Brock B. Bernstein,* Robert W. Smith,* and Bruce E. Thompson

SAMPLING DESIGN AND REPLICATION FOR BENTHIC MONITORING

In locations where there is believed to be an ecological risk to sea animals, benthic infaunal monitoring is often mandated. In spite of long experience with such monitoring in many locations, there is still scientific controversy about what should be required (e.g., number of stations, parameters to be measured, sampling frequency, and replication) to actually detect changes caused by contamination.

The uncertainties arise from the often complex relationships between the effects of man and the natural variability which may obscure them. The task of detecting and quantifying impacts is simplified if monitoring programs are designed with specific questions in mind and structured around an appropriate statistical model. A statistical model is a mathematical description of the sources of variability and the ways they affect the measurements taken to document impact. The model in turn determines the placement, timing, and number of measurements. An inappropriate statistical model may prevent the detection of an impact that in fact exists.

Our purpose in this paper is to present the results of power tests and optimization analyses performed to evaluate several parameters commonly used in monitoring programs. Evaluation was done using three statistical models for the detection of impact. All analyses were performed using existing data from benthic infaunal studies in southern California. We found that, in general, it is infeasible to detect changes in any one species without very large numbers of replicates. Changes in measures of community structure (diversity, evenness, numbers of species, total abundance) can be detected with few replicates, but, as we will demonstrate, these parameters are ambiguous measures of community change. We recommend instead, tests utilizing a multivariate ecological distance index that incorporates information about the distribution and abundance of all or most species in the community. We will also demonstrate that it is not efficient to take more than two

*EcoAnalysis, Inc., Ojai, Calif.

replicate samples at any one point in space and time when using the multivariate distance index.

METHODS

Power Tests

The power of a statistical test is the probability that it will detect a difference that, in fact, exists. In slightly more technical terms, the power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis when a difference actually exists. Power tests are useful because they permit a quantitative assessment, before resources are actually committed, of whether or not a proposed monitoring design has a reasonable chance of detecting changes deemed to be important. These tests also permit the evaluation of alternative levels of sampling effort.

Power tests require estimates of the error term in the statistical test being evaluated, as well as of the size of the change or impact to be detected. The requirement for a priori estimates of predicted change forces program designers to think carefully about the type and magnitude of change it is important to detect. If these parameters are not set, the ability of the monitoring program to detect change is arbitrarily dependent on the intensity of sampling.

The variances required for the power tests herein were calculated with raw data from benthic monitoring surveys in southern California (Table 1). Data were restricted to the 50- to 100-m-depth zone, where impacts from municipal wastewater outfalls are greatest. Previous authors have recognized three zones along outfall gradients (Bascom et al. 1978; Thompson 1982). We have followed this zonation, but use the terms "least contaminated," "transition," and "most contaminated" to classify sites. It should be noted that the "most contaminated" zone includes sites in the Zone of Initial Dilution (ZID) as defined by the Environmental Protection Agency (EPA).

Power formulas are reviewed in Cohen (1977). Since the calculation of power for interaction effects in factorial designs can be complex, we have verified results with simulations that measure power directly. We emphasize that "replicate" has different meanings in each of the three statistical models described below. The term "instantaneous replicate" has been used to refer explicitly to samples taken at the same place and time.

Statistical Models Tested

We evaluated three generalized statistical models, described below,

Table 1. Summary of data sets used in the analyses. All samples were screened through a 1.0-mm sieve. The data for each outfall were colected during monitoring surveys by personnel from the agencies listed.

