

# 7 Monitoring and assessment

## 7.1 Introduction

This chapter deals with the practicalities of collecting and analysing data for the measurement and evaluation of water quality — on the one hand, by measuring biological indicators; on the other hand, by measuring the more traditional physical and chemical indicators, including toxicants. Much of this chapter presupposes a good background knowledge of the issues involved with selecting sample sites, the timing and frequency of sampling events, and some basic principles of statistics and the design of experiments and surveys. Much of this background is provided in the companion document *Australian Guidelines for Water Quality Monitoring and Reporting* (ANZECC & ARMCANZ 2000), the Monitoring Guidelines.

The Monitoring Guidelines lays out the framework and general principles for a water quality monitoring program. Though the present chapter is self-contained in terms of its coverage of monitoring and assessment, its principal aim is to complement, not duplicate, the Monitoring Guidelines. To this end, this chapter highlights some key issues for the users of the Water Quality Guidelines that are either very specific to their needs, or that expand upon some of the general topics introduced in the Monitoring Guidelines. Sections 7.1 and 7.2 generally follow the layout of the Monitoring Guidelines while Sections 7.3 and 7.4 provide more specialist information for monitoring using biological and physical-chemical indicators respectively. The chapter is structured as follows:

- Section 7.1: Introduction, issues associated with integrated assessment, and the framework for a monitoring and assessment program with reference to the introductory steps that set the monitoring program objectives.
- Section 7.2: This section describes the remainder of the monitoring framework. Firstly, recommendations are provided for combinations of biological and physico-chemical indicators to apply to different situations. Then, some generic issues that are common to both biological and physico-chemical approaches are discussed. For example, the choice of design for a monitoring or assessment program depends partly on whether or not there are data that pre-date a putative impact and on whether or not there are appropriate control sites. Section 7.2 also recapitulates the steps needed for defining objectives and selecting candidate indicators.
- Section 7.3: A description of issues that are specific to biological indicators.
- Section 7.4: An outline of issues for physical and chemical stressors and toxicants in water and sediment. For many of the non-biological indicators, the first step is to compare test data with a guideline trigger value; the procedure is detailed in Section 7.4.4.

### 7.1.1 Integrated monitoring strategies

Traditionally, physical and chemical methods alone are used to assess water quality by indirectly estimating ecological impairment. Numerical guidelines are set according to the response of biota from different taxa to individual chemicals, derived from single-stressor toxicity tests conducted under controlled laboratory

conditions. The derivation of ‘global’ guideline values, though conceptually simple, faces a major challenge in that data derived under experimental conditions may not be relevant to complex *real world* ecosystems. Nevertheless, direct measurement of physical and chemical water quality parameters as a surrogate for ecological health has the advantages of:

- conceptual simplicity,
- established technology,
- explicit numerical objectives,
- the ability to acquire meaningful quantities of data relatively quickly,
- comparatively low costs.

Biological indicators have a shorter history of use in monitoring in Australia and New Zealand. Their development has been intellectually challenging and has evoked considerable debate. This explains in part the slower acceptance of biological indicators in environmental monitoring even though the principle is inherently sound. Biological monitoring programs, and, to a lesser extent, monitoring with physical and chemical parameters, can be labour intensive, prone to quality control failures unless special care is taken, and may require data collection over an extended period, depending on the statistical design requirements. Environmental monitoring generally, however, has developed with improvements in the way sampling is conducted and in application of appropriate statistical techniques. Appendix 4, Volume 2 contains a case study that illustrates the importance of fully optimised designs in terms of spatial and temporal controls applied to indicators.<sup>a</sup> This case study concludes by considering the balance that negotiating parties may be faced with in applying optimised designs to early detection and biodiversity indicators in an essentially unmodified aquatic ecosystem (a condition 1 ecosystem).<sup>b</sup>

*a See Sections 7.3 and 7.4 for more detail; also Chapters 3, 4 and 6 of the Monitoring Guidelines*  
*b Section 3.1.3*

As discussed in earlier chapters, these Guidelines emphasise an integrated approach to monitoring, using an appropriate mix of indicators suited to the primary management aims. Physico-chemical and biological indicators should be regarded as complementary to each other. Two issues involved in this integration are firstly, the rationale for integrated monitoring and assessment and ways to achieve integration; and secondly consideration of ways to defray costs. These are summarised briefly in turn.

#### 7.1.1.1 Enhancing inferences

*c Sections 3.1.6 & 3.2.1.1*

1. As discussed elsewhere,<sup>c</sup> it is widely acknowledged that only studies that include the biota can define or be used to assess the overall effect of waste waters on these organisms and the ecological health of ecosystems. Management goals are typically biologically-based, so organisms are the management end-point. This position holds even if the methods used for determining global numerical guidelines, including surrogates for biological end-points such as water chemical analytes, are acknowledged as having broad validity. A combination of biological and physico-chemical assessment enhances the confidence in correctly attributing causes to any observed change in water quality: biological variables integrate effects of past and present exposure and directly assess progress in achieving the management goals;

physico-chemical variables are the explanatory variables in the cause–effect relationship.

- a See box 7.1.1
- b Section 7.2.1.1/1
2. Efforts should be made, wherever possible, to examine and incorporate the results of similar types of study conducted in the region. Whether the results of the additional studies are examined alone or are combined with those from the study in question, inferences can be enhanced.<sup>a</sup>
  3. Sometimes samples may be gathered and processed in a manner that allows the results to be used for different purposes, each providing additional interpretative information. An example of this is provided below<sup>b</sup> where the advantages of combining stream macroinvertebrate samples and data from quantitative and rapid biological assessment studies are outlined.
  4. Users need to be aware always of standard operating procedures that may be in place at the regional scale and beyond. Comparison of results with those from other studies is always enhanced where a common sampling and measurement protocol is used.

**Box 7.1.1 Enhancing inferences and defraying costs in environmental monitoring programs**

Whatever the indicators used in a monitoring program, savings in resources can be made in the experimental design if data from control sites are shared amongst different bodies conducting similar monitoring programs in the region. Apart from the advantage of cost sharing, combined results can then be included in formal meta-analyses (analyses which combine the results of many similar studies) and thereby allow stronger inferences to be drawn (see also Section 7.2.5).

**7.1.1.2 Defraying costs**

The availability of resources is recognised as a major constraint in meeting the level of monitoring recommended in these Guidelines. Ways to defray costs must always be considered. Some examples to consider in this respect include:

- c See box 7.1.1
- d Section 3.1.3.2
1. As far as possible, ensure that there is a common sampling program for collection of data on different indicators. Other than providing greater interpretative value for the data gathered, this will reduce logistical costs (e.g. transport etc.).
  2. Share costs with similar monitoring programs being conducted in adjacent areas.<sup>c</sup>
  3. Incorporation of biological assessment in environmental monitoring programs may lead to cost-savings for industry if ‘no-observable-effects’ in biological responses are found, despite values for physico-chemical indicators that might be ‘high’ or which may exceed the recommended guidelines.<sup>d</sup> Use of the decision trees for physico-chemical indicators can also lead to cost-savings for industry; the first of the two case studies included in the *Introduction* to the Water Quality Guidelines provides such an example.

*a See Section 3.1.3*

*b Section 7.2.1.2/1*

*c Section 7.2.1.2/2*

Section 7.2.1.1 recommends the type and number of indicators that should be incorporated in a water resource monitoring and assessment program, depending upon ecosystem condition (condition 1, 2 or 3 ecosystems).<sup>a</sup> These recommendations need to be augmented in two special cases as described in Section 7.2.1.2. These special cases are situations where there is inadequate baseline data<sup>b</sup> and situations that call only for a broad-scale assessment of ecosystem health.<sup>c</sup>

The final balance of indicators to be measured at a site will rest with local jurisdictions and stakeholders after they have considered factors such as the nature of contaminants, the ecosystem type, the issues of concern, level of protection, availability of baseline data and resource constraints. While the constraints of resources are acknowledged, local jurisdictions still have a responsibility to ensure their water quality monitoring programs are sufficiently adequate to give unambiguous results from which confident conclusions can be drawn.

## 7.1.2 Framework for a monitoring and assessment program

Although water quality monitoring with physical and chemical indicators differs in philosophy and techniques from monitoring with biological indicators, the approaches both rely on sound practice in environmental science, including:

- explicit written definition of the sampling site, project objectives, a hypothesis and the sampling protocol that will support the work;
- the definition of sampling sites, sampling frequency, and spatial and temporal variability that will permit appropriate statistical methods to be used;
- rigorous attention to field and laboratory quality control and assurance;
- incorporation of a pilot study to test the sampling protocol and determine spatial and temporal variability.

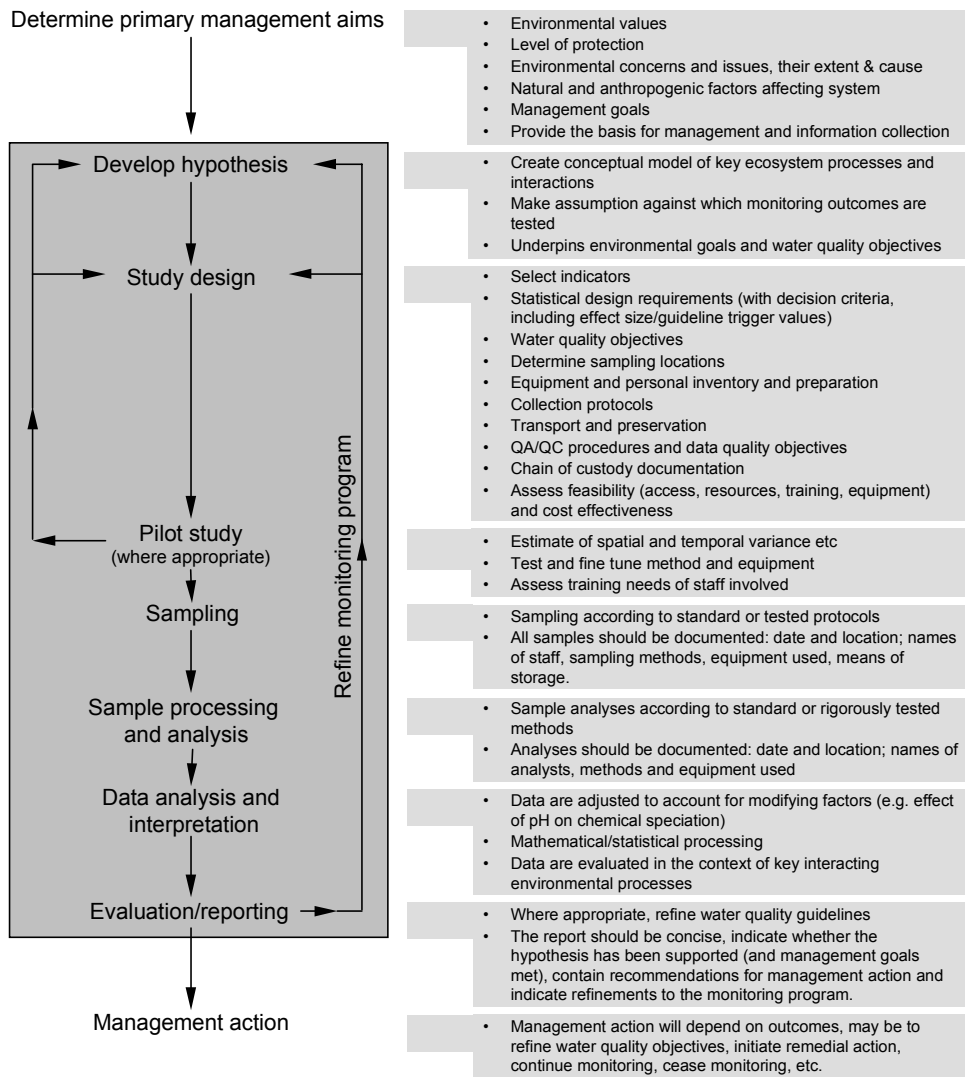
*d See Figure 2.1.1*

Figure 7.1.1 outlines the basic steps involved in developing a program for monitoring and assessing both biological and physico-chemical aspects of water quality. This figure is consistent with the framework for the Monitoring Guidelines, as portrayed in figure 1.1 of those Guidelines. The framework figure shown in the Monitoring Guidelines is necessarily general in nature while figure 7.1.1 of the current Guidelines has adapted the Monitoring Guidelines framework to incorporate aspects of the management framework outlined in Chapter 2.<sup>d</sup>

*e Chapter 2 and Section 3.1.1.1*

The first step of the framework, determining the primary management aims, has been described in earlier chapters of these Guidelines.<sup>e</sup> Determining these aims will enable stakeholders to develop an appropriate conceptual model of key ecosystem processes and interactions. By doing this they can identify assumptions against which monitoring outcomes can be tested, and develop appropriate working hypotheses whose predictions can be tested using the data that the program collects — Step 2 of the monitoring framework (figure 7.1.1). Step 2, developing a hypothesis, is discussed earlier in these Guidelines<sup>f</sup> and in Chapter 2 of the Monitoring Guidelines. Step 1 of the Monitoring Guidelines framework (figure 1.1), ‘Monitoring Program Objectives’, combines the first two steps from the Water Quality Guidelines framework of figure 7.1.1. The remainder of this chapter is concerned with the other steps in figure 7.1.1. Background information that supports the material presented here is provided in the Monitoring Guidelines.

