OCS Study
MMS 99-0062

Long-term Monitoring of Biological Parameters at a Proposed Produced Water Discharge: Application of a BACIPS Assessment Design

Final Technical Summary

Final Study Report

U.S. Department of the Interior
Minerals Management Service
Pacific OCS Region

Long-term Monitoring of Biological Parameters at a Proposed Produced Water Discharge: Application of a BACIPS Assessment Design

Final Technical Summary

Final Technical Report

Author

Craig W. Osenberg
Principal Investigator

Prepared under MMS Cooperative
Agreement Nos. 14-35-0001-30471 & 14-35-0001-30761
by
Southern California Educational Initiative
Marine Science Institute
University of California
Santa Barbara, CA 93106

U.S. Department of the Interior
Minerals Management Service
Pacific OCS Region

Camarillo
March 1999

Disclaimer

This report has been reviewed by the Pacific Outer Continental Shelf Region, Minerals Management Service, U.S. Department of the Interior and approved for publication. The opinions, findings, conclusions, or recommendations in this report are those of the author, and do not necessarily reflect the views and policies of the Minerals Management Service. Mention of trade names or commercial products does not constitute an endorsement or recommendation for use. This report has not been edited for conformity with Minerals Management Service editorial standards.

Availability of Report

Extra copies of the report may be obtained from:
U.S. Dept. of the Interior
Minerals Management Service
Pacific OCS Region
770 Paseo Camarillo
Camarillo, CA 93010
phone: 805-389-7621

A PDF file of this report is available at:
http://coastalresearchcenter.ucsb.edu/SCEI/

Suggested Citation

The suggested citation for this report is:

Craig W. Osenberg. Long-term Monitoring of Biological Parameters at a Proposed Produced Water Discharge: Application of a BACIPS Assessment Design. MMS OCS Study 99-0062. Coastal Research Center, Marine Science Institute, University of California, Santa Barbara, California. MMS Cooperative Agreement Numbers 14-35-0001-30471 and 14-35-0001-30761. 52 pages.

Table of Contents

FINAL TECHNICAL SUMMARY

STUDY TITLES:
Study I. Ecological Effects of Chronic Exposure to Produced Water: A Field Test;
Study II. Environmental Effects of Produced Water: A BACIP Field Assessment

REPORT TITLE: Long-term monitoring of biological parameters at a proposed produced water discharge: application of a BACIPS assessment design

CONTRACT NUMBERS: Study I: 14-35-0001-30471; **Study II:** 14-35-0001-30761

SPONSORING OCS REGION: Pacific

APPLICABLE PLANNING AREA(S): Southern and Central California

FISCAL YEAR(S) OF PROJECT FUNDING:
Study I: 1989-90; 1990-91; 1991-92; 1992-93; 1993-94
Study II: 1994-95

COMPLETION DATE OF REPORT: March 1999

COSTS: FY 89-90 - $65,000; FY 90-91 - $66,531; FY 91-92 - $47,454; FY 92-93 - $75,000; FY 93-94 - $75,000; FY 94-95 - $86,000

CUMULATIVE PROJECT COST: STUDY 1: $328,985**; STUDY II:** $86,000

PROJECT MANAGERS:
Study I: [1]C.W. Osenberg, [2]S.J. Holbrook, [2]R.J. Schmitt
Study II: [3]M.H. Carr, [2]S.J. Holbrook, [1]C.W. Osenberg

AFFILIATION: [1]University of Florida; [2]University of California, Santa Barbara; [3]University of California, Santa Cruz

ADDRESS: [1]Department of Zoology, University of Florida, Gainesville, FL 32611; [2]Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA 93106; [3]Department of Biology, University of California, Santa Cruz, Santa Cruz, CA 95064

PRINCIPAL INVESTIGATORS:
Study I: [1]C.W. Osenberg, [2]S.J. Holbrook, [2]W.W. Murdoch, [2]R.J. Schmitt
Study II: [3]M.H. Carr, [2]S.J. Holbrook, [3]P.T. Raimondi, [1]C.W. Osenberg

KEY WORDS: Santa Barbara Channel; produced water; BACI; BACIPS; field assessment; assessment design; long-term monitoring.

BACKGROUND:

Population densities and many other environmental variables of interest vary tremendously among different sites and at different times. As a result, it can be difficult to discern the biological effects of produced water (or any other perturbation being studied) from other sources of spatial and temporal variation (which may arise naturally or from other anthropogenic activities) (Osenberg *et al.* 1994). Prior to the start of our project, environmental effects of produced water had been poorly studied (Neff 1987, Spies 1987), due in part to the application of flawed assessment design (NRC 1990). Furthermore, vast amounts of produced were being discharged into coastal waters. As a result, the study of produced water was identified as a critical gap in the study of environmental effects of oil and gas production (Boesch and Rabalais 1987). We proposed to use the Before-After-Control-Impact Paired Series assessment design (BACIPS: Stewart-Oaten *et al.* 1986, Schmitt and Osenberg 1996) to quantify potential ecological effects associated with the nearshore discharge of produced water from a coastal facility located near Gaviota, California.

OBJECTIVES:

The original goal of this project was to provide an unambiguous test of the localized ecological effects of produced water. We focused on the application of the Before-After-Control-Impact Paired Series (BACIPS) assessment design, and aimed 1) to obtain a good time series of data prior to the discharge of produced water, and 2) to provide logistical support for other projects investigating specific processes operating at the chemical, physical and demographic levels.

Because the BACIPS design requires an extensive time series of data prior to an anthropogenic activity (such as produced water discharge), unforeseen changes to the planned activity can compromise successful execution of a BACIPS study (Piltz 1996). Indeed, the produced water outfall upon which we focused our studies, never went into full operation. As a result, our objectives were modified to 1) advance the theoretical developments of the BACIPS assessment design (Stewart-Oaten *et al.* 1992, Osenberg *et al.* 1994), 2) increase the application of the BACIPS design when suitable opportunities arise (Osenberg and Schmitt 1994, Schmitt and Osenberg 1996), and 3) use our long-term series of Before data to estimate natural spatial and temporal variation and evaluate its implications for the success of BACIPS designs. In particular, we estimated the statistical power of BACIPS designs to detect impacts on chemical-physical, and individual-based and population-based biological parameters (Osenberg *et al.* 1992, 1994).

DESCRIPTION:

We proposed to study a produced water discharge located near Gaviota, California (discharge was supposed to occur in ~27 m water depth). We initiated our Before sampling in 1988 and expanded our program during the subsequent two years. The field program focused on the enumeration of benthic infauna, epifauna, and demersal fishes, growth and tissue production of mussels transplanted to the study site, and characterization of a variety of chemical and physical attributes (e.g., temperature, sedimentation rates, grain size of sediments, trace metal concentrations in the water column and sediments). The more complex chemical analyses were done in conjunction with colleagues at UC Santa Cruz and UC Davis. Field sampling for many of the field parameters continued through October 1995 when it was concluded that produced water would never be discharged at the site. We used the resulting time series of data to examine patterns of temporal and spatial variability in environmental data as a means to evaluate the statistical power of subsequent BACIPS designs.

In addition to the fieldwork, we also expanded the conceptual development of the BACIPS design and encouraged the application of BACIPS to other environmental studies. We accomplished this through publications in peer reviewed journals and books, publication of a Special Feature in *Ecological Applications*, the organization of and participation in workshops and meetings, and the publication of a book (Schmitt, R.J. and C.W. Osenberg. 1996. *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego. 401 pages). The book represented a major undertaking and comprised a large portion of our activities during the latter stages of this project. We also helped train over 30 students and field and laboratory assistants, and incorporated our results in several new courses that were developed at UC Santa Barbara and UC Berkeley.

SYNOPSIS OF MAJOR FINDINGS:

In our field studies, we 1) successfully developed and implemented the "before" phase of the assessment design, 2) obtained time-series data on population densities of infaunal and epifaunal organisms (as well as basic physical and chemical parameters), and 3) obtained field samples for analysis by colleagues at UC Santa Cruz and UC Davis. We used the time series data to estimate natural temporal variability and used these data, together with data from a study of another produced water outfall, to estimate the statistical power of BACIPS assessment designs. Between-site differences in chemical - physical parameters (e.g., elemental concentration) and in individual-based biological parameters (e.g., body size) were quite consistent through time, whereas differences in population-based parameters (e.g., density) were more variable. The magnitude of effects was estimated to be greatest for population-based parameters and least for chemical - physical parameters, which tended to balance the statistical power associated with these two parameter groups. Individual-based parameters were intermediate in estimates of effect size. The ratio of effect size to variability (and thus statistical power) was greatest for individual-based parameters and least for population and chemical - physical parameters. The results suggest that relatively few of the population and chemical - physical parameters could provide adequate power given time constraints of most studies.

We also evaluated the assumptions of the BACIPS design and explored related analytical issues. For example, Carpenter *et al.* (1989) proposed Randomized Intervention Analysis (RIA), arguing that it was robust to violations of assumptions required by parametric tests of BACIPS data (such as normality, equal variances, additivity and independence). If true, RIA would greatly increase the applicability of BACIPS. However, contrary to these assertions, our analyses showed that randomization tests are unlikely to be more valid than parametric tests (Stewart-Oaten *et al.* 1992). The assumptions that time and location effects be additive and that Impact-Control differences be independent through time are critical to both parametric and randomization tests. These assumptions must be tested, and if they are violated, remedial action should be taken. In general, this will require a long time-series of data that enables appropriate model development and testing. We have addressed these issues using our "before" time series and have shown that serial correlation in the time series of differences was relatively small in our data set and did not appear to vary among the parameter groups (Osenberg *et al.* 1994).