Data Set; Source	Sampler	Time	Space	Utility
SCCWRP replicate analysis (RA); Thompson (1982)	0.1-sq m Van Veen	1979-1981, single sampling time	58-61 m, Santa Monica Bay to Orange County	Measures of replicate variability
White Point (WP); Los Angeles County Sanitation Districts monitoring data	0.04-sq m Shipek; 0.1-sq m Van Veen	1972-1982, semi-annually	60-305 m, Palos Verdes Peninsula	Measure of replicate and time x location variability
Point Loma (SD); City of San Diego, Point Loma lab monitoring data	0.1-sq m Van Veen	1979-1983, quarterly	60-81 m, San Diego	Measure of replicate and time x location variability
Santa Ynez survey (SY); Dames and Moore, Inc., Los Angeles, Calif.	0.1-sq m Smith-McIntyre	1981, summer	50-400 m, Santa Barbara Channel, nearshore	Measure of replicate variability
SCCWRP Orange County (OC); CSDOC monitoring data	0.1-sq m Van Veen	1981-1983, four surveys	100-600 m, nearshore	Measure of time x location variability
Santa Monica Bay (Hyperion) (HY); Los Angeles City Bureau of Sanitation monitoring data	0.1-sq m Van Veen	1983, winter and summer	60-80 m, nearshore	Measure of time x location variability

designed to detect changes under different assumptions of natural background variability. In general, there are at least four types of variability which may need to be accounted for in monitoring designs. There is the variance among measurements taken at the same place and time (the instantaneous replicate variability), as well as that among sampling locations, and among times. An important additional source of variability, which is most often ignored in monitoring designs, is the

time-by-location variability (Bernstein and Zalinski 1983); this is a measure of the degree to which the difference between locations varies naturally over time.

T-test. This simple model tests for difference between two groups of measurements. The underlying null hypothesis is that there is no difference in the mean value of some parameter in the two groups of measurements, as measured against replicate variability. Replicates are the measurements within each group.

Two-way ANOVA. This fixed-effects, two-way analysis of variance was presented as an optimal impact study design by Green (1979). The schematic below shows hypothetical mean values in each cell that demonstrate an impact:

Condition	**Location**	
	Impact	Control
Before	12.0	12.0
After	6.0	12.0

Green (1979) shows that the criterion for documentation of a detectable impact is a statistically significant interaction effect in the analysis of variance between location (here, Impact versus Control) and condition (Before versus After). In the example above, this is shown as a reduction in the Impact-After cell mean that does not occur in any other cell mean; in other words, the relationship between Impact and Control changes from the Before to the After. The null hypothesis is, therefore, that there is no change in the relationship between Control and Impact locations from the Before to the After conditions, as measured against replicate variability. In this model, replicate refers to the observations in each of the four cells of the design. The two-way ANOVA is appropriate when the time-by-location variance is small or nonexistent, or when trends over time are not of interest.

Difference model. If the time-by-location variance is large, multiple surveys over longer periods of time must be conducted to accurately detect any change. Bernstein and Zalinski (1983) proposed a model to include the fluctuations over time that are important characteristics of most natural systems. If not included in the statistical model, these can confuse the interpretation of monitoring studies. This model is presented schematically below:

Condition		Location		
		Impact	Control	Δ
Before	t_1 ⋮ t_n			ΔB_1 ⋮ ΔB_n
After	t_1 ⋮ t_n			ΔA_{n+1} ⋮ ΔA_{2n}

In this model, results of sampling at several times are nested within each condition. The t_n is the number of times both locations are sampled during each condition, and the Δ represents the difference between Impact and Control locations evaluated at each time. When several instantaneous replicates are sampled at each location, the average value of the replicates is used.

The presence of an impact can be assessed by using a simple t-test to test the null hypothesis of no difference between the Before and After sets of difference scores. This test is equivalent to the test for interaction between location (Impact versus Control) and condition (Before versus After) in a 2 x 2 mixed model analysis of variance; it is, however, much simpler to compute. The underlying null hypothesis in the difference model is that the difference between Impact and Control locations, averaged over several times in the After condition, does not change from the difference between Impact and Control locations, averaged over several times in the Before condition. A significant impact will thus only be found if it can be shown to exist over and above those short-to-moderate-range temporal changes which are known to occur naturally. In this model, replicate refers to the number of TIMES sampled in each condition. Thus, the level of sampling effort associated with a "replicate" in this model could be vastly different from that for either of the other two models.