*f Section 2.2.3*



**Figure 7.1.1** Procedural framework for the monitoring and assessment of water quality (the shaded area).  
 (Adapted from the Monitoring & Reporting Guidelines and the framework for designing a wetland monitoring program adopted by the Ramsar Wetland Convention (Ramsar Convention 1996, Finlayson 1996))



## 7.2 Choosing a study design

This next step of the monitoring framework (figure 7.1.1) includes the selection of indicators and requirements for experimental design, including the determination of guideline values. General descriptions are provided in Chapter 3 of the Monitoring Guidelines. However, the earlier chapters of the Water Quality Guidelines and its two support volumes are the main reference sources for indicator selection and determining guideline values, and these aspects are not discussed further. This section recommends a balance of indicators to apply to different situations for aquatic ecosystem protection (Section 7.2.1) and provides specific advice for experimental design using indicators from all environmental values (Sections 7.2.2 and 7.2.3).

### 7.2.1 Recommendations for combinations of indicators for aquatic ecosystems

#### 7.2.1.1 Recommendations for each ecosystem condition

This section makes some basic recommendations for the number and mix of indicators that should be used in integrated monitoring for each of the ecosystem conditions.

##### 1. Sites of high conservation value (condition 1 ecosystems)

For high conservation value sites, the goal for a water quality assessment program should include four–six of the following aspects: (i) for contaminants other than nutrients, ‘whole effluent’ toxicity testing to determine a safe dilution at which effluent may be discharged; (ii) water and sediment physico-chemistry; (iii) an ‘early detection’ indicator for either water or sediment (whichever is deemed to harbour greater risks to aquatic organisms arising from the fate and persistence of waste substances); (iv) a quantitative biodiversity indicator; and (if applicable and available) (v) a community metabolism indicator and (vi) a rapid biological assessment (RBA) indicator (see rationale below).

Ideally, for early detection (item (iii) above) a biological indicator of the type described in Section 3.2 (in particular, table 3.2.2) would be used for monitoring. It is acknowledged, however, that such indicators have at present been developed for only a relatively narrow range of conditions and regions. Until such indicators have been further developed and are more widely available, it is important, nevertheless, to adhere to the *principle* of early detection in monitoring and to consider alternative approaches to meeting this important assessment objective. For example, in some situations, adherence and responsiveness to very conservative chemical criteria and their trends may be more protective of ecosystems than even very sensitive biological tests. Alternatively or in addition, in Section 3.2.1.3/2 it was suggested that early detection and predictive capabilities would be enhanced by placing additional sampling sites for any indicator in ‘mixing zones’ — effectively measuring gradients of spatial disturbance.

The quantitative biodiversity indicator (item (iv) above) should be selected from Section 3.2. It would normally be expected that some species-level data would be gathered for relevant biodiversity indicators in regions of high conservation value. As a complement to the measurement of the quantitative biodiversity indicator, there could be situations where it would be advisable to also collect data for an RBA indicator. In some respects, results gathered for RBA can be better than

results from many quantitative approaches because they provide information about the ecological importance of effects. As stated in Section 3.2.1.3/3, RBA programs that have regional coverage and that encompass a full disturbance gradient can provide regional context for the gathered data. Data gathered for RBA indicators would not normally be expected to detect minor or subtle impacts, and for this reason they should never be measured in isolation from quantitative indicators at sites of high conservation value (nor, in most cases, at sites in slightly–moderately disturbed systems).

Measurement of quantitative *and* RBA indicators need not add significantly to the costs of a monitoring program. For example, replicate quantitative samples from stream macroinvertebrate communities at a site could initially be processed as prescribed for the AUSRIVAS RBA approach (e.g. live-sorted, see Method 3A(iii), Appendix 3 of Volume 2) and then the residue could be preserved for later laboratory processing in the usual (quantitative) manner. An initial pilot study could be required to reconcile the sampling effort needed in the field to serve both RBA and quantitative approaches. RBA data gathered from several sites would be incorporated into, and assessed against, broader regional or state/territory AUSRIVAS models.

### 2 Slightly to moderately disturbed systems (condition 2 ecosystems)

For slightly–moderately disturbed sites, the recommended water quality assessment program has the same four–six aspects prescribed in Section 7.2.1.1/1. For measurement of biodiversity indicators, species-level data may not be necessary.

### 3 Highly disturbed systems (condition 3 ecosystems)

For highly disturbed sites, it is recommended that a monitoring program includes (i) water and sediment physico-chemistry, (ii) a rapid broad-scale and/or quantitative biodiversity indicator, depending upon the nature and degree of contamination and level of sensitivity to impact required (selected from Section 3.2.2), and (iii) (if applicable and available) a community metabolism indicator.

#### 7.2.1.2 Combinations of indicators for two likely special cases

a See Section 7.2.1.2/1

b Described in Section 7.2.2 & in Ch 3, Monitoring Guidelines

c Section 7.2.1.2/2

In addition to choosing an appropriate set of indicators for an integrated program according to the ecosystem type, there are two situations that are likely to arise in many applications. In the first situation,<sup>a</sup> there are insufficient baseline (i.e. ‘pre-impact’) data to implement ‘before–after’ type sampling designs.<sup>b</sup> The second situation applies where broad-scale assessment of ecosystem health is the goal of the program.<sup>c</sup>

#### 1 Sites where an insufficient baseline sampling period is available

d Section 3.2.4.1

If it is not possible to gather sufficient baseline data, the Guidelines recommend additional monitoring, including a greater number of indicators and/or sites for ‘early detection’ and biodiversity measurement (i.e. the ‘multiple lines of evidence’ concept<sup>d</sup>). Some recent proposals to help formalise the use of ‘multiple lines of evidence’ are described in Chapter 3 (Section 3.2.3) of the Monitoring Guidelines.

e Section 7.2.1.1

i. For sites where development is planned, it is recommended that more extensive biological assessment procedures be incorporated than those outlined above.<sup>e</sup> This would include, for contaminants other than nutrients, a ‘whole effluent’ toxicity testing program for determining a safe dilution at which effluent could be discharged. For such situations, further protocols for early detection and biodiversity indicators will recommend the collection of data from a larger

a See also  
Sections 7.2.2  
& 7.2.3 below

number of ‘control’ and ‘to-be-disturbed’ sites than would otherwise be gathered, so that stronger inferences may be drawn about impact by way of disturbance gradients.<sup>a</sup>

- ii. At sites where there are existing developments, adequate baseline data were never gathered; the project approval phase pre-dated more stringent discharge licensing conditions that have subsequently been imposed by regulators. Use the same water quality assessment indicators as for Part (i) above, modified for *a posteriori* conditions.
- iii. For *a posteriori* monitoring of accidental discharges, use the same water quality assessment indicators as for Part (i) above, modified for *a posteriori* conditions.

## 2 Broad-scale assessment of ecosystem health

Applications of broad-scale monitoring procedures include assessments of biological water quality for planning purposes, the setting of goals for remediation and rehabilitation programs, and the monitoring and assessment of broad-scale impacts such as diffuse pollution. For such sites, it is recommended that a monitoring program includes (i) water and (if appropriate) sediment physico-chemistry, and (ii) data compatible with national RBA programs (e.g. AUSRIVAS).

## 7.2.2 Broad classes of monitoring design

b Section 7.4.4

This section describes the choices of broad classes of designs of monitoring programs which are available under different scenarios. Note that for the majority of the physical and chemical stressors and toxicants, the initial step in assessment is to compare data from the test waterbody or system with guideline trigger values.<sup>b</sup>

c Section 7.1.2

The design of a program for monitoring or assessing water quality is crucial. As described above, this step presupposes well articulated primary management aims and appropriate working hypotheses whose predictions can be tested using the data that the study collects.<sup>c</sup>

However, as described below, the types of program design depend on the context within which the investigation is taking place. The context can limit the choices and inferential strength of the program design.

d See the  
Monitoring  
Guidelines,  
Chapters 3, 4,  
5 and 6

There are five broad classes of program design (figure 7.2.1; modified after Green 1979). The choice depends on whether the disturbance (putative environmental impact) has already occurred, and whether any control sites are available for inclusion in the program. When designing any program for monitoring and assessment, professional statistical advice should be sought before the data are collected. All the designs outlined in this section have assumptions, and often involve sophisticated statistical procedures.<sup>d</sup>

Most water quality assessment and monitoring will take place relative to a definable event, which is called a *disturbance* in figure 7.2.1. This will often be a potential environmental impact (e.g. construction of a new outfall, change in land-use), but may correspond to change in activity to improve water quality (e.g. installation of a new treatment plant, initiation of controls on fertiliser use). If the disturbance has not already occurred, then there is scope to collect appropriate data before the disturbance; furthermore, if there are control areas or sites, then this leads to the strongest class of monitoring and assessment designs, called the ‘Before–After

a More detail provided in Section 7.3.2 and figure 7.2.1 for comparisons of similarity measures

Control–Impact’ family of designs (BACI) (case A in figure 7.2.1). Wherever possible these designs should be used, especially where the opportunity exists to incorporate appropriate controls (the so-called MBACI designs of Keough & Mapstone 1995). The logic that underpins this family of designs is described in Section 3.2.2.1 of the Monitoring Guidelines. In general terms the MBACI design (where multiple control sites are included) provides the strongest inferences. A potentially important embellishment for systems with unidirectional flow is to use matched pairs of sites (upstream and downstream) in disturbed and control locations (MBACI-P of table 7.2.1). In this scenario it is the differences between upstream and downstream sites that are compared.<sup>a</sup>

b See the Monitoring Guidelines Section 3.2 and Section 7.3.3 below

However, two common situations often arise. Either there are no appropriate control sites, in which case inferences about the event need to be based on changes through time alone (case B in figure 7.2.1); or the program has to commence after the event, in which case inferences need to be based on spatial pattern alone (case D in figure 7.2.1). Inferences based on spatial pattern alone will usually need to include reference sites or sites that provide a yardstick against which to compare the site that is being disturbed. AUSRIVAS, the rapid biological assessment procedure based on stream macroinvertebrates, can be viewed as a special case of this class of design.<sup>b</sup> Similarly, a variety of techniques can be used for basing inferences on changes through time alone.

c This is described in more detail in Section 7.3.3 below

A further case, called *a posteriori* sampling, can arise; see case B in table 7.2.1. Some chemicals and toxicants are so unusual that they can only come from human activity (e.g. some specialised pesticides, some unusual isotopes). Detection of these substances after a disturbance has occurred may be sufficient to infer environmental impact, without the need to collect any data from before the disturbance or from spatial control or reference sites. This is likely to be highly unusual, and exceptional care would need to be taken to convince all stakeholders that the substance concerned was unequivocally linked to the disturbance.<sup>c</sup> Moreover, very good evidence would need to be assembled from auxiliary studies to establish that concentrations of the substance below the detection level of the laboratory analysis were ecologically harmless.

d See also the Monitoring Guidelines Section 3.2

Occasionally, monitoring or assessment programs are initiated when the timing or location of the disturbance is unknown. This leads to two further types of study design that are not considered in any further detail in these Guidelines. *Baseline studies* (case C of figure 7.2.1) refer to those carried out before an event has occurred, where the goal is to attempt to detect unanticipated changes or trends in the environment. Broad-scale water quality monitoring networks as well as well-planned developments exemplify this approach. *Investigative studies* (case E of figure 7.2.1) are made in response to a perception that some environmental change has occurred; their goal is to determine the timing or nature of the change. Examples include studies carried out after unexpected fish kills or research programs investigating the extent and severity of acid rain.<sup>d</sup>

Finally, management for rehabilitating or restoring disturbed sites has some special problems that need to be taken into account when designing a monitoring and assessment program. They are outlined in box 7.2.1 below, ‘Issues for restoration and rehabilitation’, while box 7.2.3 outlines the related procedure of ‘bioequivalence testing’ which is appropriate for hypothesis testing in these programs.

## 7.2.3 Checklist of issues in refining program design

Once the broad category of design has been selected (Section 7.2.2 above), there are a number of issues that need to be addressed, preferably in consultation with a statistician, to refine the design and ensure that data will be collected properly for the valid application of the chosen statistical methods. (See the Monitoring Guidelines Chapter 6 for a discussion of the basic statistical issues, and Chapter 3 for discussion of site selection and the scope of the sampling program in space and time.) The following sections seek to highlight the most prominent issues.