STUDY PRODUCTS:

PUBLICATIONS (listed chronologically; * indicates most important publications: reprints included with the Final Study Report):

*Stewart-Oaten, A., J.R. Bence and C.W. Osenberg. 1992. Detecting effects of unreplicated perturbations: no simple solution. Ecology 73:1396-1404.

Osenberg, C.W., S.J. Holbrook and R.J. Schmitt. 1992. Implications for the design of environmental assessment studies. Pages 75-90 *In* P.M. Grifman and S.E. Yoder (eds.) *Perspectives on the Marine Environment*. USC Sea Grant, Los Angeles, California.

Osenberg, C.W. and R.J. Schmitt. 1994. Detecting human impacts in marine habitats. Ecological Applications 4:1-2.

*Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecological Applications 4:16-30.

*Schmitt, R.J. and C.W. Osenberg (editors and contributing authors). 1996. *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

*Osenberg, C.W. and R.J. Schmitt. 1996. Detecting ecological impacts caused by human activities. Pages 3-16 in R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1996. Detection of environmental impacts: natural variability, effect size, and power analysis. Pages 83-108 *in* R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego. [Republication of Osenberg *et al.* 1994, Ecological Applications 4:16-30.]

Schmitt, R.J., C.W. Osenberg, W.J. Douros, and J. Chesson. 1996. The art and science of administrative environmental impact assessment. Pages 279-291 *in* R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press.

Ambrose, R.F., R.J. Schmitt, and C.W. Osenberg. 1996. Predicted and observed environmental impacts: can we foretell ecological change? Pages 343-367 *in* R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

Canestro, D., P.T. Raimondi, D.C. Reed, R.J. Schmitt, and S.J. Holbrook. 1996. A study of methods and techniques for detecting ecological impacts. Pages 53-67 in *Methods and Techniques of Underwater Research, Proceedings of the American Academy of Underwater Scientists Symposium*, AAUS, Nahant, MA.

RESEARCH PRESENTATIONS:

Papers and posters presented at regional and national meetings and workshops:

Herrlinger, T.J. and C.W. Osenberg. 1989. Demographic and behavioral responses of benthic marine organisms to produced water discharge: sensitive indicators and links to population dynamics. Third Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, University of California, San Francisco (poster).

Osenberg, C.W., R.J. Schmitt and S.J. Holbrook. 1989. An impact assessment design for detecting ecological effects of discharged produced water. Third Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, University of California, San Francisco (poster).

Osenberg, C.W., A. Stewart-Oaten and J.R. Bence. 1990. The analysis of environmental impact data resulting from the Before-After-Control-Impact-Paired (BACIP) design. Fourth Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, University of California, Santa Barbara (poster).

Osenberg, C.W., R.J. Schmitt and S.J. Holbrook. 1990. Differential ability to detect environmental impacts arising in ecological systems. Fourth Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, University of California, Santa Barbara (poster).

Osenberg, C.W., R.J. Schmitt and S.J. Holbrook. 1991. Implications for the design of environmental assessment studies. Symposium on *The Marine Environment*. 100th Anniversary of the Southern California Academy of Sciences. University of Southern California, May 1991. (invited)

Osenberg, C.W., S.J. Holbrook and R.J. Schmitt. 1991. The power to detect unreplicated perturbations varies among physical, chemical and biological parameters. Ecological Society of America, San Antonio, Texas.

Osenberg, C.W., S.J. Holbrook, and R.J. Schmitt. 1991. The spatial scale of ecological effects associated with point-source discharges. Fifth Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, San Francisco, California. (poster)

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook and D. Canestro. 1992. Spatial scale of ecological effects associated with an open coast discharge of produced water. International Produced Water Symposium, San Diego, California.

Osenberg, C.W., R.J. Schmitt and S.J. Holbrook. 1992. Detection of environmental impacts: power analysis and spatial inference. Symposium on *The design of environmental impact assessment studies: Conceptual issues and application*, held at the Second International Temperate Reefs Symposium, Auckland, New Zealand, January 1992.

Osenberg, C.W., A. Sberze, and J. Dai. 1993. Assessing ecological impacts of human activities in aquatic environments. Seventh Annual Research Symposium of the UC Toxic Substances Research and Teaching Program, University of California, Santa Cruz. November 1993 (poster).

Canestro, D., P.T. Raimondi, D.C. Reed, R.J. Schmitt, and S.J. Holbrook. 1996. A study of methods and techniques for detecting ecological impacts. Annual meeting of the *American Academy of Underwater Scientists*.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, C.M. St. Mary, and T.W.-M. Fan. 1998. Effects of produced water on mussel growth and production: application of the BACIPS design. 27th Benthological Ecology Meetings, Melbourne, Florida, March 1998.

Other (Invited Seminars and Workshops)

Osenberg presented results of this research during invited seminars at:

Department of Geography, University of California at Santa Barbara, November 1991.

Department of Zoology, Oregon State University, January 1994.

Department of Wildlife Ecology and Conservation, University of Florida, March 1997.

Department of Biology, University of Michigan, Ann Arbor, September 1997.

Department of Biological Sciences, Dartmouth College, September 1997.

Department of Zoology, University of Florida, September 1998.

The Design of Environmental Impact Assessment Studies: Conceptual Issues and Application. Schmitt and Osenberg organized this symposium, which was held at the Second International Temperate Reef Symposium, in Auckland, New Zealand (January 1992) and featured speakers from Australia, New Zealand and the United States. The US contingent included several SCEI investigators and an MMS scientist.

UC Toxic Substances Research and Teaching Program (Coastal Environmental Program). All Principal Investigators were intimately involved in the UC Toxic Substances Research and Teaching Program (Coastal Environmental Program), and attended annual workshops throughout the UC system between 1989 and the present. In addition to attendance at the workshops, MMS-sponsored research was often the focus of research presentations.

Coastal Toxicology Research in California. Schmitt organized this workshop involving University of California, local, state and federal agency personnel. Santa Barbara, CA. February 1993. (Osenberg was an invited participant)

Meta-analysis, interaction strength and effect size: application of biological models to the synthesis of experimental data. Osenberg designed and organized a working group at the National Center for Ecological Analysis and Synthesis, Santa Barbara, California, which met five times between July 1996 and May 1998. Applications of meta-analysis include the synthesis of environmental impact assessments.

Florida Big Bend coastal research workshop: toward a scientific basis for ecosystem management. Sponsored by Florida Sea Grant, UF Department of Fisheries and Aquatic Sciences, USGS (Florida Caribbean Science Center), and Suwannee River Water Management District. Steinhatchee, Florida, May 1997. (Osenberg was an invited participant)

Oral Presentations at Workshops

"Integrated study of environmental impacts: linking environmental changes with biological responses". Annual workshop of the UC Toxic Substances Research and Teaching Program (Coastal Environmental Program). Bodega Marine Laboratory, January 1990. (Osenberg)

"Statistical considerations in environmental assessment studies: the Before-After-Control-Impact-Paired design". Annual workshop of the UC Toxic Substances Research and Teaching Program (Coastal Environmental Program). Bodega Marine Laboratory, January 1990. (Osenberg)

Chair, working group on "Environmental effects of produced water discharge". Annual workshop of the UC Toxic Substances Research and Teaching Program (Coastal Environmental Program). Bodega Marine Laboratory, November 1991. (Osenberg)

"Eco-toxicological research at the University of California". Annual workshop of the UC Toxic Substances Research and Teaching Program (Coastal Environmental Program). Bodega Marine Laboratory, September 1994. (Osenberg)

Oral Presentations at Public Meetings

"Ecological and Demographic Effects of Produced Water". Southern California Educational Initiative Site Visit, University of California at Santa Barbara, April 1990. (Osenberg)

"Demographic Effects of Produced Water". Southern California Educational Initiative Site Visit, University of California at Santa Barbara, March 1991. (Osenberg)

"Ecological Effects of Produced Water". Southern California Educational Initiative Site Visit, University of California at Santa Barbara, March 1991. (Osenberg)

"Environmental Effects of Produced Water". Presentation to the Minerals Management Advisory Board Outer Continental Shelf Scientific Committee, Santa Barbara, California, March 2, 1995. (Osenberg)

FINAL STUDY REPORT

Long-term Monitoring of Biological Parameters at a Proposed Produced Water Discharge: Application of a BACIPS Assessment Design

I. GENERAL INTRODUCTION

Motivation

Population densities and many other environmental variables of interest vary tremendously among different sites and at different times. As a result, it can be difficult to discern the biological effects of produced water (or any other perturbation being studied) from other sources of spatial and temporal variation (which may arise naturally or from other anthropogenic activities) (Osenberg *et al.* 1994). Indeed, most field assessments have yielded equivocal results at best, and at the time that we initiated this study, the biological effects of produced water were poorly understood (Neff 1987). For example, previous field assessments of produced water effects, most of which have been conducted in the Gulf of Mexico region, have confounded the effects of produced water with natural variability and/or effects arising from other types of human activities (Spies 1987, Carney 1987, Osenberg and Schmitt 1996). Because produced water had been so poorly studied and because vast amounts of produced water were being discharged into coastal waters, the study of produced water was identified as a critical gap in the study of environmental effects of oil and gas production (Boesch and Rabalais 1987). Understanding environmental effects of produced water requires the application of improved assessment designs (NRC 1990).