Since the difference model addresses the question of whether the average difference between Impact and Control locations changes after the impact occurs, the appropriate denominator in the F test of the location x condition interaction is a pooled term containing both the time-by-location variance and the residual, or instantaneous replicate, variance. This term can be considered as the natural variability in the difference between location, and is identical to the within-condition variance of the difference scores. A significant impact will occur only when the condition-by-location interaction is large compared to the time-by-location variance plus residual, or instantaneous replicate, error.

Parameters Tested

Power tests were performed on three types of parameters: 1) individual (or indicator) species abundances, 2) community structure parameters, and 3) an ecological distance index based on species composition and abundance. Analysis of individual species was conducted on 8 to 15 of the most abundant and common species from each survey in each zone. Since power for individual species was consistently very low, we have presented only the mean power for all individual species in each category in Table 2. We also calculated power for four community structure parameters: diversity, evenness, total number of species, and total abundance. These measures are commonly used to assess community change, but they are not accurate indicators of differencies between communities because communities with completely different relationships to an outfall can have the same parameter value. For example, Figure 2 in Smith and Greene (1976) shows the same diversities (H') and the same total abundances at sites near the White Point outfall in Los Angeles and at other positions about 3 nmi (5.6 km) from the outfall, even though the community compositions are quite different.

In contrast to these measures of community structure, ecological distance indices have the potential for being quite sensitive to changes in community composition. We estimated power for tests based on a multivariate community distance index (Smith and Bernstein, in preparation). As the communities in two samples being compared become increasingly different, both in terms of the species present and their abundances, the distance index value increases. Since this measure incorporates information about community composition and abundances, it provides information on both the magnitude of change and whether changes in species composition are representative of more or less contaminated conditions.

In the past, ecological distances were not used to measure impacts because they do not always meet the requirements of parametric statistical tests. This is because the same sample can be used in the calculation of more than one distance value, thus violating the requirement of independence among observations. For the t-test and the two-way ANOVA models, we therefore corrected for this potential nonindependence with a solution proposed by Dyer (1978) and refined and then tested by Smith, Zalinski, and Bernstein (in preparation). We then used simulations to measure power directly for these models. The difference model did not require this correction because the distances between each pair of stations in the Before and After conditions are independent, analagous to Δ in the schematic above.

Table 2. Power of the three statistical models for a range of parameters. Power values for the t-test and two-way ANOVA are given as the probability of detecting a change of 0.1 in the ecological distance index at two replicates and of 50% of the mean for the univariate parameters at five replicates; for the difference model, values show the probability of detecting a change of 0.2 in the ecological distance index at two replicates and of 50% of the mean for the univariate parameters at five replicates

	Least Contaminated (Control)			Transition			Most Contaminated	
t-Test	**RA**[a]	**SD**	**SY**	**RA**	**SD**	**OC**	**SD**	**RA**
Individual species	0.34	0.42	0.36	0.40	0.39	0.39	0.42	0.28
Diversity	1.00	1.00	0.99	1.00	0.99	1.00	0.97	1.00
Evenness	1.00	1.00	1.00	1.00	0.99	1.00	0.97	1.00
Total number of species	0.98	0.97	0.86	1.00	0.99	1.00	0.98	0.94
Total abundance	0.82	0.88	0.75	1.00	0.87	1.00	0.72	0.83
Ecological distance	0.96			0.99				0.86
Two-way ANOVA	**RA**	**SD**	**SY**	**RA**	**SD**	**OC**	**SD**	**RA**
Individual species	0.23	0.31	0.26	0.28	0.28	0.33	0.34	0.19
Diversity	1.00	0.99	0.98	1.00	0.99	1.00	0.95	0.99
Evenness	1.00	1.00	1.00	1.00	0.98	1.00	0.95	1.00
Total number of species	0.60	0.77	0.65	0.95	0.97	1.00	0.92	0.77
Total abundance	0.90	0.92	0.79	1.00	0.74	0.96	0.61	0.69
Ecological distance	1.00			1.00				0.97
Difference model		**SD**		**HY**	**WP**	**SD**	**HY**	**WP**	**SD**
Individual species		0.34		0.27	0.22	0.27	0.27	0.09	0.40
Diversity		0.95		1.00	0.96	1.00	0.94	0.41	1.00
Evenness		0.99		1.00	1.00	1.00	0.87	0.37	1.00
Total number of species		0.78		0.55	0.33	0.97	0.78	0.66	0.76
Total abundance		0.79		0.81	0.47	1.00	0.47	0.11	1.00
Ecological distance		1.00			1.00	0.99		0.49	0.89

[a]See Table 1 for explication of data sets.