### 7.2.3.1 Site selection and temporal and spatial scales

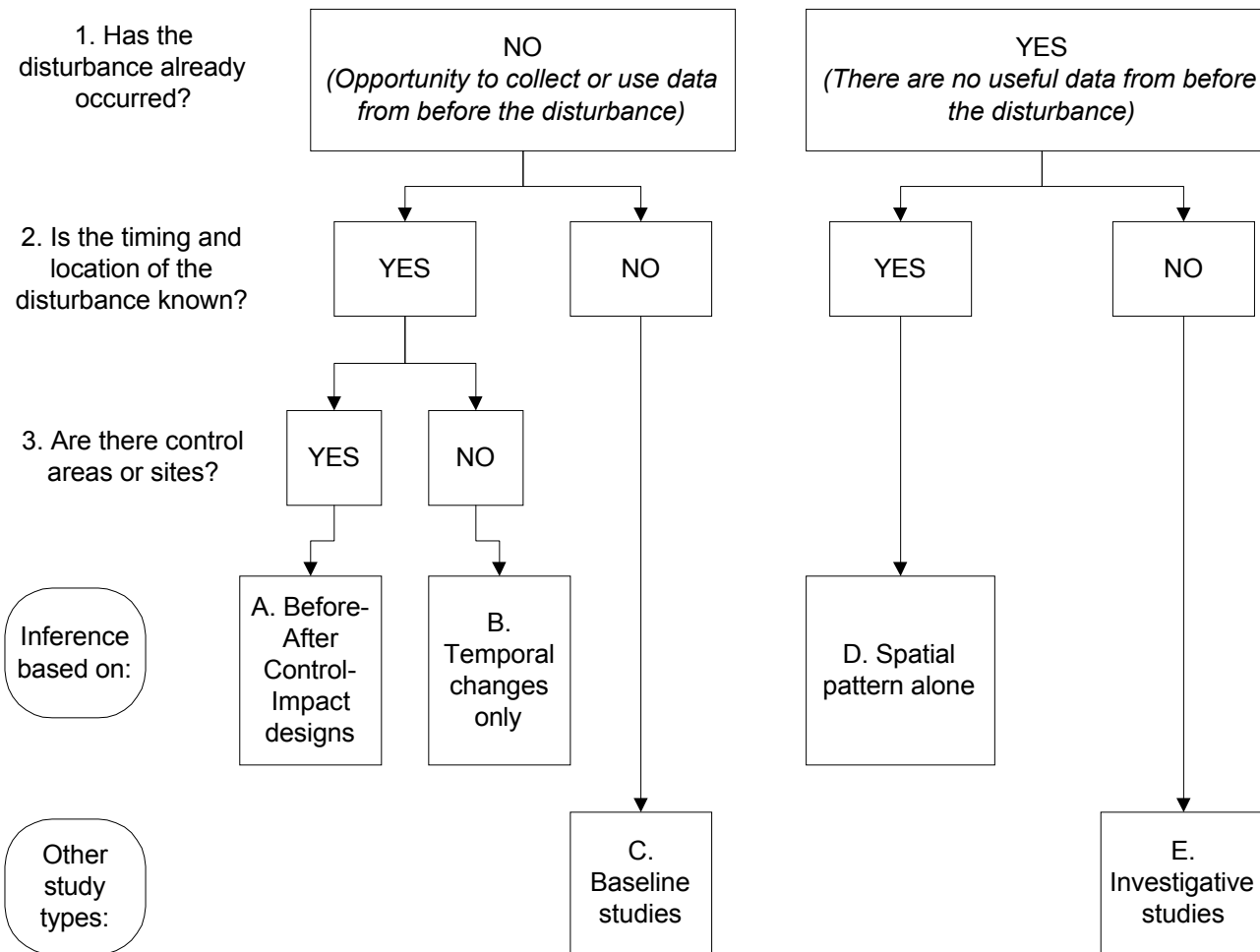
To detect impacts reliably, the size and relationship of sampling areas and the pattern of sampling in space through time need careful consideration. The assessment objective and the nature of the disturbance also affect sampling design, as well as site-specific and regional factors. It is difficult to be prescriptive, but some general guidance on the issues that need to be addressed is summarised in this section; see also discussion in the Monitoring Guidelines Chapter 3.

Independence of control and impact sites for the indicators being measured is important for all the BACI-type and spatially-based procedures (cases A and D of figure 7.2.1). If control and impact sites are too close, cross-contamination can occur which can mask changes in the indicator. What constitutes *too close* depends both on the nature of the indicator and dispersion of the pollutant. Where independence cannot be ensured, there may be procedures which can take intercorrelations between sites into account. Such procedures need expert statistical input before the data are collected.

Information on water movements is essential for planning the extent and separation of control and impact sites. Climatic and water velocity data can be combined with information on discharge and morphometry in inland waters, or data on tidal movements and oceanic circulation in marine situations, to estimate the direction and extent of mixing and dispersion of effluents. Sometimes sophisticated computer simulation models are available to assist in predicting these aspects of water movement.

Spatial variation within the site(s) to be sampled can also affect the precision of estimates of that site, which in turn can affect the outcome of any formal significance tests. Often there are distinct habitats or *strata* within the sites, and variation within the strata should be quantified in any sampling area; a single sample unit from each stratum is inadequate. Several sample units should be taken within the smallest scale of systematic variation, and often sites are sufficiently large that they require several levels of successively finer spatial resolution to be nested within each of the control and impact sites (e.g. Morrisey et al. 1992). Such sub-sampling improves the precision of the estimates of interest, and a good pilot study using a thorough, hierarchical design is essential for estimating which scales of variation are important and, consequently, the most cost effective sampling strategy likely for the final design<sup>a</sup> (theory: Sokal & Rohlf (1981), Andrew & Mapstone (1987), McPherson (1990); examples: Morrisey et al. (1992), Downes et al. (1993)). In addition, the behaviour of data is likely to be better at higher levels in a sampling hierarchy: data are more likely to be normally distributed and the influence of zeroes in the data is diminished as a result of the central limit theorem (Keough & Mapstone 1995).

*a These issues and sampling strategies to deal with them are described in Section 3.4 of the Monitoring Guidelines*



**Figure 7.2.1** Flow chart depicting the broad categories of designs for monitoring and assessment that apply in different contexts. Only categories A, B and D are discussed in detail in this document. See also table 7.2.1 and the Monitoring Guidelines Section 3.2.

**Table 7.2.1** Broad categories of design from figure 7.2.1 relevant to the Guidelines listed together with the assessment objectives that could be fulfilled by each category. Possible designs within each of the three broad categories (A, B and D) are tabulated with a brief description and commentary together with examples and references to other publications.

#### A. Inference based on the BACI (Before–After Control–Impact) family of designs

These designs are suitable for the following assessment objectives:

- early detection,
- biodiversity or ecosystem-level response.

Where comparable control sites exist and there is sufficient lead time before the disturbance, the MBACI design should be preferred unless the prevailing situation requires one of the other BACI designs described here. The general logic of the BACI family of designs is outlined in Section 3.2 of the Monitoring Guideines.

Possible designs	Description and notes	Examples and references
MBACI	<p>Before–After Control–Impact design with Multiple control areas and (if possible) &gt;1 impact area. Preferred design because of increased confidence that differences between control and impact areas are not due to peculiarities between single control and impact areas.</p> <ul style="list-style-type: none"> <li>– May be modified to MBACI-P (where P stands for pairing of sites) if indicator is best expressed in terms of differences between paired sites.</li> <li>– Short-term and long-term impacts require careful planning of frequency of sampling. Variation/trends amongst areas/times may be modelled using regression, covariates, or dynamic simulation and permutation methods.</li> </ul>	<p>The general principles behind this design are outlined in Section 3.2 of the Monitoring Guidelines (see also figure 3.3 in that document). Keough &amp; Mapstone (1995; 1997) provide a full description and discussion.</p> <p>Faith et al. (1995) discuss principles of MBACI-P designs.</p>
'Beyond BACI' designs	<p>Elaboration of MBACI designs with additional nested components in time and/or space. Appropriate where the spatial and/or temporal scale of the impact is unknown or where changes in the pattern of <i>variation</i> of the indicator are more important than detecting changes in the average value of the indicator.</p>	<p>Underwood (1994) describes the most elaborate models based on ANOVA; general principles could be extended to other statistical techniques with more flexible assumptions (e.g. general linear models).</p>
BACIP (single control site)	<p>Before–After Control–Impact, Paired differences. Applicable if there is limited scope for spatial replication (e.g. one 'control' and one 'impact' site). Required if seasonal or other temporal changes in response are known to occur <i>OR</i> if temporal behaviour of response is unknown. Differences may consist of multivariate dissimilarities.</p> <ul style="list-style-type: none"> <li>– Set of differences before and after impact compared using Student's t-test (or equivalent) (Appendix 4, Vol 2).</li> <li>– Modelling trends or thresholds and/or inclusion of covariates (Appendix 4, Vol 2).</li> </ul>	<p>Illustrated in figure 3.4 of the Monitoring Guidelines. Described in detail by Stewart-Oaten et al. (1986; 1992). Example provided in Humphrey et al. (1995).</p> <p>Faith et al. (1995) discuss multivariate modification.</p>
Modifications		
Simple BACI	<p>Before–After Control–Impact; only one sampling event prior to impact. Applicable only if seasonal or other temporal changes in the indicator have been demonstrated to not occur.</p> <ul style="list-style-type: none"> <li>– Of dubious value because only one sampling event prior to the disturbance leads to a high chance of confounding with natural changes unrelated to the disturbance.</li> </ul>	<p>Described by Green (1979) and in figure 3.2 of the Monitoring Guidelines. This design critiqued by Hurlbert (1984) and Stewart-Oaten et al. (1986).</p>

**Table 7.2.1** (continued)**B. Inference based on temporal change alone**

These designs are suitable for the following assessment objectives:

- early detection,
- biodiversity or ecosystem-level response.

These designs should be used if no comparable control sites exist. They assume that any changes in the behaviour of the indicator after the disturbance are solely attributable to the disturbance (see Section 3.2 of the Monitoring Guidelines). Other lines of evidence, such as would be gathered under an integrated monitoring program, would strengthen inferences from these designs (see table 3.2 in the Monitoring Guidelines).

Possible designs	Description and notes	Examples and references
Intervention analysis	Disturbance is regarded as an intervention and applicable when a long time series of data has been collected before the supposed impact which can be used as a baseline to compare to data collected after the disturbance. Applicable when no suitable control sites can be found that are comparable with the supposed impact site.	Welsh & Stewart (1989) and Thompson et al. (1982) exemplify intervention analysis applied to chemical and biological indicators respectively.
Trend analysis	Objective is to describe any trend in the chosen indicator. There are several methods that can be used to estimate trends, including those below. <ul style="list-style-type: none"> <li>– Time series analysis in which, if the data sequence is long enough and sampling sufficiently frequent, temporal autocorrelations can be modelled and treated appropriately.</li> <li>– Control charting and allied techniques derived from statistical process control can be used to measure changes of means and variance of the values of an indicator relative to notional 'action thresholds'.</li> <li>– GAMs (Generalised Additive Models) relatively new, advanced group of procedures which replace linear functions with unspecified 'smoothers' that are suggested by the data themselves.</li> </ul> <p>Robust smoothing is useful for displaying trends in data with extreme values or outliers.</p>	See Section 6.3 of the Monitoring Guidelines for brief, general descriptions and references for all the techniques listed here. <p>Gilbert (1987) and Galpin &amp; Basson (1990) provide overviews of the complexities of applying trend analyses to water quality data.</p>
<i>A posteriori</i> sampling	Applicable only if measured response (especially chemical or biochemical marker) is unequivocally related to the effluent (Section 3.2.3; otherwise not elaborated upon).	

**Table 7.2.1** (continued)**D. Inference based on spatial pattern alone**

These designs are suitable for the following assessment objectives:

- biodiversity or ecosystem-level response
- broad-scale assessment

These designs assume that disturbed sites and undisturbed sites had similar values of the indicator before the disturbance (see Section 3.2 of the Monitoring Guidelines). Other lines of evidence, such as would be gathered under an integrated monitoring program, would strengthen inferences from these designs; see table 3.2 in the Monitoring Guidelines.

Possible designs	Description and notes	Examples and references
Conventional statistical designs (e.g. ANOVA, ANCOVA)	<p>Comparisons are made between disturbed and undisturbed sites.</p> <ul style="list-style-type: none"> <li>– Pairing of sites upstream and downstream of disturbance and comparison of these differences with differences from matched pairs in undisturbed water bodies can strengthen the inference.</li> <li>– Matching disturbed site(s) with undisturbed site(s) is essential but sometimes difficult. Use of covariates can assist in adjusting for moderate background differences between sites.</li> </ul>	<p>Discussed by Underwood (1993); examples described by Green (1979)</p> <p>Davies &amp; Nelson (1994) provide an example comparing differences between matched paired sites on streams subjected to different forestry operations.</p>
	<ul style="list-style-type: none"> <li>– For multivariate indicators (e.g. measures of community similarity) analysis of similarity (ANOSIM) techniques are available for some basic designs. Future developments in permutation and randomisation testing are likely to expand the complexity of designs that can be analysed.</li> </ul>	<p>Legendre &amp; Legendre (1998) provide a general overview of a variety of multivariate techniques appropriate for similarity data. Clarke and Green (1988) and Clarke and Warwick (1994) explain ANOSIM and give some examples.</p>
Analysis of 'disturbance gradients'	<p>Several sites can be identified with a range of severity of the disturbance. Inferences are drawn from correlation of disturbance (or surrogate disturbance) variables with values of the indicator. A variety of techniques can be used including those below.</p> <ul style="list-style-type: none"> <li>– Regression relates the strength of disturbance to the response of the indicator.</li> <li>– Spatial statistical designs and methods can be useful where the inference is based on estimating parameters collected over a contiguous area. The sampling intensity for such methods is often demanding.</li> <li>– For multivariate indicators, spatial statistical techniques are becoming available based on permutation and randomisation tests. Again the sampling intensity can be demanding.</li> </ul>	<p>Basic description provided in Section 6.5 of the Monitoring Guidelines.</p> <p>Cressie (1993) and Rossi et al. (1992) detail some of the conventional spatial statistical techniques. Thrush et al. (1994) provide an example from marine benthos.</p> <p>Legendre &amp; Legendre (1998) provide a recent overview of a range of promising techniques with copious references to published applications.</p>
Predictive models based on spatial controls only	<p>Detection and assessment by predictive modelling (e.g. AUSRIVAS).</p> <ul style="list-style-type: none"> <li>– At present only AUSRIVAS for macroinvertebrates in rivers and streams has been developed. This method relies on a network of reference sites against which test sites (those thought to have been disturbed) are compared. Test sites may not have been sampled contemporaneously with reference sites, so this method makes a large assumption that there is low inter-annual variation in the family-level composition of macroinvertebrate communities. Because inferences are based on family-level spatial data, this method is likely to be sensitive to only moderate to large impacts.</li> </ul>	<p>AUSRIVAS is outlined in Section 7.3.3 and by Schofield &amp; Davies (1996); it is derived from the British RIVPACS system, the mechanics of which are described by Wright (1995).</p> <p>The applications and limitations of AUSRIVAS in the context of these Guidelines are described in Sections 3.2 and 7.3.3.</p>

**Box 7.2.1 Issues for restoration and rehabilitation**

Many of the principles identified in the accompanying sections also apply to designing programs for monitoring and assessing the extent of biological recovery after an environmental impact has occurred. The formal process of setting criteria for making decisions (Section 7.2.3.3) often receives little attention in rehabilitation and restoration programs and is an area meriting further exploration (e.g. Maguire 1995).