We proposed to use the Before-After-Control-Impact Paired Series assessment design (BACIPS: Stewart-Oaten *et al.* 1986, Schmitt and Osenberg 1996) to quantify potential ecological effects associated with the nearshore discharge of produced water. Application of the BACIPS design permits the separation of the produced water "signal" from natural "noise". In its simplest design, BACIPS requires simultaneous sampling of at least one Control site (away from the outfall) and at least one Impact site (near the outfall) several times Before and again After discharge of produced water. During the Before period, the difference between the Impact and Control sites estimates natural spatial variability, and thus the expected difference during the After period if the intervention has no effect. The estimated difference from the Before period is then compared statistically to the actual difference observed during the After period. If assumptions underlying BACIPS have been satisfied (Stewart-Oaten *et al.* 1986, 1992), a statistically significant result is taken as evidence of an environmental impact. The size of the impact and the confidence in this estimate can also be estimated using the Before and After time series.

Objectives

The original goal of this project was to provide statistically reliable information on the existence and magnitude of localized ecological effects that result from the nearshore discharge of produced water. We focused on the application of the Before-After-Control-Impact Paired Series (BACIPS) assessment design, and aimed 1) to obtain a good time series of data prior to the discharge of produced water, and 2) to provide logistical support for other projects investigating specific processes operating at the chemical, physical and demographic levels. These other projects were also funded through the Southern California Educational

Initiative, as well as the related UC Toxic Substances Research and Teaching Program. Through these collaborations and the application of a rigorous statistical design we hoped to provide an unambiguous test of the localized effects of produced water and associated impacts.

Because the BACIPS design requires an extensive time series of data prior to an anthropogenic activity (such as produced water discharge), unforeseen changes to the planned activity can compromise successful execution of a BACIPS study (Piltz 1996). Indeed, the produced water outfall upon which we focused our studies, never went into full operation. As a result, our data do not provide a test of produced water effects. Our studies do, however, provide a comprehensive time series of data that can be used to provide critical information needed to evaluate field assessment designs and thus guide the design and application of future assessment studies. As a result, our objectives were modified. Instead of focusing on the estimation of produced water effects, we aimed to 1) advance the theoretical developments of the BACIPS assessment design (Stewart-Oaten *et al.* 1992, Osenberg *et al.* 1994), 2) increase the application of the BACIPS design when suitable opportunities arise (Osenberg and Schmitt 1994, Schmitt and Osenberg 1996), and 3) use our long-term series of Before data to estimate natural spatial and temporal variation and evaluate its implications for the success of BACIPS designs. In particular, we estimated the statistical power of BACIPS designs to detect impacts on chemical-physical, and individual-based and population-based biological parameters (Osenberg *et al.* 1992, 1994).

Basic Approach

We proposed to study a produced water discharge located near Gaviota, California. Because we planned to use the BACIPS design, it was vital that we had sufficient time to collect Before data prior to the discharge of produced water. We were fortunate to have advance warning about this facility and the proposal to discharge produced through a diffuser located at a depth of ~27 m. We initiated our sampling in 1988 under funding from the UC Toxic Substances Research and Teaching Program. We sampled three sites: a Control site (~1600 m upcurrent from the diffuser), a Near Impact site (~50 m downcurrent from the diffuser) and a Far Impact site (~250 m downcurrent). In 1989 (the first year of the SCEI program), we expanded our sampling and began processing the field samples. The field program focused on the enumeration of benthic infauna, epifauna, and demersal fishes, growth and tissue production of mussels transplanted to the study site, and characterization of a variety of chemical and physical attributes (e.g., temperature, sedimentation rates, grain size of sediments, trace metal concentrations in the water column and sediments) – see Osenberg *et al.* 1994 for discussion of methodology (see Section V.). The more complex chemical analyses were done in conjunction with colleagues at UC Santa Cruz and UC Davis: we made the field collections and turned the samples over to them for subsequent analysis. Table 1 summarizes the types of samples that have been collected as part of this study.

Field sampling for many of the field parameters continued through October 1995 when it was concluded that produced water would never be discharged at the site. Since that time, no additional samples have been collected, and some samples collected during the last year were never processed. Despite our inability to perform a test of produced water impacts at the Gaviota study site, we used the resulting data to examine patterns of temporal and spatial variability in environmental data as a means to evaluate the statistical power of subsequent

BACIPS designs. In a companion study, we were able to conduct an "unplanned" BACIPS study when a produced discharge abruptly went out of service (see Schmitt and Osenberg, *Ecological Responses to, and Recovery From, Produced Water Discharge: Application of a BACIPS Assessment Design*). Furthermore, the infaunal samples have been made available to other projects designed to evaluate in more detail patterns of spatial and temporal variation and the effects of taxonomic aggregation (Carr *et al.*, Detecting Ecological Impacts: Effects of Taxonomic Aggregation in the Before-After/Control-Impact Paired Series Design).

Table 1. Samples Collected From Gaviota Study Sites As Part Of BACIPS Design

Sample Type	First Sample Date	No. Sampling Dates	No. Dates Processed[1,2]
Infaunal Density[3]	Feb 88	45	42
Band Transects	Feb 88	46	46
Emergence Rates[3]	Jun 89	31	31
ReEntry Rates[3]	Jan 90	26	26
Lytechinus Quadrats	Feb 88	45	45
Lytechinus Size/GSI	Feb 89	32	32
Grain size	Oct 89	22	22
Organic Matter	Oct 89	25	25
Sediment Traps	Dec 89	28	28
Temp./Currents	May 89	50	50
Mussel growth	Sep 89-Apr 90	10	10
Samples sent to colleagues at UCSC and UCD[2]:			
Sediment chemistry	Apr 89	20	11
Water column elements	Apr 89	20	12
Water column organics	Nov 92	20	0
Mussel body burdens	Apr 90	10	0

[1]This gives number of dates on which samples have been processed. In a small number of cases, this may include dates in which data were not available from all three sites (e.g., when *Lytechinus* densities were 0 at the Far Impact site, we could not estimate size-distributions).
[2]We do not know how many of the chemical samples have been processed. We have indicated only those samples for which we have received data.
[3]The invertebrate samples are only sorted to rough taxonomic levels. Species level identifications will be obtained for a subset of samples under a new project by Carr, Holbrook, and Osenberg.

In addition to the fieldwork, we also expanded the conceptual development of the BACIPS design and encouraged the application of BACIPS to other environmental studies. We accomplished this through publications in peer-reviewed journals, publication of a Special Feature in *Ecological Applications*, the organization of and participation in workshops and meetings, and the production of a book (Schmitt, R.J. and C.W. Osenberg. 1996. *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego. 401 pages). The book represented a major undertaking and comprised a large portion of our activities during the latter stages of this project. We also helped train over 30 undergraduate and graduate students and post-graduate field and lab assistants.

In the following sections, we provide several reprints of major publications resulting from this study. One of our primary accomplishments was the publication of our book, which contains chapters contributed from many SCEI investigators (and MMS personnel). This book

represents the most current and rigorous treatment of field assessment designs and has received excellent reviews (e.g., Fairweather 1996, Rachlin 1996). For example, Rachlin (1996) noted that the "book provides an intelligent and comprehensive overview of the state of the art and science of ecological impact assessment as it is currently practiced and, more importantly, how it should be practiced." (Rachlin, 1996). This book represents a general treatment of the theoretical issues that motivated the SCEI program and more specifically focuses on the application of the BACIPS design, a central focus of this specific project.

II. Contents of: R.J. Schmitt and C.W. Osenberg (eds.). 1996. *Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats*. Academic Press, San Diego.

III. The need for sound field assessments.

Osenberg, C.W. and R.J. Schmitt. 1996. Detecting ecological impacts caused by human activities. Pages 3-16 *in* R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

DETECTING ECOLOGICAL IMPACTS CAUSED BY HUMAN ACTIVITIES

Craig W. Osenberg and Russell J. Schmitt

Ecologists and environmental scientists have long sought to provide accurate scientific assessments of the environmental ramifications of human activities. Despite this effort, there remains considerable uncertainty about the environmental consequences of many human-induced impacts, particularly in marine habitats (e.g., NRC 1990, 1992). This is especially surprising when one considers the vast amounts of capital and human resources that have been expended by industry, government, and academia in reviewing, debating, and complying with, a plethora of environmental regulations, which often require extensive study and documentation of environmental impacts. As we face an ever increasing number of environmental problems stemming from human population growth, it is critical that we achieve better understanding of the effects of humans. Due to the variety of human activities that potentially affect ecological systems, it also is imperative that we discriminate among effects of specific types of disturbances (rather than focus on an overall effect without regard to the particular sources), so that we can identify and give adequate attention to those that are most "harmful" (e.g., having the biggest effects, or which affect the most "valuable" resources). This requires approaches that can isolate effects of particular activities from nonhuman sources of natural variation as well as background variation caused by other anthropogenic events. Such approaches should reduce the uncertainty that underlies the documentation of effects of anthropogenic impacts and thus facilitate solutions to many of these problems.

Uncertainty surrounding the effects of anthropogenic activities arises from limitations imposed during the two scientific processes that comprise environmental impact assessment: (i) the predictive process, aimed at detailing the likely impacts that would arise from a proposed activity (most recently termed "Risk Assessment"; Suter 1993), and (ii) the postdictive process, aimed at quantifying the actual impacts of an activity (sometimes called "retrospective risk

Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats, edited by R. J. Schmitt and C. W. Osenberg

assessment," and which we will refer to as "Field Assessment"). Instead of standing alone, these two processes should proceed in tandem and build upon each other; resolution of many environmental issues requires both quantification of impacts, as well as accurate prediction of future impacts. Neither process substitutes for the other. For example, a prediction reveals little unless the prediction is an accurate indicator of actual effects; the accuracy of predictions can only be assessed after repeated tests (i.e., comparison with the actual outcomes). Similarly, documenting an impact that has already occurred yields only a limited ability to improve environmental planning (i.e., by avoiding environmental problems, or facilitating environmentally safe activities) unless we use this information to construct or refine frameworks that enable us to accurately anticipate future environmental impacts (e.g., predict their magnitude, and know the likelihood of such impacts based on the type of activity, its location, or the system being affected). The development of such frameworks is crucial to sound decisions being made before an activity occurs.

