://kellia.nioz.nl/loicz/.

3.5.3.4. Diversity Indices

Diversity indices usually require a count of the total number of individuals and a total count for each of the taxa. The taxa need to be separated but not necessarily identified. Separation is often at the species level but it is sometimes at the generic or family level. A higher diversity, i.e. the presence of more taxa within a given number of animals, is taken to signify a healthier ecosystem. A disadvantage is that measurements require taxonomic skills, are tedious and require large numbers of samples to achieve statistical significance.

Despite this, diversity indices have been widely advocated as measurement parameters of ecosystem health for a variety of reasons:

- they are seen as a useful way to condense complex data and thus aid interpretation;
- people with little biological expertise can easily understand them, and can gather the data to create some of them;
- they are a more generic measure than physical and chemical measures;
- they allow comparisons between sites or times where collections have been made using different sample sizes, methods or habitats.

The combination of evenness and taxonomic richness in a diversity index supposedly indicates the state of the community. It seems to be generally accepted that index values decrease with decreasing water quality. Low diversity is taken to indicate a stressed community that tends to be unstable.

In practice, there are several matters for concern with the application of diversity indices. First, statistical anomalies occur with some indices as a result of the assumptions on which they are based. Simpson's index is based on the assumption that in more diverse communities there is a lower probability that individuals chosen at random will belong to the same taxon (Simpson 1949). However, this assumption disregards the possibility that members of the same taxon will be clumped for reasons of microhabitat, breeding, or behaviour.

Second, in the diversity indices of Gleason and Margalef which are similar and are based on guesses at fitting curves to species abundance distributions (Gleason 1922; Margalef 1958), the assumption is made that the number of individuals is directly proportional to the area sampled. Of course, these indices are likely to be highly dependent on sample size. There is some doubt about the biological meaning of frequency distributions, and it is not clear how environmental stress (including contaminants) will affect the relationship.

Third, a problem with the most widely used diversity indices, those derived from information theory, is the rather tenuous biological significance of the measure. For example the Shannon index (Shannon 1948) reaches its maximum value when all species are evenly distributed. Biologically, this is assumed to be the most desirable situation, although it contradicts the evidence provided by the log-normal distribution for many different communities. A range of factors other than contaminants may affect this type of diversity index, including sampling method, sample size, depth of sampling, duration of sampling, time of year, and taxonomic level used. Diversity indices based on information theory should be interpreted and compared with caution.

In summary, the measurement of diversity and the effect of contaminants on a diversity index both must be resolved before these indices can be used and the results interpreted effectively. When such indices are applied, the predicted effects of contaminants on the ecological attributes supposedly measured by the index are rarely stated. This criticism might also be applied to other indices and to most other approaches.

The use of diversity indices in Australia has tended to focus on those groups for which the taxonomy is best known, and which are abundant and taxonomically rich enough to provide reliable measures of diversity. Of particular interest have been aquatic macroinvertebrates. However, the extent of the

work on predictive modelling using macroinvertebrates (such as in AUSRIVAS, in which both sides of the O/E ratio are measures of the composition of the fauna) has overshadowed the use of diversity indices in Australia.

3.5.3.5. Biotic Indices

Biotic indices usually have been developed empirically as a means for assessing the effects of contaminants, mostly in rivers. Many are specific to a site and contaminant type (usually organic). The calculation of biotic indices usually requires:

- a total count of individuals or total counts of taxa;
- counts (or biomass measurements) of specific groups such as all insects and tubificid worms, or the number of mayflies, stoneflies and caddisflies;
- detailed lists of the responses of different taxa to contaminants; or
- division of invertebrates into groups with different feeding strategies.

Biotic indices are more clearly related to the conditions that led to their development than are diversity indices.