In the majority of rehabilitation and restoration programs, there will not be reliable data collected over long time periods before the environmental impact. Thus the main problem in setting the criteria for making decisions for such programs lies in defining appropriate targets for the chosen indicators by which the success of a program can be judged. If there are no pre-disturbance data at all, then the sampling program should include appropriate undisturbed sites that can act as reference sites for the disturbed area. This, of course, entails making assumptions about similarity in behaviour of the indicator over time in the affected area and the control areas in the absence of the disturbance (Section 7.2), and there is a danger that the reference sites will not represent a realistic target for the affected area (Wiens & Parker 1995). Furthermore, there are likely to be situations where there are no appropriate reference sites, and the target reference condition will need to be set by other means (Section 3.1.4). Setting targets in these situations is difficult, and will often involve subjective judgements from expert panels and/or stakeholders. For example, suppose a target value is set for an indicator and, after the prescribed time since rehabilitation, the indicator has still not reached the target value; there are no logical grounds for determining whether the rehabilitation has failed or the target was set too high.

In all cases, there will need to be extensive liaison between managers and stakeholders to ensure that appropriate indicators are selected and that targets are appropriate for the constraints and context of the impact under consideration (Maguire 1995). Within the framework provided in Chapter 7, the following four issues need to be considered.

First, the indicators selected will need to accurately reflect the nature of the change desired. Rehabilitation programs sometimes can concentrate on obvious, but inappropriate indicators. Norris (1986) provides a salutary example where remediation of a disused mine site focused on obvious terrestrial and riparian works (as indicators of remediation success) which did not result in any improvement in the biological attributes of the river. The nature of the desired change will also depend on the time-lags between implementing a management action and the response of the indicator. For example, changes to land use on a catchment may take longer to result in a change in algal community composition than closing a sewage outfall; thus sampling programs and decision criteria will need to be geared towards gradual changes in the former and relatively abrupt changes in the latter.

Second, the context of the desired change needs to be considered in concert with the size of the effect that needs to be detected so that timely alterations to the management of the remediation program can be made (Section 7.2.3.3). For example, a program to assess the success of a clean-up operation after an accidental oil spill will need tightly specified effect sizes and timelines if legal action about compensation payments depends on the success of this operation. By contrast, the rehabilitation of a large mine area that has been a source of serious pollution for many decades may need intermediate goals as various phases of the rehabilitation process are implemented and their success is assessed. As a result, timelines and targets may need to be re-set as rehabilitation proceeds.

Third, the relative risks and cost of committing a Type I or Type II error need to be considered carefully (Section 7.2.3.3), especially in circumstances where pre-impact baseline data are limited and/or control or reference areas are few (see box 7.2.3, 'Application of bioequivalence testing', for the meaning of Type I and Type II errors under this form of hypothesis testing).

Fourth, the choice of analytical procedures and the scope of the conclusions (Sections 7.2.2 and 7.2.6) will be limited by the availability of appropriate reference or control data. In some cases, where multiple sites are to be rehabilitated over long time spans, MBACI designs could be implemented (Section 7.2.2), although the use of bioequivalence testing procedures under such complex designs may need further statistical development (McDonald & Erickson 1994). Conversely, monitoring the recovery of an indicator after an isolated accident such as a toxic chemical spill limits the potential analytical options and strength of the inferences (Wiens & Parker 1995, see also Section 7.2.2).

*a Section 3.4, Monitoring Guidelines, also discusses these issues*

Just as spatial patterns must be considered in sampling design, so must patterns in time. These patterns may be predictable (e.g. periodic behaviour of tides) or episodic (e.g. floods), and range in scale from many years (e.g. El Niño–Southern Oscillation) to diurnal or even shorter time-scales. Again, as with spatial variation, failure to account for temporal patterns can confound impacts with natural events, and similar sampling strategies are called for to estimate these patterns.<sup>a</sup>

### **7.2.3.2 The importance of good pilot data**

Sampling programs can be costly, and it is important to try to optimise the sampling program so as to address the hypotheses posed by the program (the feedback loop in figure 7.1.1). Good pilot data collected before the monitoring program commences are therefore highly desirable in the absence of a validated historical database. Note that the number of samples acquired in pilot programs should be as large as is feasible to provide accurate estimates of variation; pilot data using small sample numbers yield unreliable information that may lead to poor decisions in optimising the sampling program. The design of the pilot sampling protocol must be as detailed and thoughtful as for the main project, though it should be remembered that to refine a sampling protocol is one of the principal objectives of a pilot study. Optimisation decisions based on a well designed pilot study will be more soundly based and hence defensible. Another advantage of a pilot study is that it gives field staff site-specific training, and allows anticipation of potential hazards and logistical problems. Most practitioners recommend that a significant fraction of total project resources should be dedicated to a pilot study; Keith (1991) recommends 10–15%.

### **7.2.3.3 Setting criteria for decisions**

The values of indicator variables usually respond to disturbances in a continuous fashion (e.g. the ‘dose–response relationship’ of toxicology). As explained in Section 3.1.7, somewhere along the continuum a value of an indicator needs to be chosen which forms the criterion for making a decision which will precipitate some management response.

This section outlines the procedures for setting such decision criteria, in three steps. First it explains the use of hypothesis testing in this process; then it describes the three stages that need to be addressed when setting decision criteria. In outline, the first stage involves deciding what sort of change to look for in the indicators, in the context of the environmental assessment objectives. The second stage involves translating this change in the indicator into a quantifiable effect size. The third stage involves assessing the risk of making a Type I error (giving false alarm) or a Type II error (giving false sense of security)<sup>b</sup> in the light of the consequences or costs of making either of those errors (in a purely scientific and/or social value sense). It is important that these three interconnected stages are discussed and iterated with the stakeholders interested in the results of any monitoring or assessment program. The negotiations should be undertaken before implementing a monitoring or assessment program so that effect sizes, error rates and costs are identified explicitly. It is also best to discuss several indicators simultaneously in this process because, inevitably, some indicators may prove to be more cost-effective than others in detecting change.

*b See Sections 3.1.7 and 3.2.4*

*The use of hypothesis testing*

*a See also the Monitoring Guidelines Sections 2.4.2 and 6.4.2; the latter touches on alternative procedures*

These Guidelines generally adopt a statistical hypothesis testing approach to determine whether the values of the chosen indicators have exceeded guideline values. Users should be aware that hypothesis testing is not the only statistical procedure that can be used in making inferences from water quality data.<sup>a</sup> (Note that this is a separate issue to the requirement for general working hypotheses such as those described in Section 7.1.2 above that identify key assumptions against which monitoring outcomes can be tested.) Some background on the criticisms of hypothesis testing and the rationale for using it in water quality monitoring are given in box 7.2.2, ‘Hypothesis testing in environmental monitoring and assessment’.

### **Box 7.2.2 Hypothesis testing in environmental monitoring and assessment**

There has been some argument against the use of hypothesis testing tools in environmental assessment programs (e.g. Suter 1996). While hypothesis testing is not always either necessary or appropriate, much of the argument about its use (or misuse) is mis-directed. The argument is, in part, that hypothesis tests will only tell us after the event that something ‘dreadful’ has happened. However, the real issue is not whether hypothesis testing is appropriate, but whether the criteria by which tests are made (and for which sampling programs are designed) are sufficient or appropriate. An appropriately designed and executed sampling program intended to detect early warning signals will provide early warning whether analysed through hypothesis testing models or other procedures. Therefore it is important to make a satisfactory definition of objectives and decision criteria for each monitoring program. In real life there is a continuum or spectrum of conditions from undisturbed to disturbed; defining statistical boundaries along this spectrum to specify changes that are ‘acceptable’ and unacceptable to the stakeholders (as is done by the AUSRIVAS model bands) is strongly advocated, especially where clear ‘break-points’ in the meaning of ecological variables are not well documented.

A related issue is whether the inferences of impacts or changes should be based on dichotomous alternatives or a continuum of conditions. Suter (1996) and Stewart-Oaten (1996a,b) infer that hypothesis tests are constrained to test only two alternatives. However, there is no reason why those alternatives cannot be but two of a range of conditions along a continuum, the test being to (perhaps progressively) detect whether a response variable has moved from one condition to the next. In this case, the dichotomous test would be used only to test whether a particular threshold along a continuum had been crossed.

In summary, it is most important to choose appropriate ‘performance criteria’ for impact assessments or monitoring programs. If the criteria by which a management action will be triggered are inappropriate or insensitive or too coarse, then the issue of which tool to choose for statistical analysis becomes irrelevant.

For hypothesis testing to be useful in making decisions, the user needs to negotiate how much change in the indicator represents ‘background noise’. In formal terms this means stipulating the null hypothesis (‘no change’) in terms of an effect size, as explained in the next section. That is, the null hypothesis is best thought of as the condition representing no *important* change in the value of the indicator, where ‘importance’ is determined by the context of the problem being monitored or assessed. Similarly, Type I and Type II errors are minimised by setting a suitable level of statistical significance when testing differences or change.

a See Section  
3.1.4

Some management programs will be oriented towards restoration or rehabilitation. In these circumstances the monitoring program will be seeking to prove that the values of the indicators are similar to those defined by the reference conditions.<sup>a</sup> In formal terms such programs will be trying to prove the null hypothesis (no ‘important change’ from reference conditions), although this is formally impossible. Hypothesis testing frameworks have been developed for such situations (they are sometimes called ‘bioequivalence tests’ in medicine and toxicology) and these are outlined briefly in box 7.2.3, ‘Application of bioequivalence testing’.

### **Box 7.2.3 Application of ‘bioequivalence testing’ for environmental restoration**

Where statistical hypothesis-testing procedures are being used to analyse the data, it may be useful to re-cast the test using the framework of ‘bioequivalence testing’. This procedure has been used in medical contexts (e.g. Westlake 1988, Chow & Liu 1992) and has recently been applied to environmental restoration in the USA. It is clearly explained by McDonald and Erickson (1994).

The problem with testing for recovery using a conventional hypothesis test is that the investigator is attempting to ‘prove’ the null hypothesis that the selected indicator in the disturbed site(s) has the same value as in the control or reference sites. However, failure to reject a null hypothesis does not constitute proof.

Tests of bioequivalence solve this problem by recasting the question so that the undesirable outcome, that the disturbed site differs substantially from the reference (i.e. the sites are *not* ‘bioequivalent’), becomes the ‘null hypothesis’<sup>15</sup> and evidence is sought to reject this hypothesis in favour of the alternative, that the impacted site is similar to the reference (i.e. the sites *are* ‘bioequivalent’). Formally, the hypotheses are framed in terms of the ratio of the values of the indicator in the disturbed site and the reference site. If recovery has been achieved, the ratio should be sufficiently close to 1, and there should be strong evidence against the ‘null hypothesis’ which is then rejected in favour of the alternative after conducting the appropriate statistical test.

Under bioequivalence testing, a Type I error results in incorrectly deciding that the sites are bioequivalent when they still differ by an important amount (i.e. inadequate recovery, a false sense of security), whereas a Type II error results in deciding that the sites still differ when in fact they are similar (i.e. adequate recovery, false alarm). Note that with this technique stakeholders still must negotiate an effect size; users need to stipulate how different sites can be before they are declared ‘non-bioequivalent’. In formal terms a critical value of the ratio of the indicator between the sites needs to be stipulated.

#### ***Stage 1: The nature of the change and its context; the use of hypothesis testing***

The criteria used for making a decision depend on the level of protection assigned. As explained in Section 3.1.3 the level of protection depends on the condition of the ecosystem (condition 1, condition 2 or condition 3); specific guidance on how the level of protection affects decision criteria is given in Section 3.1.3.2 and table 3.1.2, while Section 3.1.8 elaborates on condition 3 ecosystems.

<sup>15</sup> Technically, the term ‘null hypothesis’ is usually reserved for the equivalence of a test statistic under different conditions, whereas in a bioequivalence test the investigator is quantifying the evidence against a proposition of non-equivalence under the different conditions.

In real life there is a continuum or spectrum of conditions from ‘undisturbed’ to ‘disturbed’. Any indicator’s response to disturbance is likely to vary according to the strength of that disturbance, whereas the decision about whether or not an impact has occurred is a point on that continuum. In other words, managers and stakeholders need to determine how much change from the unimpacted or pre-impact condition is acceptable.