To date, we have only a limited ability to accurately predict the ecological consequences of many anthropogenic impacts (e.g., Culhane 1987, Tomlinson and Atkinson 1987, Buckley 1991a, 1991b, Ambrose et al., Chapter 18). For example, audits of environmental impact assessments (i.e., comparisons of predicted impacts with those actually observed) often have found relatively good agreement between predictions and reality when focused on physical or engineering considerations (e.g., the amount of copper discharged from a wastewater facility), but poor or limited agreement when focused on biological considerations (e.g., impacts on population density). As Ambrose et al. (Chapter 18) point out, the poor agreement stems both from lack of quantitative (or often qualitative) predictions of ecological change in the predictive phase of assessments, as well as lack of knowledge of the actual impacts (due to the absence or poor design of Field Assessments). Advances in Risk Assessment are likely to promote more precise predictions and thus reduce the first hurdle; however, the second will continue to plague us until Field Assessment studies are designed that better isolate effects of human activities from other sources of variation. Furthering this latter goal is a major theme of this book.

The field of environmental impact assessment is quite broad, requiring expertise from a diversity of fields, including physics, chemistry, engineering, toxicology, ecology, sociology, economics, and political science. In assembling the components of this book, we did not seek a comprehensive treatise on impact assessment. Instead, we focused on a narrower but central topic: the Field Assessment of localized impacts that potentially affect ecological systems (we further use marine habitats to provide the context for the discussions). We chose this conceptual focus because a wealth of books have appeared in the past 10 years that deal well with other aspects of environmental assessment (e.g., Risk Assessment: Bartell et al. 1992, and Suter 1993; general overviews of EIA: Westman 1985, Wathern 1988, Erickson 1994, Gilpin 1994; general introduction to monitoring: Spellerberg 1991; see also Petts and Eduljee 1994). None of these

books, however, deals with the issue of Field Assessments in more than a cursory way (but see NRC 1990 for a good introduction to the role of environmental monitoring); there are no detailed discussions of sampling designs that can most reliably estimate the magnitude of the impacts and quantify the power of the designs to detect impacts, nor are there evaluations of institutional and scientific constraints that limit the application of such designs. This book is designed to fill that gap, and in so doing, provides grist for future discussion and advances that are critically needed to better understand effects of anthropogenic activities.

Two major tenets, which we elaborate below, underlie this book: (i) that Field Assessments are absolutely essential to understanding human impacts, in part, because they complement, and provide field tests of, predictions provided by Risk Assessment; and (ii) that improved sampling designs are critical to improving the quality and utility of results obtained from Field Assessments.

The Need for Field Assessments

The emerging field of Ecological Risk Assessment (Bartell et al. 1992, Suter 1993) has led to a tremendous increase in the precision and explicitness of predictions of anthropogenic impacts on ecological systems. These predictions are often based on models derived from laboratory studies of toxicological effects, transport models that describe the movement of contaminants, and population models that attempt to couple physiological and demographic changes with shifts in population dynamics and abundances. However, no degree of sophistication of such models can guarantee the accuracy of the predictions. The quality and applicability of Risk Assessment can only be judged by the degree to which its predictions match the impacts that actually occur. This requires estimation of the magnitude of the impact, not just its detection, and thus requires a Field Assessment that is able to separate natural spatial and temporal variability from variation imposed by the activity of interest. This is not a trivial problem, and many previous assessments have failed in this regard. The small number of successes that exist are too few to permit any sort of rigorous evaluation of Risk Assessment models.

As Risk Assessment models become more complex and sophisticated, it is possible that they will be championed as the final step in environmental impact assessment; follow-up Field Assessments might be deemed a waste of effort (redundant with the effort devoted to obtaining the predictions). While this is a worthy (but elusive) goal, no Risk Assessment model, no matter how sophisticated, is currently capable of accurately predicting ecological change in response to an anthropogenic activity. As mentioned above, this, in part, is due to the lack of knowledge about the actual response of many systems to anthropogenic disturbances, and therefore the general inability to compare predicted and observed change.

Uncertainty in predictions from Risk Assessment models often is acknowledged but typically is limited to two sources: (i) uncertainty about the actual

value of parameters that are estimated from the studies that underlie the Risk Assessment model; and (ii) uncertainty about the environmental inputs to the model (e.g., how much freshwater runoff will enter an estuary during an upcoming year). Estimates of these sources of error often are incorporated into a Risk Assessment analysis to estimate the uncertainty associated with the prediction(s) of the model. Another, perhaps more important source of uncertainty rarely is examined: the uncertainty that the model chosen exhibits dynamics that are quantitatively (or even qualitatively) similar to the dynamics exhibited by the actual system. For example, improved laboratory techniques might provide improved quantification of the effect of a toxicant on the fecundity of a focal organism. This information might then be used in a model that links toxicant exposure with fecundity, and fecundity with population growth. However, even if the effect of the toxicant can be accurately extrapolated to field conditions, there is little guarantee that the connection between fecundity and population dynamics has been modeled correctly. More generally, the predicted dynamics may bear little resemblance to the observed dynamics, not because of uncertainty in the laboratory measurements, but due to uncertainty in the structure of the model into which the laboratory data are embedded. Addressing this uncertainty requires extensive field data, including information on the link between physiological changes and behavior (e.g., habitat selection, mate selection, reproductive condition), demographic consequences (e.g., changes in survival, birth rates, migration), population-level responses (e.g., shifts in age-structure, temporal dynamics), community responses (e.g., due to shifts in the strengths of species interactions), and ecosystem properties (e.g., feedbacks between biotic shifts and the physio-chemical aspects of the environment). Ultimately, these field data, together with laboratory data and the Risk Assessment model(s), must be integrated and then tested via comparison with actual responses to specific human activities. This last step requires properly crafted Field Assessment designs of sufficient power to distinguish the effects of the activity from a diverse set of other processes that drive variation in ecological systems.

The (In-)Adequacy of Exisitng Field Assessment Designs

The Goal of Field Assessments

A basic goal of a Field Assessment study is to compare the state of a natural system in the presence of the activity with the state it would have assumed had that activity never occurred. Obviously, we can never know, or directly observe, the characteristics of a particular system (occupying a specific locale at a specific time) in both the presence and absence of an activity. Thus, fundamental goals of the assessment study are to estimate the state of the system that would have existed had the activity not occurred, estimate the state of the system that exists with the activity, and estimate the uncertainty associated with the difference

between these estimates (Stewart-Oaten, Chapters 2 and 7). The inability of most studies to accomplish these goals has, in large part, led to tremendous uncertainty regarding the environmental consequences of anthropogenic activities. We briefly review some of these design considerations, beginning with an often misunderstood approach borrowed from modern field ecology—the manipulative field experiment.

The Role of Field Experiments

Manipulative field experiments (with spatial replication of independent subjects, and randomized assignment of subjects to treatment groups) is a common and powerful tool of field ecologists. However, field experiments can do very little to resolve the specific goal of Field Assessments. This issue (i.e., the application of experimental design to assessment) has clouded much of the debate about the design of Field Assessment studies (e.g., Hurlbert 1984, Stewart-Oaten et al. 1986). While field experiments may provide crucial insight into the functioning of systems and the role of particular processes (typically acting over limited spatial and temporal scales), a field experiment cannot reveal the effects of a specific activity on the system at a specific locale at a specific time, which is often the focus of a Field Assessment. A field experiment could provide a powerful way to determine the average effect of a process (e.g., an anthropogenic activity) defined over replicates drawn at random from a larger population of study (assuming that we could conduct such a replicated experiment on the appropriate spatial and temporal scale). However, this field experiment could not tell us about the effect of the treatment on any one of the replicates. Yet, this is analogous to the problem faced in Field Assessments.

To illustrate, consider the possible environmental impacts related to offshore gas and oil exploration, specifically those associated with the discharge of drilling muds. We could conduct an experiment to address whether oil exploration has localized effects on benthic infauna inhabiting a particular region, say the Southern California Bight, by (i) randomly selecting a subset of sites within the Bight and allocating these between "Control" and "Treatment" groups; (ii) drilling exploratory wells and releasing muds in our Treatment sites; (iii) quantifying the abundances of infauna in the Control and Treatment sites after a specified amount of time (e.g., 1 year); and (iv) determining if there is sufficient evidence to reject the null hypothesis of "no effect" (e.g., are the means of the two groups sufficiently different to be unlikely to have arisen by chance?) using standard statistical procedures (e.g., a *t*-test).

Clearly, this is an unlikely scenario (few oil companies would be willing to have a group of ecologists dictate where they will conduct their exploration), but in some situations, such an opportunity might exist. If so, then we will be in a tremendous (and enviable) position to estimate the average local effect of oil exploration on benthic fauna inhabiting the Southern California Bight. While such a study would provide invaluable information, the results would say nothing

about the effect of a single oil platform drilling exploratory wells at a specific site within the Bight. Indeed, it is possible that a significant (and biologically important) treatment effect could be found in our experiment even if there were no effects at a majority of the Treatment sites (so long as the remaining Treatment sites were sufficiently affected). There are, of course, situations where knowledge of the average affect of an activity would be quite useful (e.g., the administrative process of Environmental Impact Assessment). However, in most Field Assessments we are less concerned with the average effect and more concerned with the specific effects of a particular project at a specific locale. This is analogous to the experimentalist pondering the effect of the treatment on a single replicate (and not a collection of replicates).