Some examples of biotic indices and community indices are these: the Invertebrate Community Index (ICI) (DeShon 1995); the Rapid Bioassessment Protocols used by USEPA (Shackleford 1988; Plafkin et al. 1989; Barbour et al. 1992, 1995, 1996; Hayslip 1993; Smith and Voshell 1997); and various Macroinvertebrate Community Indices (MCI), including quantitative and semi-quantitative applications, used in New Zealand (see also section 3.5.2.4; Stark 1985, 1993, 1998; Collier et al. 1998). The Index of Biotic Integrity (IBI), originally developed for fish (Karr 1981; Karr et al. 1986; Miller et al. 1988), uses species richness, abundance, community structure, and the health of the individual animals as metrics. It has been tried in Australia (Harris 1995; Harris and Silveira 1999), but its potential application is still under investigation. A Benthic macroinvertebrate Index of Biotic Integrity (B-IBI) (Kerans and Karr 1994; Fore et al. 1996) has also been developed.

The metrics used in these indices evaluate aspects of community composition and of structure and processes within the macroinvertebrate assemblages. Although the indices have been developed first for a particular region, they are typically applicable over wide geographic areas with minor modification (Barbour et al. 1995).

The USEPA has developed the Rapid Bioassessment Protocol for Use in Streams and Rivers (Plafkin et al. 1989); it assesses community diversity as a measure of water quality. Contaminants are indicated by the absence of contaminant-sensitive benthic macroinvertebrate groups (Ephemeroptera, Plecoptera, and Trichoptera) and the dominance of contaminant-tolerant groups (oligochaetes or chironomids). Overall, a paucity of benthic macroinvertebrates may indicate impairment. However, nutrient levels are naturally low in pristine headwaters and may explain the low productivity and few benthic macroinvertebrate species that are found there.

The most effective metrics are those that respond across a range of human influence (Fore et al. 1996; Karr and Chu 1999). Resh and Jackson (1993) tested the capacity of 20 benthic metrics used in 30 different assessment protocols to discriminate between impaired and minimally impaired sites in California, USA. The best measures, from their study, were the richness measures, two community indices (Margalef's and Hilsenhoff's family biotic index), and a functional feeding group metric (percent scrapers). Resh and Jackson (1993) emphasised that both the measures (metrics) and the protocols need to be calibrated for different regions of the country, and, perhaps, for different impact types (stressors). In a study of 28 invertebrate metrics, Kerans and Karr (1994) demonstrated significant patterns for 18 metrics and used 13 in their final B-IBI. Richness measures were useful, as were selected trophic and dominance metrics. One of the unique features of the fish IBI (Harris and Silveira 1999) which is not found in benthic indices is that it can incorporate metrics on individual condition. However, Lenat (1993) has advocated measures evaluating chironomid larvae deformities.

3.5.3.6. Similarity Measures

Communities of organisms at two sites can be similar or dissimilar. The similarity (or dissimilarity) can be measured and rated, and related to water quality that has been assessed using other parameters, perhaps physical or chemical.

Numerous numeric similarity (or dissimilarity) indices have been proposed. The PATN package (Belbin 1993; Belbin and McDonald 1993), developed in Australia, includes a wide range of similarity measures and techniques for examining community structure and patterns in species distribution, and association and similarity between sites and species.

3.5.3.7. Functional Feeding Group Measures

Some biotic indices are based on the assumption that the ratios of organisms with different feeding strategies will change with contamination (e.g. collectors will be more abundant than shredders under contaminated conditions), or that trophic generalist organisms will be more tolerant to contaminants than trophic specialist organisms. There is some doubt that these general rules hold true, and even that it is possible to assign taxa to different feeding strategies (Resh and Jackson 1993).

Organisms can be assigned to groups of species that feed in the same way, generally known as functional feeding groups. For example, bacteria can be photosynthetic autotrophs, or bacterivores–detritivores, or algivores, or nonselective omnivores, or saprotrophs, or raptors. Feeding groups have also been identified for fish: they can be predators, grazers, strainers, suckers or parasites. The usefulness of functional feeding groups has not been well demonstrated for benthic macroinvertebrates, and the concept is not considered reliable in this case (Karr and Chu 1997).