The environmental assessment objectives (table 3.2.1) determine the point along the continuum at which an environmental impact is deemed to have occurred. For example, monitoring based on early detection of impact will have a different emphasis from monitoring geared towards assessing the ecological importance of an impact that has already happened. For early detection, a decision must be made *before* the level of change becomes harmful; otherwise the change may be irreversible. By contrast, to assess the importance of, say, an accidental ecological impact, the monitoring team must decide whether the level of acceptable change has been exceeded and by how much. In this situation the decision criterion is at the point of harmful change rather than some smaller value. In general, however, the emphasis will be on setting the decision criteria at a level that prevents harmful effects from occurring in the first place.

*a See Section 2.1.3 for management goals*

Thus a very important part of setting decision criteria is knowing what management actions will be taken if an impact is detected. The management goals<sup>a</sup> that managers have established provide most of this context, and some of the issues that may affect these goals are outlined in Section 3.1.3.3. For example, if a condition 2 ecosystem is being managed to conserve the population of a recreationally important fish, and the threat is a persistent contaminant with the potential to reduce the fecundity of the fish, then the decision criteria for the water quality indicators need be set at values which are smaller than those which begin to affect the reproduction of the fish; this will allow sufficient lead time for managers to act before the population of fish are affected. Much scientific judgement is involved in this process, and the actual values used as decision criteria will depend on a number of modifying factors (e.g. chemical speciation of toxicants); such matters are covered in more detail for each of the broad classes of indicators in Sections 3.2–3.5.

*b Section 3.1.3.1 and table 3.1.2  
c, d See also footnote 2, page 2–9*

For ecosystem condition 1 (high conservation/ecological value) a criterion of ‘no change beyond natural variability’ is prescribed for biological indicators, physical and chemical stressors and sediments.<sup>b</sup> Operationally, this still requires users to stipulate how much change can be expected under ‘natural’ conditions, because this natural variation constitutes an acceptable level of change in the ecosystem.<sup>c</sup> Note that it is still necessary to decide on an effect size (see the next sub-section) explicitly and to ensure that sampling is intensive enough to detect effects larger than the acceptable natural changes in the chosen indicators, and avoid Type II errors. Note that the determination of the acceptable level of change may have both scientific and social elements.

For those who are new to environmental assessment, defining an acceptable level of change may seem weak, especially when management insists there must be ‘no change’ in the indicator. A criterion of ‘no change’ cannot be used operationally because it requires the user to prove the null hypothesis — an impossibility, as mentioned above. However, it is possible to state some level of change in an indicator below which it is not important to reject the null hypothesis of ‘no change’.<sup>d</sup> This requires stakeholders to be explicit about what level of change in

the indicator is regarded as harmless or acceptable. In formal terms this process involves specifying an effect size, which is described in the next sub-section.

**Stage 2: Specifying the size of the effect**

The values of all the indicators used in these Guidelines vary naturally in space and time, and estimates of their true values can only be made via samples. Accordingly some observed changes in the indicators are likely to be ecologically trivial. The problem for water quality monitoring is to detect *non-trivial* changes in the chosen indicators soon enough to allow management to act. This means that monitoring programs need to be sensitive enough to detect modest rather than large changes in the indicators.

In formal terms, therefore, we need to identify the maximum amount of change in the indicator that is tolerable before we reject the null hypothesis (no important change) in favour of the alternative hypothesis (important or unacceptable change). This level of ecologically important change is sometimes called the *critical effect size*,<sup>a</sup> but for brevity we refer to it as the *effect size*.<sup>b</sup> Some of the procedures in these Guidelines have an implicit effect size; the relationship of guideline trigger values to the concept of effect size is described at the end of this subsection.

a See box 2.3  
in Section 2.2.1

b See box  
7.2.4

**Box 7.2.4 Effect sizes are implicit in some procedures**

For some procedures, the effect size and error rates tend to be implicit in the methods and are less amenable to the procedure of using scalable decision criteria described in this section.

For example, when comparing test data to a guideline trigger value, the 'effect size' may be implied by the choice of percentiles used in the comparison. See Section 7.4.4 for a full discussion of this and the trade-offs between Type I and Type II errors made in this procedure.

Similarly, in the AUSRIVAS procedure for rivers, notions of effect size and error rates are inherent in the way the summary indices are compared with the reference conditions. See Section 7.3.3 for more discussion.

There are two components of effect size: its form and its magnitude (Cohen 1988, Mapstone 1995). The form of an effect is the statistical measure (e.g. mean, variance) that is expected to differ between control and impact sites, and the pattern of differences or trends that it is necessary to detect (Stewart-Oaten et al. 1986, Green 1989, Underwood 1991a,b). The magnitude of an effect is the size of the difference or change in mean, say, or variance that would be considered important.

It is difficult to be prescriptive about effect sizes in ecological assessment for two reasons. Firstly, there is little information about the relationships between contaminants and biological indicators in field conditions, especially in Australia and New Zealand. Secondly, the degree of change that is important depends on the environmental and social values that stakeholders are seeking to protect. Strategies for setting an effect size are discussed in box 7.2.5, 'Some suggestions for setting effect size'. This is not an exhaustive list, and other strategies may arise as experience in planning programs with these procedures increases.

**Box 7.2.5 Some suggestions for setting an effect size**

Where an indicator has intrinsic socio-economic value (e.g. it is a commercially or recreationally important species), then effect sizes can be set to ensure sustainable use of that indicator. However, many biological indicators have been selected because they are more sensitive than commercial species or because they are thought to be *ecologically* important rather than of economic value. For example, seagrass is not used directly by humans in Australia and New Zealand, but is an important indicator because of the habitat it provides and the number of species it supports.

Existing research, or similar impacts, preferably in comparable regions, can provide information about the relationship between the indicator and size of potential impact, especially if existing impacts can be found on a gradient from mild to extreme. For example, a variety of sewage treatment plants may be present in a river basin with differing degrees of sewage treatment. Pilot data relating indicator levels and type of treatment could be used in stakeholder consultations to correlate stakeholders' expectations of acceptable sewage treatment with change in the indicator. In some cases, simulation models can use these data to estimate how much an indicator might change under different scenarios.

For many indicators in ecosystems in Australia and New Zealand, however, such background data are unlikely to be available. This will inevitably involve some judgement by the planners of a program, and an arbitrary but conservative effect size will need to be specified (e.g. Humphrey et al. 1995). This should be done explicitly, and at the beginning of the program. Any change to the effect size later in the program must be openly and explicitly negotiated and fully justified on scientific grounds.

Once the level of acceptable change has been negotiated, the degree of change in the indicator may need to be set to a smaller value so that management actions can be implemented before harmful and irreversible effects occur. When the effect size is being set, such issues as the fate and persistence of the contaminant and time-lags between a contaminant event and a measurable change in the biological indicator should be considered. Allied to these issues are selection of appropriate indicator(s),<sup>a</sup> and assessment of the relative costs of erroneously missing an effect of the stipulated size (Type II error) and erroneously concluding an impact occurred when, in fact, it did not (Type I error) (see next subsection).

*a See Section 8.1*

For the non-biological indicators in Sections 3.3–3.5, the guideline trigger values listed are the best currently available estimates of ecologically low-risk levels for those indicators.<sup>b</sup> These trigger values make an implicit statement about effect size: data from the test waterbody which are lower than the trigger value are thought to pose little risk to the ecosystem. Depending upon the management goals, stakeholders may need to negotiate different trigger values. There will also be situations where trigger values are exceeded. In these cases, more complex monitoring designs are called for,<sup>c</sup> and the steps outlined here for negotiating effect sizes will need to be followed.

*b Section 7.4.4*

*c Section 3.1.5*

**Stage 3: Specifying the error rates relative to the costs of those errors**

Having stipulated an effect size, the stakeholders then need to minimise the risk of two potential outcomes — in statistical terms, the Type I and Type II errors. These errors can arise because the indicators we use are sampled rather than measured completely, meaning that we are working with information which is necessarily incomplete. The first potential error is to declare that an impact has occurred

a See box 2.3  
in Section 2.2.1

(i.e. the effect size has been exceeded), when really there has been no actual change that was ecologically important. The second potential error is to miss an ecologically important change. The probabilities of each error are conventionally denoted by the Greek letters  $\alpha$  (for Type I) and  $\beta$  (for Type II).<sup>a</sup>

The challenge is to ensure that sufficient data are collected to detect the change stipulated in the effect size while, on the other hand, not expending too many resources on sample sizes that will detect ecologically trivial changes in the indicator. Inevitably, resources are scarce, so all monitoring programs will need to balance these two errors.

Conventionally,  $\alpha$  has been set at 0.05 or smaller and few programs have stipulated  $\beta$  (Toft & Shea 1983, Fairweather 1991, Mapstone 1995). Although some recommendations for  $\alpha$  and  $\beta$  are conservative default values for ecosystem conditions 1 and 2 (for biological indicators in Section 3.2.4), it must be emphasised that ideally these error rates should be *negotiated* rather than accepted uncritically. The most important part of this negotiation is to ensure that the *balance* between these two types of errors is acceptable to stakeholders in the process. To this end, these Guidelines recommend Mapstone's (1995, 1996) proposal that the *ratio* of these two errors is negotiated as part of refining the design of a monitoring program. This process requires iteration between stakeholders, but should be transparent, accountable and, above all, should take place before the final monitoring or assessment program is put in place (Mapstone 1995).

In outline, the choice of  $\alpha$  and  $\beta$  involves four steps. First, establish the relative importance or cost of the consequences of each type of error. Second, set the ratio of the critical Type I and Type II errors relative to their costs (if there is insufficient information to estimate the costs of the errors, Mapstone suggests they should be weighted equally). Third, negotiate desired values of  $\alpha$  and  $\beta$  with reference to the ratio established in the previous step with the stakeholders. Fourth, design a sampling program to meet the desirable Type II error rate,  $\beta$ , established in the previous step, given the effect size which has been specified earlier; this allows the sample size and details of the design to be finalised. Mapstone (1995) details two alternative decision procedures that can be followed once data have been collected and analysed.

Ideally, these negotiations should include a number of potential indicators simultaneously. In the process of balancing Type I and Type II errors, some indicators will inevitably prove much more costly than others if the two error rates are to be kept low. In such cases, stakeholders are faced with a choice: either discard the costly indicators in favour of those that will detect the stipulated effect size more cheaply, or, if the costly indicators have to be included in the program for some reason, increase the sizes of the two errors while maintaining the *ratio* between the errors. The only way to reduce the sizes of these errors is to increase the sampling intensity. Maintaining the ratio between the errors ensures that Type I errors are not minimised at the expense of increasing Type II errors — i.e. that the monitoring program does not lose power to detect an important change at the expense of being conservative about the probability of incorrectly declaring that an important change has occurred.

In two situations in these Guidelines, this negotiation of the balance between Type I and Type II errors is implicit; these are outlined in box 7.2.4.

The choice of the best sampling program is not a trivial issue. The strongest evidence will come from designs that have extensive baseline data collected before the suspected or potential impact takes place, and will involve simultaneous monitoring in multiple control sites. The weakest evidence will result from programs with limited or no pre-impact data. In all situations, inferences and assessments can be strengthened by including multiple lines of evidence.<sup>a</sup> The power of any statistical tests employed may be improved by including multiple indicators in a multivariate analysis, depending on the pattern of responses amongst the indicators (Green 1989).

<sup>a</sup> See Section 3.2.4.1

## 7.2.4 Sampling protocols and documentation

From figure 7.1.1, once the sampling program has been finalised, sampling can begin. This should take place according to standard or tested protocols. Section 8.1 and Appendix 3 of Volume 2 provide a list of protocols for biological indicators, while Section 8.3.6 outlines sources for protocols to be used in direct toxicity tests for toxicants. Procedures for sediment toxicity testing seem to be less well developed, but references to and guidance through the recent literature are provided in Section 8.4.3. Protocols for measuring physical, chemical, biological and ecotoxicological parameters of sediments are described in general terms in Section 3.5 and Appendix 8 of Volume 2, and Chapter 4 of the Monitoring Guidelines, with references to detailed literature.

Quality assurance and quality control (QA/QC) procedures should be part of any sampling protocol. Quality control (QC) and quality assurance (QA) are different but related concepts. In the context of these Guidelines, *quality control* means devising and implementing safeguards to minimise the corruption of data. These safeguards must be installed at every step of the process from project definition to the decision on whether measured concentrations compare acceptably with the guidelines. *Quality assurance* means testing the effectiveness of these safeguards.

In any QA/QC program, chain of custody documentation is essential to ensure that errors can be traced. Chapter 4 of the Monitoring Guidelines discusses QA/QC in some depth for key points for chemical, physical and toxicant indicators; Section 7.4.3 below refers to that source.

## 7.2.5 Sample processing and analysis

Analysis here refers to the processing of sample units (e.g. field or laboratory measurement of analytes in a water sample, counting and identifying invertebrates in a benthic sample) rather than the statistical analysis of the resulting data. As with sampling, standard or rigorously tested protocols should be used; many protocols also detail methods of analysis. Because of their reliance on often complex, rigorous laboratory procedures, more specific guidance on analytical procedures is provided for physical and chemical stressors, toxicants and sediments.<sup>b</sup>

<sup>b</sup> See also Sections 7.4 and 7.4.3

Again, QA/QC procedures are often described in protocols, and QA/QC is also discussed in Chapter 5 of the Monitoring Guidelines. The monitoring team should document at least the analytical steps and the date and location of the analyses, the identities of the analysts, the methods used and the type and status of any equipment used for the analysis.