Furthermore, an oil company certainly does not randomly select sites for exploration. It always could be argued that part of the selection criteria included the need to find sites that not only yield oil or gas, but also are sites in which oil and gas could be found and extracted without any environmental damage; resolution of the issue thus requires specific information about specific locales. Therefore, we require a tool more powerful, or at least more specific, than the replicated field experiment with randomized assignment.

Instead of manipulative field experiments, the basic tools used in Field Assessments involve monitoring of environmental conditions. Many such monitoring designs bear superficial resemblance to one another, but differ in some fundamental aspects. In the next section, we draw on discussions from Underwood (1991) and Osenberg et al. (1992) to clarify some of these distinctions. We illustrate the basic elements of the most commonly used assessment designs, and summarize results from studies with which we have been associated to illustrate where these studies can go wrong.

The Control-Impact Design

Perhaps the most common Field Assessment design involves the comparison of a Control site (a place far enough from the activity to be relatively unaffected by it) and an Impact site (i.e., near the activity and thus expected to show signs of an effect if one exists); a common variant involves a series of Impact sites that vary in their proximity to the activity. This sort of design often is part of the monitoring program required by regulatory agencies for various coastal activities. Environmental parameters typically are sampled at the two sites (with multiple samples taken from each site), and an "impact" is assessed by statistically comparing the parameters at the Impact and Control sites.

We illustrate this approach in Figure 1.1a, which shows that the density of a large gastropod (*Kelletia kelletii*) was significantly greater at a Control site (1.6 km from a wastewater diffuser) than at either of two Impact sites (located 50 and 250 m from the diffuser). This difference might be taken as evidence that the discharge of wastewater had a negative effect on the density of the gastropod.

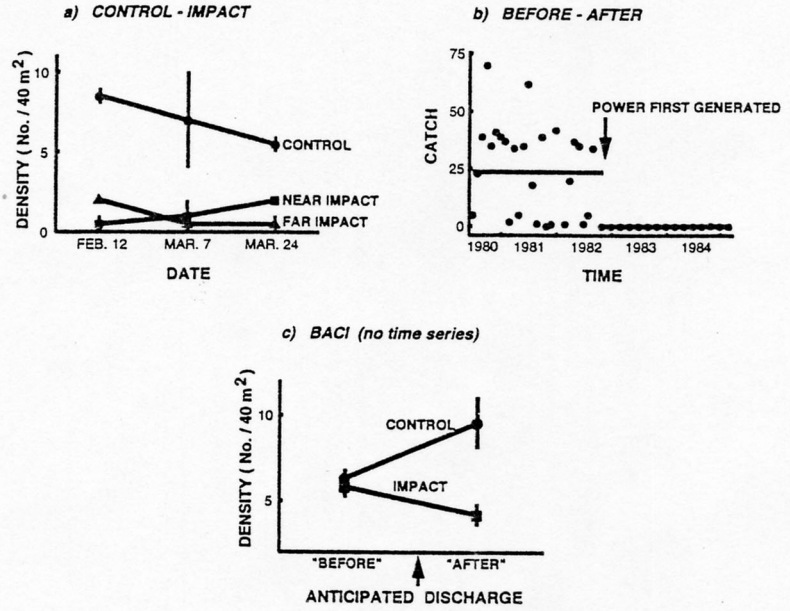

Figure 1.1. Three commonly used assessment designs that confound natural variability with effects of the anthropogenic activity. (a) The Control-After design showing the density of the snail *Kelletia kelletii* at three sites over time. The Near (square) and Far (triangle) Impact sites are located 50 and 250 m downcurrent of a wastewater outfall; the Control site (circle) is 1500 m upcurrent. These data were collected prior to discharge of wastewater. Shown for each date are the mean and range of gastropod density (n = 2 band transects per site). (b) The Before-After design showing density (catch per otter trawl) of pink surfperch *Zalembius rosaceus* over time at a location 18 km from the San Onofre Nuclear Generating Station (SONGS). The arrow indicates the first date on which power was generated by two new units of SONGS. Mean densities during the Before and After periods are indicated by the solid lines. (c) The Before-After-Control-Impact (BACI) design of Green (1979) showing the density of seapens *Acanthoptilum* sp. at two sites. The Control site is located 1500 m upcurrent, and the Impact site 50 m downcurrent, of a wastewater outfall. Because of permitting and production delays, discharge of wastewater did not begin when expected; all data were collected prior to discharge. Shown are means (± SE) using all observations within a period as replicates. The figure is adapted from Osenberg et al. (1992).

However, these data were collected prior to the discharge of wastewater. Thus, these differences observed during the Before period simply indicate spatial variation arising from factors independent of the effects of wastewater. To be applied with confidence, the Control-Impact design requires the stringent and unrealistic assumption that the two sites be identical in the absence of the activity. However, ecological systems exhibit considerable spatial variability, and it is extremely unlikely that any two sites would yield exactly the same result if sampled sufficiently. This design fails to separate natural *spatial* variability from effects of the activity.

The Before-After Design

An alternative design requires sampling of an Impact site both Before and After the activity; this avoids problems caused by natural spatial variation. Here, a significant change in an environmental parameter (e.g., assessed either by comparison of one time Before and one time After using within site sampling error as a measure of variability, or sampled several times Before and After and using the variation in parameter values through time as the error term) is taken as evidence of an "impact". Figure 1.1b provides an example for a fish, pink surfperch (*Zalembius rosaceus*), sampled Before and After the generation of power from new, seawater-cooled units of a large nuclear power plant (DeMartini 1987). The precipitous decline in abundance of pink surfperch is suggestive of a dramatic and detrimental impact from the power plant. However, these data are from a Control site 18 km from the power plant (a similar pattern also was seen at an Impact site: DeMartini 1987). Instead of indicating an impact, these data simply reflect the effect of other processes that produce temporal variability (in this case, it was an El Niño Southern Oscillation event that began at the same time as initiation of power generation: Kastendiek and Parker 1988). Applied in this way, the Before-After design fails to separate natural sources of *temporal* variability from effects of the activity.

A more sophisticated Before-After design is possible, and a classic example of intervention analysis (Box and Tiao 1975) provides both an illustration of its successful application and helps identify why the approach is limited for most ecological studies. Box and Tiao estimated the influence of two interventions (a traffic diversion and new legislation) on the concentration of ozone in downtown Los Angeles. Their procedure required that they (i) frame a model for the expected change; (ii) determine the appropriate data analysis based on this model; (iii) diagnose the adequacy of the model and modify the model until deficiencies were resolved; (iv) make appropriate inferences. Their analysis provided estimates of the effect of each intervention on ozone concentration.

There are several features of their system/problem that facilitated their successful analysis: (i) there was a long and intensive time series of ozone samples (hourly readings were available over a 17-year period, which included several years during the pre- and postintervention periods); (ii) the dynamics of ozone concentration were fairly well behaved, with repeatable seasonal and annual patterns; (iii) the number of pathways for the production and destruction of ozone were relatively few. These features contrast markedly with many ecological systems, where (i) we often have little expectation of how the system is likely to respond; (ii) data are sparse (time series are short, and intervals between sampling are long); and (iii) population density (for example) can be influenced by a multitude of processes (including a variety of mechanisms driven by abiotic factors and a wealth of mechanisms involving interactions with other species, each of which is also influenced by a variety of factors, including the effects of the focal activity). Certainly, such an approach might provide a powerful way to

assess responses of biological systems to interventions (Carpenter 1990, Jassby and Powell 1990), but currently it remains limited due to the paucity of detailed knowledge about dynamics of ecological systems. In cases where data from an unaffected Control site are available, we may be able to incorporate them into such time series analyses to compensate for the sparseness and complexity of ecological data (Stewart-Oaten, Chapter 7: see below).

Before-After-Control-Impact (BACI) Designs

One possible solution to the problems with the Control-Impact and Before-After designs is to combine them into a single design that simultaneously attempts to separate the effect of the activity from other sources of spatial and temporal variability. There are a variety of such designs. In the first, which we refer to simply as BACI (Before-After-Control-Impact), a Control site and an Impact site are sampled one time Before and one time After the activity (Green 1979). The test of an impact looks for an interaction between Time and Location effects, using variability among samples taken within a site (on a single date) as the error term. Data from our studies of a wastewater outfall (Figure 1.1c) demonstrate such an interaction; the decline in the density of seapens (*Acanthoptilum* sp.) at the Impact site relative to the Control site suggests that the wastewater had a negative effect on density of seapens. However, discharge of wastewater at this site was delayed several years, and did not occur when first anticipated. Thus, the observed changes were due to other sources of variability and were not effects of the wastewater. This design confounds effects of the impact with other types of unique fluctuations that occur at one site but not at the other (i.e., Time × Location interactions). Unless the two sites track one another perfectly through time, this design will yield erroneous indications that an impact has occurred.

To circumvent this limitation of Green's BACI design, Stewart-Oaten et al. (1986; see also Campbell and Stanley 1966, Eberhardt 1976, Skalski and McKenzie 1982) proposed a design based on a time series of differences between the Control and Impact sites that could be compared Before and After the activity begins. We refer to this design as the Before-After-Control-Impact Paired Series (BACIPS) design to highlight the added feature of this scheme (see Stewart-Oaten, Chapter 7 and Bence et al., Chapter 8). In the original derivation of this design (e.g., Stewart-Oaten et al. 1986), the test of an impact rested on a comparison of the Before differences with the After differences. Each difference in the Before period is assumed to provide an independent estimate of the underlying spatial variation between the two sites in the absence of an impact. Thus, the mean Before difference added to the average state of the Control site in the After period yields an estimate of the expected state of the Impact site in the absence of an impact during the After period: i.e., the null hypothesis. If there

were no impact, the mean difference in the Before and After periods should be the "same" (ignoring sampling error). The difference between the Before and After differences thus provides an estimate of the magnitude of the environmental impact (and the variability in the time series of differences can be used to obtain confidence intervals: Stewart-Oaten, Chapter 7 and Bence et al., Chapter 8).