Feeding groups represent a way of assessing the dynamics of food supplies in a water body and the balance of feeding strategies (food acquisition and morphology) in the fauna assemblages. An imbalance in functional feeding groups generally indicates stressed conditions. Specialised feeders, such as scrapers, piercers, and shredders, are relatively sensitive organisms and are thought to be well represented in healthy streams. Generalists, such as collectors and filterers, have a broader range of acceptable food materials than specialists (Cummins and Klug 1979), and thus are more tolerant to contaminants that might alter availability of certain food. However, filter feeders are also thought to be sensitive in low-gradient streams (Wallace et al. 1977).

3.5.3.8. Taxonomic Richness

Taxonomic richness generally decreases with decreasing water quality. The number of individuals and biomass may increase or decrease, depending on the type of contaminant and the organisms involved. Considerable work is needed to classify adequately the responses of different taxonomic groups to contaminants, an increasingly difficult task because new chemical contaminants continue to appear, and because many effluents (including sewage) are becoming more complex.

3.5.3.9. Stream Community Metabolism

The stream community metabolism approach is based on the concept that movement of organic carbon through an ecosystem can be used as a measurement parameter of stream community metabolism. This in turn provides an indication of ecosystem health.

Two biological processes affect the movement of carbon: production (via photosynthesis) and respiration. It is argued that community metabolism is sensitive to small changes in water quality (particularly organic contaminants and sedimentation) and to riparian conditions that affect light input. As a result of this sensitivity, the stream community metabolism approach may be able to detect a disturbance early, before it is manifest in changes in organism assemblages (e.g. macroinvertebrate community composition).

The two key community metabolism components measured are gross primary production (P) and respiration (R). The ratio of these measures provides a measurement parameter of stream status and health. Undisturbed forest stream sites are typically heterotrophic (e.g. $P/R<1$) and so are net

consumers of carbon. Sites in which the P/R ratio >1 (i.e. fundamentally autotrophic ecosystems) have been found either in cleared catchments or in sites that have been nutrient enriched. Consequently, the P/R ratio is an index with ecological meaning in freshwaters.

Community metabolism is best measured by monitoring water oxygen concentration. In systems of high or rapid metabolism, whole-river measurements can be made using the two-station (e.g. Odum 1956) or single-station technique over 24 hours (e.g. Bunn et al. 1997). In systems with low metabolic rates or high re-aeration due to turbulence (e.g. forested upland streams), closed system procedures are recommended (e.g. Davies 1997). These can be conducted over 24 hours (Davies 1997) or over short time periods in full sunlight followed by no light (Hickey 1988).

The approach uses Perspex chambers placed over the selected habitat and pushed into the substrate. Chambers require an oxygen sensor and a recirculation pump to maintain a flow similar to that outside the chamber. Oxygen concentration is recorded over an appropriate period to allow an accurate calculation of both P and R. A specific habitat can be chosen as a measurement parameter of stream health; then changes in community metabolism are more likely to result from ecological impact than from differences in habitat. Bunn et al. (1999) recommend the use of cobble habitat as the one in which variation in community metabolism best reflects catchment characteristics. Alternatively, the range of habitats present in the stream can be monitored and, using areal weighting, the metabolism of an entire reach can be measured.

The stream community metabolism approach has been applied in Jarrah forests of south-west Western Australia (WA), and in the Johnstone River in north Queensland and in the Mary River catchment in south-east Queensland.

Bunn et al. (1999) contend that gross primary productivity (GPP) could be used to infer the type of impact occurring; increased GPP indicating nutrient enrichment and catchment clearing, and depressed GPP indicating sedimentation or degradation in water quality. Initial results from the Mary River catchment suggest clear links between community metabolism and contaminants or catchment condition. There appears to be potential for development of stream community metabolism as a measurement parameter of stream ecosystem health

3.5.3.10. Quantitative Ecological Assessment

A ‘quantitative’ method refers to one that permits rigorous and fair tests of the potential impacts under consideration; typically, conventional statistical tools are employed to attach formal probability statements to the observations; see the Water Quality Guidelines section 3.2.1.3 and section 8.1.1.3.