Similar care in QA/QC should also be used during data entry and data management. Most modern database software packages provide value-checking routines, and clear procedures should be established to manage, track, back-up and archive data files. Clear documentation of the features of the data (e.g. the units that the data are entered in, codes used for missing or ‘below detection limit’ data) need to be kept with the data files.

### 7.2.6 Data analysis, evaluation and reporting

*a See Chapter 6 of the Monitoring Guidelines*

The first step in evaluating the data will be the formal statistical analysis. For some indicators, the data may need to be adjusted to account for modifying factors (e.g. effect of pH on chemical speciation). The process of analysing the data is also iterative, with the first step being to examine the distributions of the variables and to check for outliers,<sup>a</sup> to see whether the data meet the assumptions of the intended analysis. Sometimes transformation of the data can solve distributional problems. Most statistical procedures have a second diagnostic stage after the procedure has been applied (e.g. examination of residuals after fitting a regression or general linear model). If these diagnostics show that the assumptions of the procedure have been violated, alternative statistical models may need to be developed. Chapter 6 of the Monitoring Guidelines discuss these issues, while the involvement of professional statisticians is invaluable in ensuring the rigour of these analyses.

Once the statistical analyses have been completed, the results need to be interpreted in the context of the key interacting environmental processes and the environmental assessment objectives of the program. Reporting of the results needs to be clear, concise, unambiguous and timely to allow management to act on the results. It is essential to disseminate the results to stakeholders in a form that is readily understandable, and some general recommendations are given in Chapter 7 of the Monitoring Guidelines. Reporting will often include recommendations on modifications to the program if it is a continuing program, thereby closing the feedback loop in figure 7.1.1 (this chapter).



## 7.3 Specific issues for biological indicators

This section addresses issues specific to biological indicators that need to be borne in mind when designing monitoring and assessment programs. Section 7.3.1 outlines the issues for *univariate* indicators: these consist of a *single* response variable such as the density or biomass of phytoplankton, measures of community metabolism, or chemical/biochemical markers in aquatic organisms. Multivariate indicators<sup>a</sup> refers to measures of community composition or structure where the response variable is usually based on some measure of community similarity which, in turn, is computed from the abundance (structure) or presence or absence (composition) of many taxa within the ecosystem. Examples include measures of the community structure of diatoms, macroalgae and invertebrates. AUSRIVAS, the rapid biological assessment technique for Australian rivers, is also based on multivariate community composition data, but is a special case of a design class where inferences are based exclusively on spatial controls alone. Section 7.3.3 discusses how the outputs of AUSRIVAS relate to the issues raised in Section 7.2.

a See Section 7.3.2

### 7.3.1 Issues for univariate indicators

Most of the key issues are raised in Section 7.2 and in Chapters 3 and 4 of the Monitoring Guidelines. Univariate indicators are easily analysed using conventional and novel statistical procedures, provided the key assumptions are met.<sup>b</sup> However, two issues are worth emphasising.

b Ch 6 of the Monitoring Guidelines

First, many of the classical techniques of statistical analysis assume independence of sample units through space and time. Biological indicators may violate these assumptions temporally because of the longevity of indicators or spatially because of dispersal or behaviour of indicators. If these phenomena are likely within the monitoring program, then professional statistical advice should be sought to either adjust the sampling regime or select statistical modelling tools that can accommodate these spatial and/or temporal autocorrelations (Legendre & Legendre 1998).<sup>c</sup>

c See also the Monitoring Guidelines Sections 6.5.2, 6.6.1

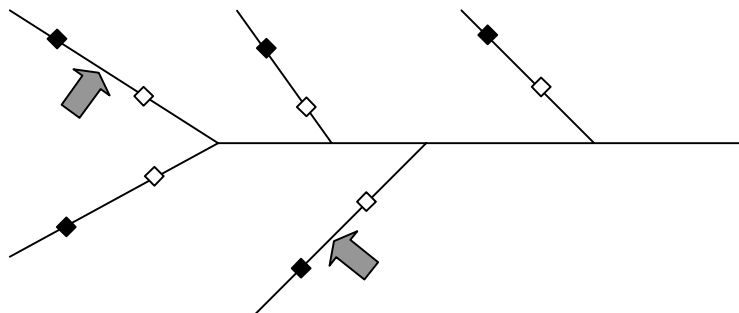
Second, data which consist of counts of organisms sometimes result in a large number of zero values (i.e. when there are no organisms in the sampling unit). The frequency distributions of such data are typified by a 'spike' at zero, then a mode at some larger, non-zero value. Assuming that the sampling unit or device is appropriate for the size and behaviour of the organism (most of the protocols recommended in Chapter 8 give advice on sizes of sampling units; for a more thorough discussion see, for example, Andrew & Mapstone 1987), such data are usually problematic for most statistical techniques. Some recent advances have been made in this area; as this is still an active developing area of applied statistics, professional advice should be sought when choosing and using these techniques.

### 7.3.2 Issues for multivariate indicators

Multivariate data for biological indicators in these Guidelines typically consist of either the presence or absence of taxa or their abundances across the sample units. These data can then be summarised as similarities (or dissimilarities) between each pair of sample units. The Bray–Curtis measure, among a few others, has been demonstrated to be the best choice for such biological data (Faith et al. 1987), and

sometimes transformation of the data is desirable before the similarity measure is computed, as described in the protocols in Appendix 3 (Vol 2).

The result of these computations is a triangular matrix of similarity values. These are not easily analysed in conventional statistical procedures such as analysis of variance or regression. However, one situation which is amenable to the use of similarity measures in more conventional procedures is where the ‘control’ and ‘impact’ sites can be paired, on rivers for example. In figure 7.3.1, there is a pair of sites on each tributary river which are comparable in terms of habitat and separated by similar distances on each river. The similarities between the upstream and downstream sites on each river could be computed for a number of times before the disturbance; if the similarities computed between the paired sites decreased after the start of the impact relative to the similarities between paired sites on the control rivers, then an impact is likely to have occurred. Examples using this design are Faith et al. (1995) and Davies and Nelson (1994).



**Figure 7.3.1** Schematic diagram of a river system with paired upstream (black diamonds) and downstream (white diamonds) sites on each tributary. Grey arrows indicate locations of disturbances.

Pairing of sites in this way is not always possible, however. Permutation tests that are analogues of some of the simpler conventional techniques (Smith 1998, the ANOSIM of Clarke & Green 1988, Clarke & Warwick 1994) have been used (e.g. Smith 1994). Significance testing of multivariate data based on similarity measures using permutation tests is rapidly developing (see Legendre & Legendre 1998 for an overview; Legendre & Anderson 1999 for an attempt at analysing multifactorial data). It is likely that methods for the analysis of similarity data in complex designs will become available in the near future.

The more conventional methods of analysing dissimilarity data have focused on displays of the data via such techniques as multidimensional scaling, principal components analysis and correspondence analysis (review: Legendre & Legendre 1998; brief description of principal components analysis and multidimensional scaling in Sections 6.5.4 and 6.6.3 of the Monitoring Guidelines). Inferences have been made purely on the basis of striking patterns in such displays, and Green (1979) argued that obvious patterns in such graphs were likely to correspond to large impacts. Such a procedure obviously lacks the sensitivity required for some assessment objectives. However, these displays remain an important tool for interpretation and communication after a formal hypothesis test via a randomisation or permutation procedure (Clarke & Warwick 1994).

### 7.3.3 Use of AUSRIVAS

#### 7.3.3.1 Outline of AUSRIVAS

AUSRIVAS is a rapid biological assessment procedure developed for rivers and streams (Schofield & Davies 1996) similar to the British RIVPACS system (Wright et al. 1993). It currently uses macroinvertebrate data, but the use of other taxa (e.g. diatoms and fish) is being researched. Its applicability to wetlands and to New Zealand rivers is also being investigated.

AUSRIVAS is a specialised example of a monitoring design that relies on spatial information alone to infer whether a disturbance has caused an impact. In general terms the problem can be stated thus: to judge whether a particular site has been disturbed by some activity or event, other, apparently undisturbed, sites that are similar in their environmental attributes must be found to act as a standard or *control* for comparison. The site suffering the supposed disturbance is designated the *test site*, while the sites acting as controls are called *reference sites*.

In AUSRIVAS a large number of reference sites with as high an environmental quality as possible have been identified across a wide variety of river types and ecosystems, sampled for their macroinvertebrates and had their habitats characterised by a standard set of physical and chemical variables that are largely unrelated to likely pollutants. This set of reference sites has then been classified according to their biota to produce groups of sites containing similar fauna. A numerical analysis has then been used to identify the environmental attributes which describe each group of reference sites. Now, any test site requiring assessment has its environmental attributes compared with those of the reference sites to determine which group or groups of reference sites it resembles most closely. The fauna of these corresponding reference sites is then compared with the test site: if the test site supports fewer taxa than are predicted by the reference sites, it is judged to be disturbed.

#### 7.3.3.2 Sampling protocol and issues about effect size and sensitivity

The AUSRIVAS protocols modify much of the advice given in Sections 7.2.3 to 7.2.6 above. Site selection, the methods of stratifying habitats within sampling sites, the timing of sampling and the analytical methods and outputs are all specified in protocols for each state and territory in Australia. The summary indices and recommendations for reporting procedures are also standardised, and the sampling, sorting and identification steps are subject to QA/QC programs. The software for analysing the data is maintained and developed at a central location accessible via the AUSRIVAS home page.

Decisions about effect size are implicit in the procedure. The degree of impact upon a site is judged by the values of summary indices relative to a stipulated percentile of the reference sites that act as spatial controls. If a site scores a value on these indices that is smaller than 90% of the values recorded for reference sites, the fauna is deemed to be lacking some of the families of invertebrates that could be expected at that site. Although designation of such a percentile threshold expresses an effect size in terms of how deviant a site is from reference conditions, it is analogous to but not exactly equivalent of the process of setting Type I and II error rates for conventional statistical procedures.

Nevertheless, the issue of how far a potentially-disturbed site can deviate from reference or control conditions before an impact is deemed to have occurred must still be resolved with stakeholders. Strategies similar to those described above for procedures based on statistical methods can be employed: i.e. use of existing information, examining the response of the indicator variable to known impacts of different degrees of severity, and use of pilot data in simulation modelling. Note that indicators used in rapid, broad-scale methods are often quite coarse (e.g. use of family-level rather than species-level identifications). Thus the threshold value at which the decision is made that an impact has occurred should take account of the harmfulness of the potential impact, its reversibility and the time-lags between an event and the implementation of actions to prevent harm. The threshold value may need to be set at a more conservative value than that deemed acceptable by stakeholders so that management has time to react and prevent irreversible harm.

### 7.3.3.3 Application and cautions

*a See box  
3.2.1, Section  
3.2.2.1/3*

AUSRIVAS has been promoted in these Guidelines as ideally suited for the rapid, cost-effective, first-pass determination of the *extent* of a problem or potential problem, e.g. as applied to broad-scale land-use issues. Earlier, a note of caution was provided for use of the method in applications other than these,<sup>a</sup> particularly for detecting impacts of a minor nature and for site-specific assessments where the method requires additional testing and the addition of more data. A perspective on some of the limitations of the approach is provided below, together with comments on ongoing data collection and proposed research and development aimed at improving the sensitivity and broadening the application of this procedure.

An important aspect of AUSRIVAS is the availability and selection of suitable reference sites. In some regions of Australia it is easy to find reference sites on rivers and streams draining relatively intact catchments. Unfortunately, large regions of Australia have been subject to broad-scale impacts and there are no 'near-pristine' sites from which to select biogeographically relevant reference sites (e.g. wheat belt of Western Australia, lowland reaches of the Murray-Darling Basin). Thus, in AUSRIVAS, the least impacted sites of such regions have been targeted to provide reference sites for setting targets for rehabilitation of the more degraded sites; however, this does not solve the problem of assessing the degree of degradation of the reference sites themselves. Without pre-impact data, this task is outside the ambit of routine prescriptive procedures and would require a variety of situation-specific case studies to arrive at some assessment. The issue is being addressed as part of the current Australia Wide Assessment of River Health (AWARH), which aims to report on the ecological condition of around 4000 Australia river sites by the year 2000 using AUSRIVAS.

A related but more tractable problem results when a test site has no close environmental equivalents in the reference database. Therefore, an important initial step in evaluating a test site is a statistical comparison between its environmental attributes and those of all the reference sites: if it has no sites with similar attributes in the reference set, no further assessment can be made, i.e. there are insufficient sites in the database that can be regarded as a 'control' (Furse et al. 1987). The current AUSRIVAS software (available from the AUSRIVAS homepage) contains a testing routine to assess whether test sites fall within the 'domain' of the existing reference site set on which the bioassessment models are based. It then must be

decided whether the test site warrants the added expense of adding sufficient comparable reference sites to the database to enable an assessment to be made.

A potential drawback is the relatively large number of reference sites that must be sampled to build reliable models for predicting the presence or absence of the target organisms. This is particularly relevant to site-specific assessments, where adequate characterisation of the local reference condition is critical. The development of AUSRIVAS is predicated on the collection of a large amount of reference site data nationally, and it is anticipated that the geographic spread, as well as the spatial density of sites, will gradually increase to improve the applicability of the predictive models.

Another important aspect for some Australian streams is the natural inconstancy of animal community composition amongst years. Thus, 'high' temporal variability of macroinvertebrate communities over large parts of Australia, particularly semi-arid and northern regions prone to drought and/or cyclonic disturbance (Humphrey et al. 2000), may reduce to some (as yet unknown) extent the sensitivity of 'static' models derived for these locations. To this end, it is recommended that test site assessment using the AUSRIVAS protocol should be done in parallel with reference ('control') site assessment to assess the degree of natural temporal change in macroinvertebrate community composition and compare it with the summary index value for the test site.