RESULTS

Power Tests

We found two main patterns in the power test results for all three statistical models (Table 2). First, there was uniformly low power for all individual or indicator species in all zones. The number of replicates required to reach a power of 0.80 at $\alpha = 0.05$ was often over 100. This finding demonstrates that it is not possible, for monitoring purposes, to detect anything but catastrophic changes in abundances, even of the most common and abundant species. Second, there was consistently high power for community structure parameters, with the exception of total abundance. Power for species richness, diversity, and evenness was often above 0.80 at just two replicates (t-test; two-way ANOVA) or two times (difference model). As a measure of sample species composition, ecological distances also showed very high power at two replicates. (The magnitude of ecological distance, or change, used in the power tests (0.1 or 0.2) is equal to or less than that which typically exists between instantaneous replicates within a given contamination zone.) There were no recognizable trends in power between the zones.

Implications for Monitoring

Sampling Design and Levels of Replication. The difference model is most appropriate for outfall monitoring programs because it 1) incorporates the time-by-location variance, which is an important feature of benthic shelf communities, and 2) provides a suitable framework for questions about long-term changes by monitoring over time. The properties of the difference model are described in more detail below.

Because the model incorporates two levels of sampling replication--at a single point in space and time (instantaneous replication) and through time (time replication)--a monitoring program based on the model requires decisions about the relative allocation of sampling effort between these two levels. That is, if a survey grid is to be sampled once a year for a given number of years, the number of instantaneous replicates to be taken at each station during each survey must be determined.

This requirement can be addressed in several ways. We first examined the use of optimization techniques that assess the relative contribution of each kind of replicate to reduction of the error variance in the analysis. Using data from transition stations at Point Loma (see Table 1), and equation 5 from Bernstein and Zalinski (1983), we estimated the increase in statistical efficiency of the difference model design as the