The rationale for using quantitative methods is that they allow the use of sampling designs based on statistical inference and, hence, the explicit identification of effect size and Type I and Type II error rates. These procedures are likely to be more sensitive to subtle impacts than those based on rapid bioassessment techniques, where the effect size and error rates are implicit in the modelling procedure. In addition, quantitative procedures based on statistical designs can be adapted to local site-specific conditions.

Where possible, paired areas should be employed in an MBACI design. The current lack of knowledge about year-to-year variations in benthic diversity in many ecosystems argue for at least three years of pre-impact baseline data wherever this is possible. Quantitative sampling methods should also be used. Further details on a protocol for quantitative assessment for macroinvertebrates is provided in the Water Quality Guidelines Appendix 3, method 3A(ii).

This protocol also provides a model for development of protocols for biological indicators such as fish, macrophytes, diatoms and other groups. The key aspects are the survey design and the sampling required to obtain the statistical power needed to produce the desired results or sensitivity in the monitoring program.

3.5.3.11. Selecting Ecological Assessment Methods

If the monitoring team decides to use ecological measurement parameters, that decision will dictate the monitoring strategy and approach required and the various protocols. The team will need to check that the objectives of the monitoring program are not jeopardised. For example, rapid biological assessment methods such as that used for AUSRIVAS may not meet the quantitative assessment requirements of a site-specific study of impact assessment. This applies to the collection, processing and analysis of samples, and to the sampling design needed to meet the statistical requirements of the study. These conflicts are discussed more fully in the Water Quality Guidelines (ANZECC & ARMCANZ 2000) which explicitly distinguishes between rapid biological assessment (RBA) and quantitative analysis, and provides different protocols for these broad types of assessment.

The choice of the right method is a crucial part of the study design. Similarly the monitoring team must choose the most appropriate approaches for the problems or issues under investigation: they could be whole ecosystem approaches, river health assessments, biodiversity indices, community indices or specific indicator species or taxa for particular water quality problems. The monitoring designs that incorporate these measures will be different in each case.

3.6. Data Requirements

Once the decisions have been made about the study type, study boundaries and measurement parameters, the data requirements need to be summarised. The data requirements include the measurement parameters, scale, geographic locations and length of study, frequency, accuracy and precision. These serve as the 'concrete' instructions for the decisions that have to be made about techniques required for data analysis (Chapter 6) and for the design of specifically tailored sampling and analysis programs (Chapters 4 and 5).

3.7. Cost-Effectiveness of Sampling Programs

It is preferable for the cost of sampling programs to be as small as possible while still meeting the stated objectives of the monitoring study. Cost-effectiveness considerations involve trade-offs between loss of statistical power for discriminating between various hypotheses and the cost of data acquisition. It is necessary to determine all the resources and associated costs required, thereby ensuring the study can be carried out. Costs of data acquisition are determined by:

- the number of sampling stations;
- the number of sampling occasions;
- the replication;
- the cost of collecting samples (staff, transport, consumables);
- the cost of analysis;
- the cost of data handling and interpretation (cost of reporting).

There is extensive information available about the optimisation of sampling programs with regard to precision and cost (Montgomery and Hart 1974; Eberhart 1976; Ellis and Lacy 1980; Short 1980; Bailey et al. 1984; Lettenmaier et al. 1984; Hayes et al. 1985; Radford and West 1986; Kratochvil 1987).

3.8. Reporting Schedules

During the study design process it is important that the primary users and the suppliers of the information agree on the reporting schedules. If the expected schedules are unreasonable,

compromise arrangements need to be made. Promising more than can be delivered within a certain time places unnecessary pressure on those doing the monitoring study, while failure to report findings on time will damage relationships between the information user and the supplier.

All stages of the monitoring program will have their own time frames that must be considered when agreeing to a reporting schedule. The monitoring of a range of river flows, for example, will take months or years; with laboratory analyses, the time frames for reporting will vary significantly depending on the analyte.

The design process should consider also the reporting needs and expectations of all other stakeholders and information users. How these might be addressed is discussed in more detail in Chaper 7.