In some regions of Australia it is clear that some reference sites are naturally depauperate; that is, the number of macroinvertebrate taxa is low. For procedures such as AUSRIVAS, where the final reporting indices are based on the ratio of the number of taxa observed to the number of taxa expected, this poses potential problems for the sensitivity and robustness of the final assessment, even at species level.

Finally, AUSRIVAS and related procedures (Reynoldson et al. 1995) are rapid assessment tools and will only detect impacts that are severe enough to eliminate taxonomic groups of organisms. The formal hypothesis testing associated with conventional statistical methods has no clear analogue here. This procedure uses a suite of reference sites to predict the expected composition of families of invertebrates at a test site; if the test site has fewer families than expected based on the distribution of reference site values, it is deemed to be disturbed. Nevertheless, several basic considerations of survey design (sample and site replication, etc.) still apply to assessments or surveys conducted with AUSRIVAS. These considerations become particularly important at small spatial scales (i.e. a specific activity, development, point-source disturbance, within a catchment) where stronger inference and greater sensitivity to impact may be required. If AUSRIVAS is to be adopted in these situations, it must be conducted in a design framework that has adequate sample and site replication to enable the study objectives to be met. If necessary, aspects of the rapid biological assessment protocol may need to be adapted or modified so that the data gathered are amenable to both AUSRIVAS and quantitative assessment.<sup>a</sup>

<sup>a</sup>  
*Complementary roles for quantitative and rapid assessment in monitoring programs are recommended in Section 7.2.1.1 above*



## 7.4 Specific issues for physical and chemical indicators (including toxicants) of water and sediment

This section outlines issues specific to physical and chemical indicators (including toxicants) of water and sediment that need to be borne in mind when designing monitoring and assessment programs. The issues in Sections 7.4.1 and 7.4.3 are comprehensively discussed in Chapters 3–6 of the Monitoring Guidelines — see table 7.4.1 for a checklist of these issues and appropriate cross-reference to the Monitoring Guidelines.

**Table 7.4.1** Checklist for sampling and analysis of physical and chemical indicators with cross-reference to details provided in the Monitoring Guidelines.

Issue	Chapter or section from the Monitoring Guidelines
Representative sampling	Chapter 3, Chapter 4
spatial boundaries	3.3.1
scale	3.3.2
duration	3.3.3
patterns of sampling	3.4.1
selection of sites	3.4.2
frequency of sampling	3.4.3
numbers of samples	3.4.4, A5.1.10
Surface water sampling	4.3.1, 4.3.2
hydrology, flow variations, runoff	3.4.3, 3.4.3.2, 4.3.1
stratification	3.4.1.2, 3.4.2.1
human effects on contaminant loads and timing	3.4.3.2
automatic samplers	3.4.3.2, 4.3.2
time of day	3.4.3, 3.4.3.2
Sediment sampling and sediment sample handling	4.3.1, 4.3.5, 3.4.2.1, 5.5.8
potential for contamination	4.3.5, 4.3.1
suspended sediments	4.3.5
Sample storage and handling	4.5, 4.6, 4.3
Chemical speciation	5.5.8.2, Tables 4.5 & 5.2
Bioavailable concentration vs total concentration	3.5
Quality assurance/Quality control in the field	4.6 and subsections
chain of custody	4.6.1
training of staff	4.7.2
quality assurance samples: blanks	4.6.3.1
quality assurance samples: replicates	4.6.3.2
quality assurance samples: spiked samples	4.6.3.3
pilot trial	3.4, 3.4.2, 3.4.4
equipment	4.6, 4.6.2, 4.6.3.1
sample transport	4.6.1, 4.6.2, 4.6.3.1
site access	3.4.2, 4.2, 4.7.1, 4.7.3
occupational health and safety	4.7 and subsections
analytes	5.3
cleaning and calibration	4.3.1, 4.3.2.1, 4.3.6, 4.6.1, 4.6.2
protocols	4.3.7, 4.6.2
Quality assurance/Quality control in the laboratory	5.5 and subsections
chain of custody	5.4.1.2
occupational health and safety	5.6 and subsections
training of staff	5.6.3
quality assurance program	5.5.5 and subsections
quality assurance samples	subsections of 5.5.5
matrix compatibility	5.5.5.1
accurate recording of data	5.4, 6.2

In addition to the cross-references provided to sampling and analysis of sediments in table 7.4.1, a protocol describing key aspects of collection and laboratory analysis of sediment samples is provided in Appendix 8 of Volume 2, while advice on comparing sediment ‘test’ data with default guideline values is provided in Section 7.4.4.4 below.

*a See also the Monitoring Guidelines Section 6.4.3*

These Guidelines emphasise the use of guideline trigger values for assessing the environmental significance of physical and chemical indicators. The statistical procedure for comparing test data and a trigger value is described in Section 7.4.4.<sup>a</sup> The generic considerations for sampling design given in Section 7.2 also apply to physical and chemical indicators.

### 7.4.1 Hydrology and representative sampling

*b See Monitoring Guidelines Chapters 3 & 4*

Sampling of waters and sediments must be representative.<sup>b</sup> The challenge is to sample in enough detail to outline a picture of the natural variations in time and space and to reliably detect deviations from this natural ‘background’ variation.

Natural variations in surface waters and groundwaters, whether flowing or standing, can affect the values of physical and chemical indicators. For example, all water bodies can form vertical or horizontal layers of differing temperature or salinity that may or may not need to be sampled separately, according to the sampling plan. Currents and the lateral and vertical movements of different water masses also need to be considered during the planning of field sampling, analyses and study design. Natural periodicity, and the timing of industrial discharges into water bodies, and the considerable effects of runoff in inland waters can make large differences to the loads and concentrations of physical and chemical indicators, and must also be planned for.

*c See Vol 2, App 8 and Monitoring Guidelines Sections 4.3, 4.3.5*

In sediments,<sup>c</sup> the sampling plan and study design must consider the effects of natural layering, mixing, and variations in particle size and porosity on the indicator being sampled. The likelihood of disturbance and cross-contamination during sampling must not be forgotten. Suspended sediments need to be collected in a representative manner (Batley 1989), as do sediment pore waters.

For all samples, precautions must be planned and taken to prevent the values of the indicators changing during storage and transport.

### 7.4.2 Chemical speciation in water samples

The issue of the chemical form of physical and chemical indicators (that is, the compound(s) of the indicator present in the sample) are relevant regardless of the use envisaged for the water. Speciation (the form of the chemical) assumes critical importance where the environmental value concerns ecosystem protection or human health. The form of the indicators needs to be determined and those chemical species that are likely to affect the environmental value must be identified. In the past, total (i.e. unfiltered) concentrations were measured and compared with guideline values on the understanding that this approach probably overestimates the amount of deleterious form(s) of the indicator. This approach to protection may be overconservative. A refinement is to measure and compare *total filtered* concentrations. This, too, is a conservative approach (though less so) because the diversity of chemical forms of a physical and chemical indicator in the solution may have different effects on an environmental value.

There are at least two ways to resolve the speciation problem.

- Determine the indicator using an analytical method that is specific to the chemical species. While this is an improvement on using total filtered concentrations, it requires the species or range of species detected by the method to be arbitrarily defined as a surrogate for the species affecting (usually detrimentally) the environmental value. An example of this approach is the use of anodic stripping voltametry in the determination of copper. The fraction determined under operationally defined conditions is identified as labile forms which, in turn, are believed to be the forms in which copper is most bioavailable.
- Use *thermodynamic speciation modelling*. One requirement of this mathematical tool is that all aqueous chemical species that may be important to the chemical form of indicators be measured. This usually increases the analytical requirements because of the inclusion of chemical species that would not otherwise be determined. The technique requires that the system measured is in equilibrium, and that the equilibrium is the same as that existing at the time of sampling. This has implications for preservation, transport and storage of samples. The specification and interpretation of thermodynamic speciation models is complex and requires considerable facility in the use of computers, and in the interpretation of chemical data. A more detailed discussion of speciation modelling is beyond the scope of these Guidelines.

### 7.4.3 Quality Assurance and Quality Control (QA/QC)

Quality control and quality assurance were defined generically in Section 7.2.4. A specific formal statement of quality control for physical and chemical indicators is this:

The overall objective of quality control in the measurement of physical and chemical variables is the determination of the *exact* indicator concentration that existed at a specifically defined location at the time the sample was taken. In most cases this requirement extends to the chemical speciation of the indicator.

Neglect of QA/QC is probably the most important reason for the unreliability of most historical chemical data.

Protocols for field and laboratory aspects of sampling must be followed carefully, as discussed in the Monitoring Guidelines.<sup>a</sup> QA/QC begins with the choice and training of competent field and laboratory staff; it includes the choice and maintenance of field and laboratory equipment and vehicles. It extends to the checking of analytical methods and analytical performance, the tracking of each sample throughout sampling and analysis, and the accurate recording of data in the final database.

<sup>a</sup> See  
Monitoring  
Guidelines  
Chapters 4, 5  
and 6

### 7.4.4 Comparing test data with guideline trigger values

#### 7.4.4.1 Physical and chemical stressors

This section provides a summary of the approach recommended for comparing results from a test site with a guideline trigger value. Details of the method are contained in Appendix 7 of Volume 2; Section 6.4.3 of the Monitoring Guidelines touches on it also. There are a number of common statistical methods that are

potentially applicable for this purpose, although experience suggests that the assumptions underpinning many ‘conventional’<sup>16</sup> statistical tests are often violated by water quality data (and sometimes quite seriously so). Section 6.3.4 of the Monitoring Guidelines recommends transformations to correct specific problems, although the action required will depend very much on the characteristics of the data at hand. This lack of consistency in the way site-specific data may be processed and interpreted is an impediment to the development of a simple, straightforward trigger rule.

Compounding this difficulty is the usual requirement to specify the magnitude of change in a particular statistical parameter (e.g. mean, variance, percentile) that is deemed to be ‘significant’ — either ecologically or statistically or both. The quantification of a *minimum* effect size that can be claimed to be ecologically important is difficult. With respect to the trigger rule outlined in this section, this issue of ecological importance is discussed further below and more generally in Section 7.2.3.3. The important observation to note at this stage, however, is that exceedances of the trigger values are an ‘early warning’ mechanism to alert managers of a potential problem. ***They are not intended to be an instrument to assess ‘compliance’ and should not be used in this capacity.***

During the development of a suitable trigger mechanism, considerable attention was given to the following design requirements:

- explicit recognition of the inherent (and usually large) variability of natural systems;
- robustness under a wide range of operating conditions and environments;
- no, or only weak, distributional assumptions about the population of values from which the test and reference data are obtained;
- known statistical properties, consistent with and supporting the monitoring objectives of this document;
- ease of implementation and interpretation;
- suitability for visual display and analysis;
- intuitive appeal.

The recommended trigger-based approach for physico-chemical stressors may be stated as follows.

A trigger for further investigation will be deemed to have occurred when the median concentration of  $n$  independent samples taken at a test site exceeds the eightieth percentile of the same indicator at a suitably chosen reference site. Where suitable reference site data do not exist, the comparison should be with the relevant guideline value published in this document.

This rule satisfies the first dot point above since it is statistically-based and acknowledges natural background variation by comparison to a reference site. Its robustness derives from the fact that it accommodates site-specific anomalies and uses a robust statistical measure as the basis for triggering. No assumptions are

---

<sup>16</sup> In this context, the term conventional is used to denote statistical procedures based on the *general linear statistical model* having normally distributed errors.

required to be made about the distributional properties of the data obtained from either the test or reference sites. The computational requirements of the approach are minimal and can be performed without the need for statistical tables, formulae, or computer software. As demonstrated later in this section, the temporal sequence of trigger events is readily captured in a simple plot.

It should be understood that the trigger protocol is responsive to shifts in the *location* (i.e. ‘average’) of the distribution of values at the test site. While differences in shape of the reference and test distribution may be important in some instances, this is a secondary consideration that is not specifically addressed by this protocol. It is also important to note that the role of the 80<sup>th</sup> percentile at the reference site is to simply quantify the notion of a ‘measurable perturbation’ at the test site. The protocol is not a statistical test of the equivalence of the 50<sup>th</sup> and 80<sup>th</sup> percentiles *per se*. The advantages of using a percentile of the reference distribution are 1) it avoids the need to specify an absolute quantity and 2) because the reference site is being monitored over time, the trigger criterion is being constantly updated to reflect temporal trends and the effects of extraneous factors (e.g. climate variability, seasonality).

Implementation of the trigger criterion is both flexible and adaptive. For example, the user can identify a level of routine sampling (through the specification of the sample size *n*) that provides an acceptable balance between cost of sampling and analysis and the risk of false triggering. The method also encourages the establishment and maintenance of long-term reference monitoring as an alternative to comparisons with the default guideline values provided in Section 3.3 that do not account for site-specific anomalies.

The remainder of this section addresses sampling issues, data requirements, computational procedures and statistical properties associated with the proposed method. The mathematical detail associated with computation of Type I and Type II errors may be found in the Annex of Appendix 7 of Volume 2. Worked examples of the computations and performance aspects of the trigger rule are provided in Appendix 7 (Volume 2).

##### 1 Data requirements at the reference sites

Prior to implementing the trigger rule, the user will need to have addressed some data collection issues.