The BACIPS design is not without its limitations, for it also makes a set of assumptions, which if violated can lead to erroneous interpretations (e.g., due to nonadditivity of Time and Location effects or serial correlation in the time series of differences). Indeed, one of the fundamental contributions of Stewart-Oaten's work (Stewart-Oaten et al. 1986, 1992, Stewart-Oaten, Chapter 7)has been to make explicit the assumptions that underlie the BACIPS design, pointing out the importance of using the Before period to generate and test models of the behavior of the Control and Impact sites, and to suggest possible solutions if some of the assumptions are violated. Importantly, many of these assumptions can be directly tested. Of course, it is still possible that a natural source of Time × Location interaction may operate on the same time scale as the study, and thus confound interpretation of an impact. However, this problem is far less likely than those inherent to the other designs (e.g., that variation among Times and Locations be absent and that there be no Time × Location interaction).

In this volume, Stewart-Oaten (Chapter 7) and Bence et al. (Chapter 8) elaborate upon and apply a more flexible BACIPS design based on the use of the Control site as a "covariate" or predictor of the Impact state, which might have even greater applicability than the original design (which was based on the "constancy" of the differences in the Before and After periods). Underwood (1991, Chapter 9) has suggested a "beyond-BACI" approach, which incorporates multiple Controls, as well as random sampling of the study sites (thus, the "Paired Series" aspect of the BACIPS design is not present in Underwood's beyond-BACI design). Underwood suggests that the beyond-BACI design is able to detect a greater variety of impacts than the BACIPS design (e.g., detection of pulse responses as well as sustained perturbations); however, he also notes that his design is not able to deal explicitly with problems of serial correlation. By contrast, the presence of serial correlation can be directly assessed, and appropriate action taken, when applying the BACIPS design (Stewart-Oaten et al. 1986, 1992, Stewart-Oaten, Chapter 7). In a variety of important ways, Underwood's approach differs from Stewart-Oaten's and some others represented in this book (e.g., Osenberg et al., Chapter 6, Bence et al., Chapter 8). While both schools-of-thought advocate the advantages of using more than one Control site, they do not agree on the ways in which this added information should be incorporated into the analyses. We expect that debate on these issues is far from over, and hope that this book serves to further the discussion and facilitate advancements in the design and application of BACI-type studies.

The Organization of This Book

The issues highlighted above are tackled directly in the second section of this book, which is devoted to elaboration of the application and design of BACI-type studies. However, prior to the implementation of any Field Assessment a number of initial issues must be considered, and some of these are highlighted in the book's first section. For example, the general goal and purpose of the study is critical, and Stewart-Oaten's first chapter (Chapter 2) tackles the standard "*P*-value culture" that places undo emphasis on the detection of impacts, rather than estimation of their magnitude or importance. Ecological parameters must also be selected for study, and although the use of bio-indicators has been often criticized, Jones and Kaly (Chapter 3) point out that any study necessarily must select a limited number of parameters from the myriad available (thus necessitating the selection of a subset of "bio-indicators"). Once appropriate species (or parameters) are selected, sampling error can constrain our ability to discern the temporal dynamics of populations, and thus impair our ability to use time series analyses to assess ecological change (Thrush, Hewitt, and Pridmore, Chapter 4). Variability also limits the power of statistical tests of impacts, and Mapstone (Chapter 5) suggests a novel way to incorporate such a constraint directly into the permitting process by simultaneously weighting Type I and Type II errors in assessment studies.

The second section of the book (*Improving Field Assessments of Local Impacts: Before-After-Control-Impact Designs*), provides the core of the book and elaborates on the theory and application of BACI-type designs. Osenberg, Schmitt, Holbrook, Abu-Saba and Flegal (Chapter 6) provide a segue from Mapstone's discussion of statistical power by evaluating sources of error in BACIPS designs and specifically evaluating the power to detect impacts on chemical-physical vs. biological (individual-based vs. population-based) parameters. Stewart-Oaten (Chapter 7) follows with a theoretical treatment of BACI-type designs, which extends and generalizes much of the earlier research on BACI(PS). Bence, Stewart-Oaten and Schroeter (Chapter 8) apply this more general and flexible BACIPS design to data derived from an intensive study of the impacts of a nuclear power plant. In the final chapter in this section (Chapter 9), Underwood offers an alternative approach in which multiple Control sites are used to detect impacts in a different way than proposed by the other authors, and which potentially can detect a greater variety of impacts (e.g., pulses as well as sustained impacts).

While BACI-type designs offer great potential to detect impacts of local perturbations, they require sampling of a Control site(s) (a site(s) sufficiently close to the Impact site(s) to be influenced by similar environmental fluctuations, but sufficiently distant to be relatively unaffected by the disturbance). In many cases, such a control does not exist, or the significant biological effects of interest are dispersed over large spatial scales and therefore are difficult (if not impossible) to detect. In such cases, a BACI-type design or other empirical measure of

impact is unlikely to be able to quantify the effect with the desired level of precision. Therefore, we must be able to extrapolate results obtained from localized or smaller scale effects to those arising on larger spatial scales. This is the theme for the book's third section. Raimondi and Reed (Chapter 10) discuss how the spatial scales of impacts on chemical–physical parameters might differ from the scale of impacts on ecological parameters based on life-history features of the affected organisms. Larval dispersal is central to many of their points. Understanding the coupling between larval pools and benthic populations and their response to impacts will require integration of oceanographic models and ecological studies (Keough and Black, Chapter 11). Oceanographic processes may also "collect" larvae and pollutants in particular sites (e.g., along linear oceanographic features), and this aggregation of larvae in high concentrations of pollutants might amplify deleterious effects of many types of discharge (Kingsford and Gray, Chapter 12). Ultimately however, effects on larvae need to be translated into consequences at the population level, and in Chapter 13, Nisbet, Murdoch, and Stewart-Oaten provide an approach intended to provide an estimate of how local larval mortality, induced by a nuclear power plant, may impact the abundance of adults assessed at a regional level. Their work points out the critical need to better understand the role of compensation in the dynamics of fishes and other marine organisms affected by anthropogenic activities.

In the final section of the book, we return to the issue of Predictive vs. Postdictive approaches, emphasizing how and why the Predictive phase (which typically yields an Environmental Impact Report or Statement) should be integrated with the Postdictive phase (which yields the Field Assessment) to improve the overall quality of the entire process. The introductory chapter in this section (Chapter 14: Schmitt, Osenberg, Douros, and Chesson) provides a brief summary of the current state of the EIR/S process in the United States and Australia, and Carney (Chapter 15) evaluates the biological data that have been collected with regard to EIR/S studies (as well as Field Assessments). He concludes that numerous problems (including taxonomic errors, design flaws, statistical inaccuracies) plague even the most extensive studies. Piltz (Chapter 16) then elaborates on how institutional constraints can impede sound scientific investigations that require long-term monitoring. His chapter encourages both scientists and administrators to find solutions that ensure the research continuity that is required to obtain the most defensible Field Assessments. If EIR/S consistently fail to yield consensus, or if too few data exist to determine the accuracy of such studies, considerable debate can ensue. This often leads to judicial involvement in the EIR/S process, which in turn leads to tremendous effort expended on documentation during the EIR/S process, but with little additional clarity regarding the likely or actual environmental impacts of a project (Lester, Chapter 17).

Ultimately, the interplay between the Predictive process (EIR/S or Risk Assessment) and the Postdictive process (the Field Assessment) is critical to help guide our development of frameworks used to predict and understand anthro-

pogenic impacts: are our predictions accurate, and if not how might we modify our approach? Ambrose, Schmitt and Osenberg (Chapter 18), provide an audit of an intensive study of the San Onofre Nuclear Generating Station (SONGS) and compare effects that were predicted during the EIR/S process with those subsequently observed. Their findings reveal that the EIR/S process (even in such a major project) revealed little of the actual impacts; even a detailed scientific study conducted by an independent committee erred in a number of crucial ways. Their results demonstrate the need for continued vigilance in conducting well-designed monitoring studies, such as those using BACI-type analyses.

At times the tone of many of these contributions is rather critical of existing approaches and even the new approaches outlined in other chapters. It is only through healthy debate of the merits of alternative designs, and by better integration of administrative and scientific goals, that improvements in the EIR/S and Field Assessment processes occur. Indeed, despite this body of criticism, it is undeniable that refinements in our scientific tools have led to recent improvements in our understanding of anthropogenic impacts. This can only continue by avoiding complacency and by continuing to develop and refine new tools that can be used to tackle these important issues. Of course, realization of our ultimate goal depends upon expanding our basic knowledge of the dynamics and functioning of ecological systems, understanding the mechanisms by which anthropogenic activities impact these systems, and incorporating this information into models and theory to permit us to predict the occurrences of future impacts. But first, we must be better able to quantify the actual impacts that specific activities have induced in ecological systems. This simple goal is neither trivial nor commonly realized, but it is fundamental and achievable. Our hope is that this book helps to further advance understanding of the interactions between human activities and our impacts on our environment.

Acknowledgments

We gratefully acknowledge the helpful comments of S. Holbrook, C. St. Mary, and A. Stewart-Oaten. Research that led to the production of this chapter was funded by the Minerals Management Service, U.S. Department of Interior under MMS Agreement No. 14-35-001-3071, and by the UC Coastal Toxicology Program. The views and conclusions in this chapter are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.