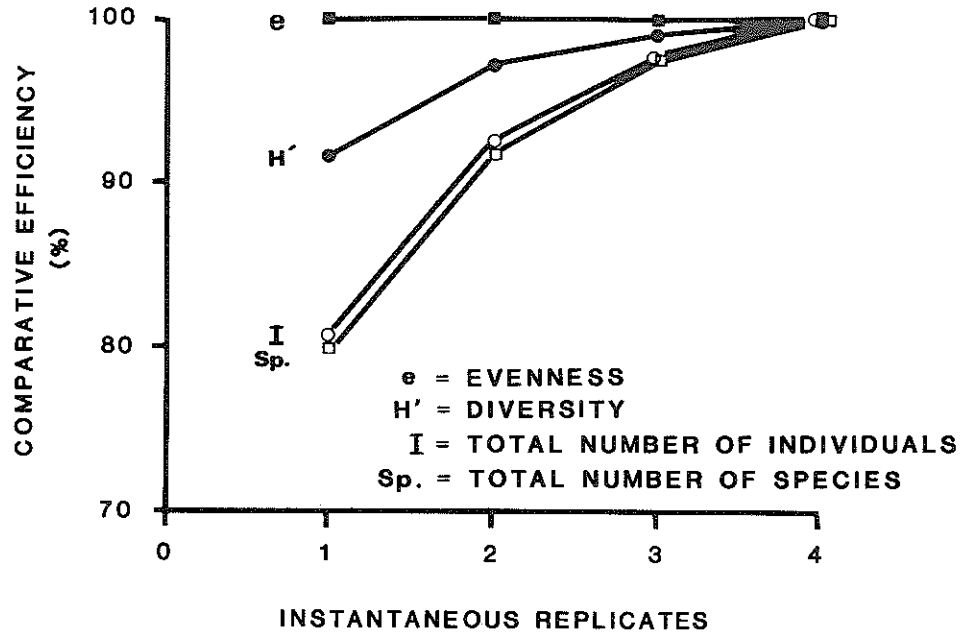

Figure 1. Increase in comparative efficiency using the difference model with additional instantaneous replicates, calculated with data from transition stations at Point Loma. Comparative efficiency is a function of relative efficiency (Sokal and Rohlf 1968, p. 289), which measures the percent decrease in the total error variance with increased replication.

number of instantaneous replicates increased from one to four (Figure 1). As described above, the error variance of this design is a composite of the time-by-location variance and the residual, or instantaneous replicate, variance. Figure 1 shows that designs with more instantaneous replicates were only slightly more efficient. The greatest increase in efficiency came between one and two instantaneous replicates.

We also calculated the power of the difference model at several levels of instantaneous replication. Using data from transition stations in the Point Loma surveys, we set the number of times in the Before and After conditions at 4. We then calculated power tests with all possible combinations of instantaneous replicates taken 1, 2, 3, and 4 at a time. The results (Figure 2) clearly show that additional instantaneous replicates contributed very little to the power of the test. In fact, for diversity, evenness, number of species, and ecological distance, each additional instantaneous replicate after 2 added at most 1% to power. We conclude that it is not efficient to allocate resources to sampling more than two instantaneous replicates; the same amount of effort would be better spent sampling and analyzing a larger and/or more complete grid of stations, or more times.