- *Reference site selection*: selection of (a) suitable reference site(s) has been addressed in Section 3.1.4.
- *Minimum data requirements at the reference site*: a minimum of two years of contiguous monthly data at the reference site is required before a valid trigger value can be established. Until this minimum data requirement has been established, comparison of the test site median should be made with reference to the default guideline values identified in Section 3.3 of this document.

##### 2 Computation of the 80<sup>th</sup> percentile at the reference site

The computation of the 80<sup>th</sup> percentile at the reference site is always based on the *most recent* 24 monthly observations. The procedure is as follows:

- (i) arrange the 24 data values in *ascending* (i.e. lowest to highest) order,
- (ii) take the simple average (mean) of the 19<sup>th</sup> and 20<sup>th</sup> observations in this ordered set.

### 3 Updating the reference site data and 80<sup>th</sup> percentile

Each month, a new reading at the reference (and test) site is obtained. The reference site observation is appended to the end of the original (i.e. unsorted) time sequence. Steps (i) and (ii) from 2 above are applied to the most recent 24 data values. Note, even though only the most recent two years of data is used in the computations, no data should be discarded.

Maintenance of the complete data record will allow longer-term statistics to be computed. For example, after five years of monthly monitoring, *all* sixty observations could be used to compute the overall 80<sup>th</sup> percentile. This could serve as a useful benchmark against which the ‘rolling’ monthly percentiles could be compared for evidence of trends.

### 4 Data requirements at the test site

A feature of the method is the flexibility it provides the user for the allocation of resources to the sampling effort. As previously mentioned, there is no fixed requirement to monitor at a reference location (i.e. the default guideline values can be applied). Similarly, the choice of sample size at the test site is arbitrary, although there are implications for the rate of false triggering. For example, a minimum resource allocation would set  $n=1$  for the number of samples to be collected each month from the test site. It is clear that the chance of a *single* observation from the test site exceeding the 80<sup>th</sup> percentile of a reference distribution which is *identical* to the test distribution is precisely 20%. Thus the Type I error in this case is 20%. This figure can be reduced by increasing  $n$ . For example, when  $n=5$  the Type I error rate is approximately 0.05. The concomitant advantage of larger sample sizes is the reduction in Type II error (the probability of a false no-trigger). So-called ‘power curves’ are provided in Appendix 7 (Volume 2) to assist in understanding the consequences upon error rates of a particular sampling strategy at the test location.

### 5 Computation of the median at the test site.

The median is defined to be the ‘middle’ value in a set of data such that half of the observations have values numerically greater than the median and half have values numerically less than the median. For small data sets, the sample median is obtained as either the single middle value after sorting in ascending order when  $n$  is *odd*, or the average of the two middle observations when  $n$  is *even*.

### 6 Ecological importance

The proposed trigger rule does not purport to define or represent an ecologically important change. As previously explained, the trigger approach is an early warning mechanism to alert the resource manager of a *potential* or *emerging* change that should be followed up. Whether or not the actual change in condition at the test site has biological and/or ecological ramifications can only be ascertained by a much more comprehensive investigation and analysis. To make this distinction clear, the concept of a **measurable perturbation** is introduced. Our *de facto* definition of a measurable perturbation is that it is the magnitude of the shift between the 50<sup>th</sup> and 80<sup>th</sup> percentiles *at a reference site*. While the definition is arbitrary, it does have broad acceptance and intuitive appeal among experts. It should also be noted that the *statistical significance* associated with a change in condition equal to or greater than a measurable perturbation would require a separate analysis.

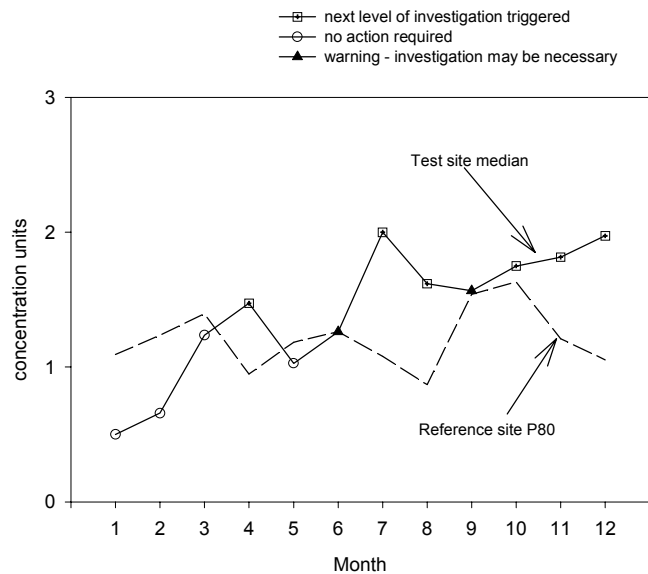
### 7 Performance characteristics

It is important that the statistical performance characteristics of any test or decision-making rule are documented and understood to avoid unduly conservative or liberal triggering.

The foregoing discussion makes no assumptions regarding the shape of the reference and test distributions. Without this knowledge, a formal calculation of Type I and Type II errors is not possible. However, as a general principle, increasing the frequency of collection of independent samples will reduce the magnitude of both errors. A more complete discussion of the performance characteristics of the recommended approach is provided in Appendix 7 of Volume 2.

### 8 On-going monitoring — the control chart

The foregoing has been provided to assist with the month-by-month comparisons. It is suggested that these monthly results be plotted in a manner indicated in figure 7.4.1 below. This provides a visual inspection of all results and helps identify trends, anomalies, periodicities and other phenomena. The methods in Chapter 6 of the Monitoring Guidelines can be used to model trends and other data behaviour if required.



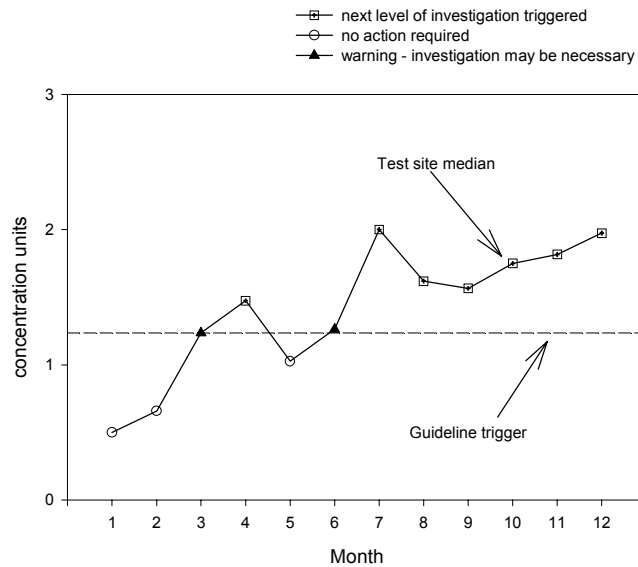
**Figure 7.4.1** Control chart showing physical and chemical data (Y axis) for test and reference sites plotted against time, and recommended actions

### 9 Comparing test data against single guideline (default values)

In the absence of suitable reference site data (as defined in step 1 above), the median of the test site data is to be compared with the default guideline value identified in Section 3.3.2.5 of this document. This guideline value has been computed as the 80<sup>th</sup> percentile of the amalgamation of a number of historical data sets across broad geographical regions. Unlike the comparison with a locally-derived 80<sup>th</sup> percentile, the guideline value is static and will not reflect any local spatial and/or temporal anomalies. Reference site monitoring is strongly advocated

if these effects are considered to represent a significant source of departure from the guideline value.

Figure 7.4.2 below illustrates the difference in control charting procedures when the guideline value is used in place of a trigger obtained using the 80<sup>th</sup> percentile from reference site monitoring.



**Figure 7.4.2** Control chart showing physical and chemical data (Y axis) for test site plotted against default trigger value, time and recommended actions

#### 7.4.4.2 Toxicants

This section describes the general needs for comparing toxicant test data with guideline trigger values. Conceptually, toxicants and ‘physical and chemical stressors’ are subcategories of the same class of potentially hazardous indicators, being properties or (usually) constituents of the aquatic environment. However, the treatment of these groups for guideline purposes is different. Specifically, toxicants are usually compared with a single default trigger value, less commonly with a background or reference distribution. The default values are prepared by analysis of a comprehensive set of available ecotoxicological data. Physical and chemical stressors at a test site are usually compared with those at a reference site. The latter reference-comparison approach, however, has its parallels in measurement programs for toxicants, as described in 1 below.

##### 1 Background data that may supplant guideline default trigger values

Some surface waters will contain concentrations of toxicants that may naturally exceed the default guideline trigger values tabulated in Section 3.4. Where this is the case and as recommended in Section 3.4.3.2, new trigger values should be based on background (or baseline) data. (Note that ‘background’ in this case, refers to *natural* toxicant concentrations that are unrelated to human disturbance.) As a matter of course, gathering of background data is always recommended, at least in

the initial stages of a water quality management program, to establish whether or not concentrations of toxicants are naturally high.

*a See step 1,  
Section 7.4.4.1*

Toxicant concentrations may vary seasonally. Because of this and the need to be confident about the best estimate of background concentrations, it is recommended that background data be gathered on a monthly basis for at least two years. In all respects, data requirements and collection are the same as for physical and chemical stressors, as described above.<sup>a</sup> Until this minimum data requirement has been established, comparison of the test site median should be made with reference to the default guidelines identified in Section 3.4.3 of this document.

For those months, seasons or flow periods that constitute logical time intervals or events to consider and derive background data, the 80th percentile of background data (from a minimum of 10 observations) should be compared with the default guideline value. This 80<sup>th</sup> percentile value is used as the new trigger value for this period if it exceeds the default guideline value provided in Section 3.4.3 of this document. Test data are compared with the new trigger values using the same principles as outlined in steps 2–8 above for physical and chemical stressors.

Where background toxicant values fall consistently below default trigger values, sampling intensity at these sites could be reduced after a suitable period (e.g. two years).

#### *2 Comparing test data with default guideline values*

*b Section  
7.4.4.1*

In practical terms, the method for comparing toxicant test data with default guideline values should be similar to the approach recommended in step 9 above for physical and chemical stressors.<sup>b</sup> However, it is recommended that a more conservative approach should apply to the comparison of toxicant test data with default guideline values. Specifically, it is recommended that action is triggered if the 95th percentile of the test distribution exceeds the default value (or stated differently, no action is triggered if 95% of the values fall below the guideline value). The more stringent approach is recommended here because, unlike physical and chemical stressors, toxicant default values are based upon actual biological effects data and so by implication, exceedance of the value indicates the potential for ecological harm. Note that because the proportion of values required to be less than the default trigger value is very high (95%), a single observation greater than the trigger value would be legitimate grounds for action in most cases, even early in a sampling program.

#### **7.4.4.3 Physical and chemical (including toxicant) data gathered from surface waters 'upstream' of the test site**

In many situations, particularly where additional human use activities are present 'upstream' of the test site of interest, the regular collection of data from upstream of the test site will be necessary. These data will be compared with the test data of interest to assist in determining the source and cause of any possible elevated toxicant concentrations found at the test site. Where there are multiple sources of toxicants along a water-course, catchment managers will need to establish appropriate data analysis and assessment procedures to apply.

#### 7.4.4.4 Sediments

*a See Section 3.5*

*b Section 3.1.4*

The application of the decision tree<sup>a</sup> reverts to reference or background site concentrations if these exceed the trigger values. The selection of an acceptable reference site for water quality studies has been discussed earlier.<sup>b</sup> Basically the same considerations apply to sediments, with the additional option, that a reference or background condition can also be established from measurements at depths in sediment cores below observed concentration excursions.

While temporal variability is used to characterise water quality parameters at a reference site, this is clearly inappropriate for sediments where the accumulation rates are typically below 1 cm/y. It is more appropriate to use spatial variability, either based on depth profiles at a test site or an appropriate number of surface sediment samples, to characterise a site. Sites will typically contain a range of grain sizes, and determining median concentrations and 80<sup>th</sup> or 95<sup>th</sup> percentile values may distort any comparison. It is important that in comparing test and reference sites, samples with a similar grain-size distribution are used. Normalising to a fine grain size (e.g. <63 µm) is inappropriate, as the normalised value will have less of an impact on biota when diluted with coarser sediments that usually contain lower contaminant concentrations.

The spatial scale over which the reference and test site measurements are taken is a matter for decision by stakeholders, based on sound scientific judgement. The heterogeneity of sediment samples with respect to contaminants largely mirrors the differences in grain size. Defining the size of the test site will be a regulatory responsibility, in terms of the spatial extent of contaminated sediment that is acceptable in the region of interest. As a guide, the spatial extent of a test site may be a geographical feature, for example, a delta or an embayment within a harbour. Alternatively, a test site may comprise a recognised ecological habitat, for instance a riffle zone in a stream or a defined area of fine sediment in a lake. In a large water body the test site might be larger than in a narrow river or creek, where biota might have difficulty in avoiding the contamination. The area of any reference site should be comparable to that of the test site, and the grain size must be similar.

Because of the poor reliability of the sediment trigger values it is difficult to be prescriptive about how these can be compared with test values. The same applies to the comparison of reference site values with test sites, where comparisons of reference median or 80<sup>th</sup> percentile with the test site median may be equally appropriate in giving an estimate of the relative concentrations, which is really all that is required in the case of sediments. However, where sediment samples within a test site clearly exceed trigger values, or are reasonably inferred to be ecologically hazardous, these Guidelines recommend additional sampling to more precisely delineate contaminated zones within the site.