References

Bartell, S. M., R. H. Gardner, and R. V. O'Neill. 1992. Ecological risk estimation. Lewis Publishers, Chelsea, Michigan.

Box, G. E. P., and G. C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association **70**:70–79.

Buckley, R. 1991a. Auditing the precision and accuracy of environmental impact predictions in Australia. Environmental Monitoring and Assessment **18**:1–24.

Buckley, R. 1991b. How accurate are environmental impact predictions? Ambio **20**:161–162.

Campbell, D. T., and J. C. Stanley. 1966. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, Illinois.

Carpenter, S. R. 1990. Large-scale perturbations: opportunites for innovation. Ecology **71**:453–463.

Culhane, P. J. 1987. The precision and accuracy of U.S. Environmental Impact Statements. Environmental Monitoring and Assessment **8**:217–238.

DeMartini, E. 1987. Final report to the Marine Review Committee. The effects of operations of the San Onofre Nuclear Generating Station on fish. Marine Science Institute, University of California, Santa Barbara.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Erickson, P. A. 1994. A practical guide to environmental impact assessment. Academic Press, San Diego, California.

Gilpin, A. 1994. Environmental impact assessment (EIA). Cambridge University Press, Cambridge, England.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. Wiley and Sons, New York, New York.

Hurlbert, S. J. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**:187–211.

Jassby, A. D., and T. M. Powell. 1990. Detecting changes in ecological time series. Ecology **71**:2044–2052.

Kastendiek, J. and K. R. Parker. 1988. Interim technical report to the California Coastal Commission. 3. Midwater and benthic fish. Marine Review Committee, Inc.

NRC (National Research Council). 1990. Managing troubled waters: the role of marine environmental monitoring. National Academy Press, Washington, DC.

NRC (National Research Council). 1992. Assessment of the U.S. outer continental shelf environmental studies program. II. ecology. National Academy Press, Washington DC.

Osenberg, C. W., S. J. Holbrook, and R. J. Schmitt. 1992. Implications for the design of environmental impact studies. Pages 75–89 in P. M. Griffman and S. E. Yoder, editors. Perspectives on the marine environment of southern California. USC Sea Grant Program, Los Angeles, California.

Petts, J., and G. Eduljee. 1994. Envionmental impact assessment for waste treatment and disposal facilities. John Wiley and Sons, Chichester, England.

Skalski, J. R., and D. H. McKenzie. 1982. A design for aquatic monitoring systems. Journal of Environmental Management **14**:237–251.

Spellerberg, I. F. 1991. Monitoring ecological change. Cambridge University Press, Cambridge, England.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "psuedoreplication" in time? Ecology **67**:929–940.

Stewart-Oaten, A., J. R. Bence, and C. W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology **73**:1396–1404.

Suter, G. W., II, editor. 1993. Ecological risk assessment. Lewis Publishers, Boca Raton, Florida.

Tomlinson, P., and S. F. Atkinson. 1987. Environmental audits: A literature review. Environmental Monitoring and Assessment **8**:239–261.

Underwood, A. J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian Journal of Marine and Freshwater Research **42**:569–587.

Wathern, P. 1988. Environmental impact assessment. Unwin Hyman Ltd., London, England.

Westman, W. E. 1985. Ecology, impact assessment, and environmental planning. John Wiley and Sons, New York, New York.

IV. Assumptions of BACIPS and the (mis-)application of randomization procedures.

Stewart-Oaten, A., J.R. Bence and C.W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. *Ecology* **73**:1396-1404.

ASSESSING EFFECTS OF UNREPLICATED PERTURBATIONS: NO SIMPLE SOLUTIONS[1]

ALLAN STEWART-OATEN
Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

JAMES R. BENCE
National Marine Fisheries Service, Southwest Fisheries Science Center, Tiburon Laboratory, 3150 Paradise Drive, Tiburon, California 94920[2] and Marine Science Institute, University of California, Santa Barbara, California 93106 USA

CRAIG W. OSENBERG
Marine Science Institute, University of California, Santa Barbara, California 93106 USA

Abstract. We address the task of determining the effects, on mean population density or other parameters, of an unreplicated perturbation, such as arises in environmental assessments and some ecosystem-level experiments. Our context is the Before-After-Control-Impact-Pairs design (BACIP): on several dates Before and After the perturbation, samples are collected simultaneously at both the Impact site and a nearby "Control."

One approach is to test whether the mean of the Impact–Control difference has changed from Before to After the perturbation. If a conventional test is used, checks of its assumptions are an important and messy part of the analysis, since BACIP data do not necessarily satisfy them. It has been suggested that these checks are not needed for randomization tests, because they are insensitive to some of these assumptions and can be adjusted to allow for others. A major aim of this paper is to refute this suggestion: there is no panacea for the difficult and messy technical problems in the analysis of data from assessments or unreplicated experiments.

We compare the randomization t test with the standard t test and the modified (Welch-Satterthwaite-Aspin) t test, which allows for unequal variances. We conclude that the randomization t test is less likely to yield valid inferences than is the Welch t test, because it requires identical distributions for small sample sizes and either equal variances or equal sample sizes for larger ones. The formal requirement of Normality is not crucial to the Welch t test.

Both parametric and randomization tests require that time and location effects be additive and that Impact-Control differences on different dates be independent. These assumptions should be tested; if they are seriously wrong, alternative analyses are needed. This will often require a long time series of data.

Finally, for assessing the importance of a perturbation, the P value of a hypothesis test is rarely as useful as an estimate of the size of the effect. Especially if effect size varies with time and conditions, flexible estimation methods with approximate answers are preferable to formally exact P values.

Key words: environmental assessment; intervention analysis; pseudoreplication; randomization tests.

INTRODUCTION

A common problem in basic ecological studies and applied environmental work is to determine whether a particular population, community, or other object of interest has changed after a perturbation to the environment. The answer is often obtained by conducting an experiment, consisting of a number of replicates, each randomly assigned to one of several treatments, and then applying standard statistical analyses.

However, replication with randomly assigned treatments is not always possible. In assessing the effects of a particular power plant, we cannot randomly assign the location of the plant, or build more than one of them. Even in basic ecological work, although we can often randomly assign the perturbation to one or several of the experimental units, costs or the unavailability of replicates may make replication infeasible, particularly in whole ecosystem manipulations (Carpenter 1989, 1990, Carpenter et al. 1989).

In general, the major goal of a study of an unreplicated perturbation is to determine whether the state of the perturbed system differs significantly from what it would have been in the absence of the perturbation. Usually the "state" of the system is the mean value of some univariate or multivariate quantity, such as the population size, average size, or life history parameters

[1] Manuscript received 11 January 1991; revised 26 August 1991; accepted 28 August 1991.
[2] Present address.

of one or more species. We will assume the quantity of interest is univariate, e.g., the population abundance of a single species in a fixed area, although many of the general points we make also apply in the multivariate case.

Because the state of the system in the absence of the effect cannot be observed after the disturbance, we need to estimate what it would have been and compare the estimate statistically with the observed (perturbed) condition. The Before-After-Control-Impact-Pairs (BACIP) design (Stewart-Oaten et al. 1986) accomplishes this by collecting samples at both the Impact site and a nearby "Control" site. These samples are paired, in the sense that the Control and Impact sites are sampled simultaneously (as nearly as possible). Replication comes from collecting such paired samples at a number of times (dates) both Before and After the perturbation.

Each observed difference (e.g., in estimated population density) between the Impact and Control sites during the Before period is taken as an estimate of the mean difference that would have existed in the After period without the perturbation. The observed Impact–Control differences, one for each sample date, constitute a time series; we compare the differences from the Before period to those from the After period; a change in the mean difference indicates that the system at the Impact site has undergone a change relative to the Control site. The general process of estimating a change in a parameter, following a perturbation, has been termed "intervention analysis" (Box and Tiao 1975).

The BACIP design allows for natural differences between the Control and Impact locations, and for changes from the Before to the After period that influence both sites the same way (e.g., resulting from a large-scale change coincident with the putative local impact). Hypothetical examples are shown in Fig. 1.

But the design does not ensure that the assumptions of standard 2-sample tests, for comparing the "Before" set of differences to the "After" set, are satisfied. For the two-sample t test, the assumptions are:

1) Additivity: Time and location (site) effects are additive (i.e., in the absence of the perturbation, the expected Impact–Control difference is the same for all dates).

2) Independence: Observed differences from different dates are independent.

3) Identical Normal Distributions: The distribution of the deviation (observed difference–mean difference) is (a) the same for each time within a period; (b) the same in the After period as in the Before period; (c) Normal.

An adequate analysis must deal with these assumptions, either by supporting them (by arguing for their *a priori* plausibility and/or carrying out tests or other diagnostic procedures) or by showing that the analysis is not sensitive to their violation. This is a messy and complicated part of the analysis, which rarely can dispel doubt altogether. Thus tests needing fewer or more plausible assumptions could be valuable.

Recently, Carpenter et al. (1989) proposed "randomized intervention analysis" (RIA), which employs a BACIP design but uses a randomization test instead of a t test to decide whether there has been a change in the difference between the impact and control sites. They argue that a "distinct advantage" of RIA is that non-Normality does not affect the test results, and imply that this solves problems of temporal trends and time lags. They add that RIA is not affected by heterogeneous variances "unlike . . . the t test," and that the effects of serial correlation will often not lead to equivocal results.