Another approach to the issue of instantaneous replication is to inspect

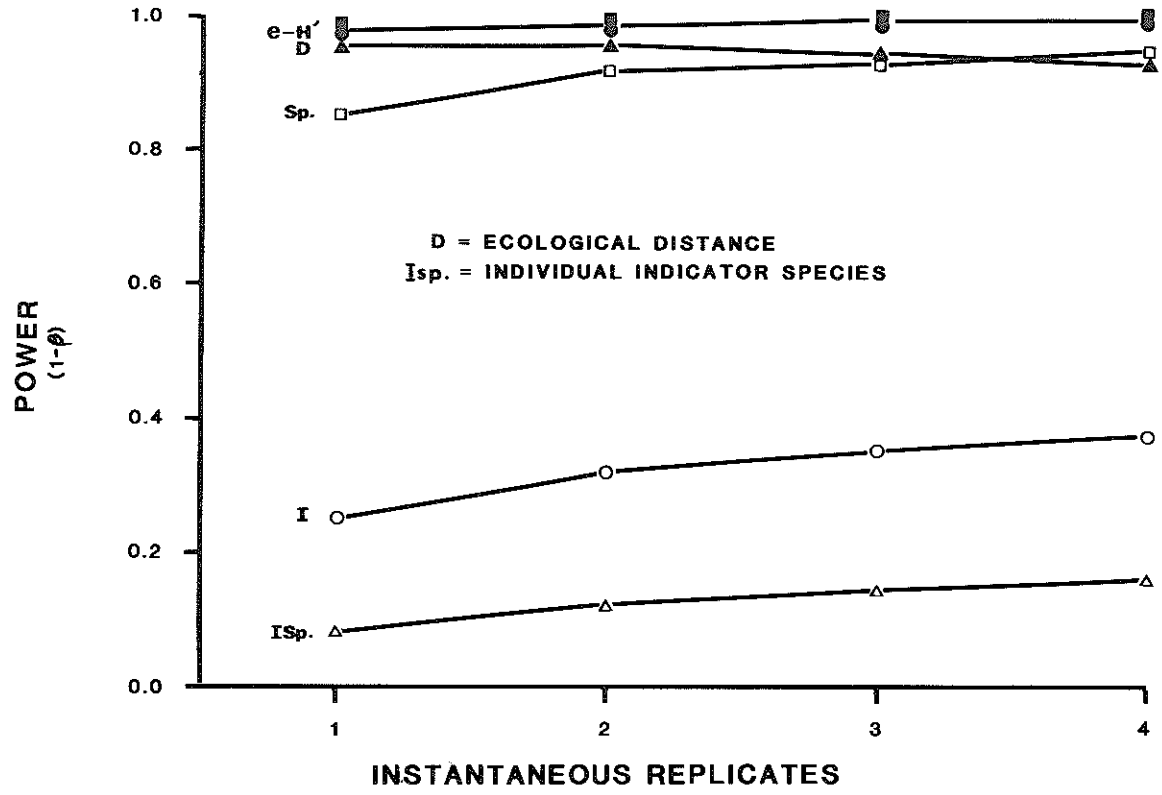

Figure 2. Increase in power of the difference model test with increasing numbers of instantaneous replicates, calculated with data from transition stations at Point Loma. Power values are given as the probability of detecting a change of 0.2 in the community distance index, or a 50% change in the mean of the univariate community parameters at $\alpha = 0.05$. (The number of times in the Before and After conditions was set at 4 for these power comparisons.)

the contribution of additional instantaneous replicates to the ability to detect patterns of community change. The community distance index values (i.e., the ecological distance matrix) are commonly used in cluster or ordination analyses to display such patterns. We performed an information gain analysis on the ecological distance matrices from the Point Loma survey and the SCCWRP replicate analysis (Thompson 1982). The relative amount of information contributed by each additional instantaneous replicate can be measured by comparing the distance matrix resulting from pooling all the replicates at a station-time with the distance matrices resulting from pooling successively fewer replicates. Information is defined as $100 \times r^2$, where r is the correlation between elements of the distance matrices being compared. The range of information content is defined as 0% at zero replicates and 100% at the maximum number (four at Point Loma; nine in the replicate analysis) of instantaneous replicates sampled.

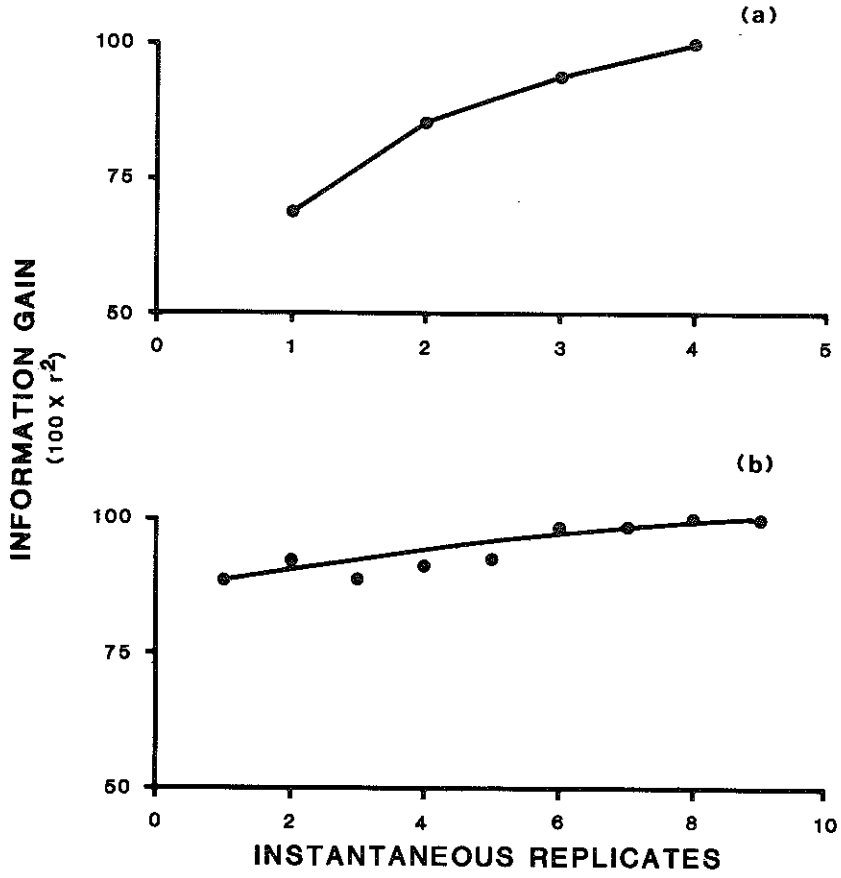

Figure 3. The information gain associated with different numbers of instantaneous replicates in the a) Point Loma survey and b) SCCWRP replicate analysis. (See text for additional details.)

Figure 3 shows information gain with successive instantaneous replicates, using all data from the Point Loma surveys and the SCCWRP replicate analysis. For the Point Loma data, 69% of the information about the community pattern contained in four instantaneous replicates was contained in one replicate, and 85% in two replicates. For the replicate analysis data, 89% of the information contained in nine instantaneous replicates was contained in one.

Direction of Change. Table 3 demonstrates that measures such as evenness, number of species, and total abundance are ambiguous indicators of change around outfalls. There is no difference in mean evenness between transition and contaminated sites. Additionally, number of species and total abundance are both likely to decrease with movement away from transition stations toward either normal or contaminated stations. In contrast, ecological distance readily distinguishes among stations along contamination gradients, because it always increases as stations from a control are compared to stations with increased levels of contamination. It is thus important to utilize a measure of community composition such as ecological distance when

Table 3. Mean values (numbers in parentheses are coefficients of variation) for community parameters in different contamination zones in southern California. Data are from the SCCWRP replicate analysis. The table also shows percent change in the value of each parameter from one zone to adjacent zones, along with the ecological distance between adjacent zones.

	Least Contaminated (Control)		Transition		Most Contaminated
Community parameters					
Diversity (H')	3.26 (7.7)		2.95 (6.7)		2.20 (9.2)
% change	←10		9→	←34	25→
Evenness	0.76 (5.7)		0.61 (6.2)		0.60 (7.3)
% change	←24		20→	←1	2→
Number of species	73.4 (15.7)		121.8 (8.7)		39.7 (19.6)
% change	←40		66→	←306	68→
Total abundance	446.2 (24.1)		1514.9 (15.0)		402.4 (21.9)
% change	←71		339→	←376	73→
Ecological distance between zones		0.30		0.50	

determining whether a particular area has changed to a less or more contaminated condition.

Establishing Levels of Detection in Monitoring Programs. As discussed above, monitoring programs must be designed to answer specific questions about levels of change in the parameters being monitored. There are, however, currently no criteria or guidelines for selecting levels of change to be detected. Since this important first step is so often neglected, we have developed the following example as an illustration of how available data can be used to set monitoring criteria.

Data from the SCCWRP replicate analysis (Table 3) show that there is a characteristic mean value and coefficient of variation (CV) for each univariate parameter in each zone. The CV is a measure of the variation in relation to the mean within a zone and can be used to set a lower limit of detection for the monitoring program. For example, the CV for number of species in the transition zone is about 9%. Based on the shape of the normal curve, this means that 68% of the measured values for number of species will be within one standard deviation (9%) of the mean. It would be an inefficient use of resources to attempt to detect changes in this parameter smaller than 9%, because changes of this magnitude are typically found among sites within the transition

zone. On the other hand, the mean number of species decreases 68% from the transition to the most contaminated zone. A monitoring program based on number of species should detect changes smaller than this in order to detect change before stations had become "most contaminated." The lower detection limit should therefore be somewhere between 9 and 68%, depending on the resources available. Similarly, a monitoring program based on the ecological distance index used here should be able to detect a change of between 0.2 (the distance between replicate stations in the transition zone) and 0.5. While this procedure does not provide a "magic number" for compliance purposes, it does provide a way to establish some guidelines based on actual data. Since each zone and parameter will yield a different range of values, it is necessary to examine site-specific data to establish such guidelines for a particular monitoring program.

SUMMARY AND CONCLUSIONS

We have emphasized the importance of thoroughly evaluating alternative monitoring programs before committing resources to them. Such evaluations should include at least the following six steps.

1. Develop explicit hypotheses about the type and size of impacts, or changes, that are to be monitored for. Without these, the ability of the monitoring program to detect change may be unrelated to the actual magnitude of the change it is important to detect.

2. Select the parameters that will be monitored as indicators of change.

3. Identify and quantify background sources of natural variability that may obscure the predicted changes.

4. Develop a statistical model that incorporates all relevant sources of variability and use it to provide a framework for the sampling and analysis designs.

5. Perform optimization analyses to allocate sampling effort effectively among different levels of sampling and replication in the design. If different levels of sampling have very different costs, optimization can also be used to derive the sampling plan that provides the most information per dollar.

6. Perform power tests to determine the ability of the sampling design to detect the changes predicted in 1. above.

We have followed this procedure, utilizing data from benthic monitoring programs around municipal waste outfalls in southern California and have arrived at several conclusions with important implications for outfall monitoring design. There are characteristic changes in community structure and composition that occur around outfalls. Commonly used community parameters (diversity, evenness, number of species, and total abundance) alone are not efficient indicators of these changes because they indicate change only in a gross feature of the community and provide no information about the direction of change (e.g., toward more or less affected communities). An ecological distance index is a much more sensitive and informative measure of community change. Monitoring programs must evaluate measures of both structure and composition to accurately understand how sewage discharge affects the benthos.

The appropriate statistical model for outfall monitoring is the difference model, in which surveys are repeated through time. Power tests with this model show that it is not possible to detect changes in individual species, even the most common and abundant ones, without inordinate amounts of sampling. It is possible to detect changes in community structure parameters, and as we have discussed, an ecological distance index provides information about the actual character of any change. Optimization and power tests with the difference model reveal that replicates at a single point in space and time (instantaneous replicates) add little to either the efficiency or the power of the design. Additional replicates beyond one or two typically added only a few percent to the statistical efficiency, and 2% or less to the power. As a result, more information can be gained by spreading samples out in space and time than by sampling more instantaneous replicates.

The body of data available from past and current monitoring programs makes it possible to rigorously design future programs by testing the utility of currently applied parameters and approaches to sampling design. We hope this report will provide a template--and a motiviation--for the design and evaluation of monitoring programs that accurately detect changes.

ACKNOWLEDGEMENTS

We wish to thank those investigators who made their data available to us, and who generously assisted us in error correction and interpretation. They include: Jan Stull, Los Angeles County Sanitation Districts; Susan Hamilton, City of San Diego Water Utilities Department; Douglas Diener and Larry Lovell, Marine Ecological Consultants; and John Dorsey, Los Angeles City Bureau of Sanitation.

LITERATURE CITED

Bascom, W., A.J. Mearns, and J.Q. Word. 1978. Establishing boundaries between normal, changed, and degraded areas. p. 81 IN: SCCWRP Annual Report, 1978, W. Bascom (ed.). Long Beach, Calif.

Bernstein, B.B., and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. J. Environ. Manage. 16:35-43.

Cohen, J. 1977. Statistical Power Analysis for the Behavioral Sciences. Academic Press, New York. 474 pp.

Dyer, D.P. 1978. An analysis of species dissimilarity using multiple environmental variables. Ecol. 59(1):117-125.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley-Interscience, John Wiley and Sons, New York. 257 pp.

Smith, R.W., and B.B. Bernstein. In Preparation. The measurement of distance in ecology.

Smith, R.W., and C.S. Greene, 1976. Biological communities near submarine outfall. J. Water Poll. Cont. Fed. 48(8) 1894-1912.

Smith, R.W., J. Zalinski, and B.B. Bernstein. In Preparation. Hypothesis testing with ecological distances.

Sokal, R.R., and F.J. Rohlf. 1969. Biometry. W.H. Freeman and Co., San Francisco. 776 pp.

Thompson, B.E. 1982. Variation in benthic assemblages. pp. 91-98 IN: SCCWRP Biennial Report, 1981-1982, W. Bascom (ed.). Long Beach, Calif.