We discuss assumptions (1), (2), and (3) in reverse order, with special reference to RIA, the standard t test, and the Welch (or Welch-Satterthwaite-Aspin) modification of the t test for unequal variances (Snedecor and Cochran 1980:97). We argue: (a) RIA's robustness to non-Normality offers little advantage: the two parametric t tests are also little affected by non-Normality unless sample sizes are very small; (b) the Welch t test is approximately valid when the Before and After distributions have different variances; the other two tests are not, unless sample sizes are nearly equal; (c) the Welch t test is approximately valid when the distributions vary within a period; the others are not, although they are approximately valid if the average Before variance is nearly the same as the average After variance; (d) if the successive differences are not independent, none of the tests is valid; they may be approximately valid if the dependence is weak (and the other assumptions hold); (e) if time and location effects are not additive, none of the tests is valid; they may be approximately valid if the effects are approximately additive.

We also discuss the general application of BACIP. We argue (1) that hypothesis testing, either classical or Bayesian, is less important than estimation of the effect's size and ecological assessment of its importance, and (2) that the appropriate statistical methods will often be unavoidably messy: effects may vary with environmental conditions that can be delineated only roughly, and estimates will depend on models, which are based partly on intuition, guesswork, and mathematical convenience, and must be supported by biological arguments and formal and informal diagnostic checks.

In what follows, we assume there are n_B Before dates and n_A After dates; on the i^{th} Before date, the estimated densities were I_{Bi} at the Impact site and C_{Bi} at the Control, for a difference of D_{Bi}. Similarly we have I_{Aj}, C_{Aj}, and D_{Aj} on the j^{th} After date. The average differences are $D_{B\cdot}$ and $D_{A\cdot}$. The randomization test takes the $(n_B + n_A)$ *values* (the D_{Bi}'s and D_{Aj}'s) as given but, under the null hypothesis, their *assignment* to "Before" or "After" is assumed to be random. The P value for

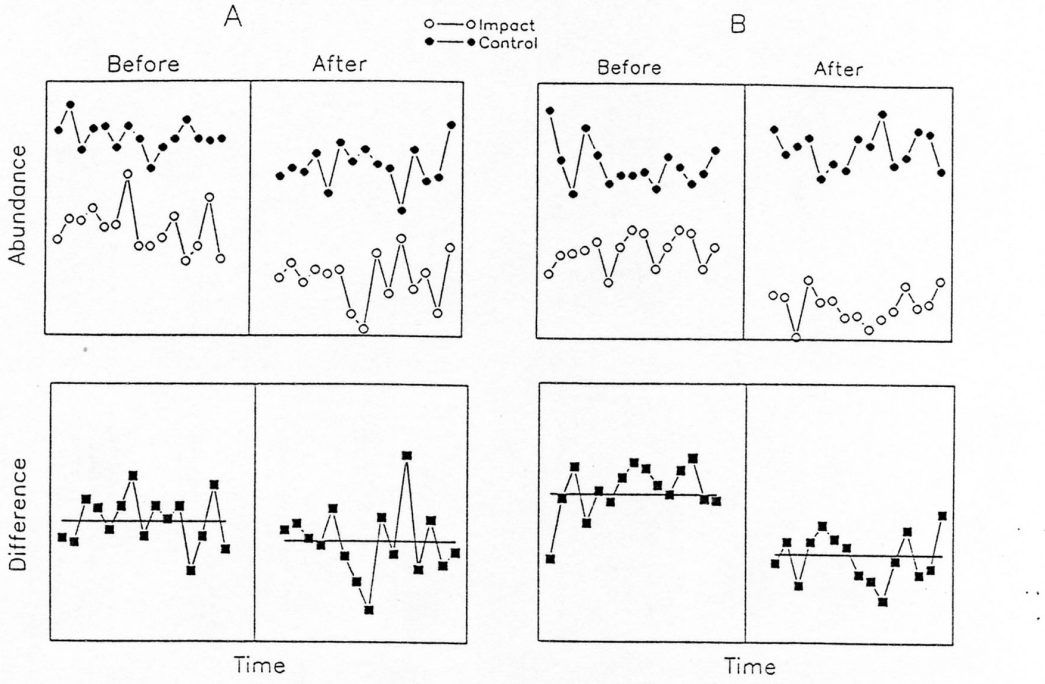

FIG. 1. Hypothetical examples of data collected using the Before-After-Control-Impact-Pairs (BACIP) design. (A) A case where average abundance is greater in the Control area than in the Impact area and where average abundance falls from Before to After. Note that the average difference between Impact and Control does not change significantly from Before to After (bottom panel), indicating that there has been no effect of the perturbation. (B) For comparison, a case where the perturbation has reduced the abundance of the species at the Impact site, leading to a decline in the difference from Before to After (bottom panel).

the test is then the fraction of the $(n_B + n_A)!/(n_B!n_A!)$ possible assignments that give a larger value of the test statistic than was actually observed (Pratt and Gibbons 1981: Chapter 6). The randomization t test uses $D_{B\cdot} - D_{A\cdot}$ (or an equivalent) as the test statistic.

Identical Normal Distributions

It is likely that one or more of these assumptions will fail. Many biological observations are non-Normal. Even without perturbation effects, distributions may well change between periods (Before and After), e.g., due to long-term weather patterns. They may also change *within* periods, e.g., the variance of an estimate of population density may be greater in summer than in winter.

Parametric t *tests*

Non-normality.—Strong evidence that the standard and Welch t tests are little affected by non-Normality comes from studies of both large and small sample sizes.

For large sample sizes, it is a direct result of the Central Limit Theorem: the usual t statistics are all approximately Normal, provided only that the parent distributions have finite variances.

For small sample sizes, there are a few analytical studies (Efron 1969, Tan 1982) providing evidence of the t test's robustness to non-Normality, but the main evidence comes from simulations, e.g., Yuen and Dixon (1973), Yuen (1974), Murphy (1976), Posten (1978, 1979), Tiku (1980), Gans (1981), Tiku and Singh (1982). Several others are reviewed by Glass et al. (1972).

A serious problem arises only from strong skewness. If $D_{B\cdot}$ and $D_{A\cdot}$ have different skewness, or have the same (non-zero) skewness but different variances, then $D_{B\cdot} - D_{A\cdot}$, the numerator of the t statistics, will have a skewed distribution. But such skewness is unlikely to be strong. Since I_{Bi} and C_{Bi} are estimates of similar things (e.g., population densities) based on similar sampling effort, they are likely to have similar skewness and variance: most of the skewness should cancel in the difference, $D_{Bi} = I_{Bi} - C_{Bi}$. More skewness is lost by averaging to get $D_{B\cdot}$, and still more in the difference, $D_{B\cdot} - D_{A\cdot}$, if these are similarly skewed, as is likely. If histograms of the D_{Bi}'s and D_{Aj}'s show pronounced skewness that is likely to persist through averaging and differencing, a modification of the Welch t test (Cressie and Whitford 1986) seems to solve the problem.

Distributions change between periods.—This creates little problem unless both variances and sample sizes are unequal, in which case the Welch t test is approximately valid, but the standard t test is not. The two

t statistics have the same numerator, $D_{B\cdot} - D_{A\cdot}$, which is approximately Normal by the Central Limit Theorem, except for the problem of skewness just described. Validity depends on the denominator, whose square should approximate the variance of $D_{B\cdot} - D_{A\cdot}$, with a relative error that approaches 0 as sample size increases.

The variance of $D_{B\cdot} - D_{A\cdot}$ is $S^2 = \sigma_B^2/n_B + \sigma_A^2/n_A$, where the σ^2's are the true variances. For the Welch t test, the denominator is

$$S_W = \sqrt{[s_B^2/n_B + s_A^2/n_A]}.$$

For the standard t test, the denominator can be written as

$$S_s = \sqrt{[M_B s_B^2/n_A + M_A s_A^2/n_B]}.$$

The s^2's are the usual variance estimates and, for $Z = A$ or B,

$$M_Z = (1 - 1/n_Z)/[1 - 2/(n_B + n_A)].$$

As n_B and n_A approach ∞, the s^2's approach the σ^2's and the M_Z's approach 1. Thus S_W^2 does approach S^2, but

$$S_s^2 \text{ approaches } s_B^2/n_A + s_A^2/n_B = R^2 S^2,$$

where $R^2 = (r\sigma_B^2 + \sigma_A^2)/(\sigma_B^2 + r\sigma_A^2)$ and $r = n_B/n_A$.

This shows that the Welch t test gives the correct level for large sample sizes; the simulations mentioned above verify this for moderate sample sizes. It also shows that, when its nominal level is α, the standard t test rejects the (true) null hypothesis that the Before mean is less than or equal to the After mean with probability approximately $\Phi(-Rz_\alpha)$, where Φ is the standard $N(0, 1)$ cumulative distribution function and z_α is the point for which $\Phi(-z_\alpha) = \alpha$. If either the sample sizes or the variances are approximately equal, $R \approx 1$ and the test is approximately valid. But if the smaller sample comes from the distribution with the larger variance, $R < 1$ and the test rejects more frequently than advertised. In the reverse case, it rejects less frequently.

One option is to use a test of equality of variances to decide whether to use the standard or the Welch t test. Simulations by Gans (1981) indicate that this is inferior to direct use of the Welch t test. In particular, it rejects too frequently when $R < 1$ and one variance is about half the other.

Distributions change within periods.—For this problem, some general results are given for estimates of location by Stigler (1976), for one-sample t tests by Cressie (1982) and for two-sample t tests by Cressie and Whitford (1986). The main large sample results are similar to those just described. The numerator of both t tests, $D_{B\cdot} - D_{A\cdot}$, is approximately Normal. (There is a condition for this, roughly that the variances not be so dissimilar that most of the variability of $D_{B\cdot}$ or $D_{A\cdot}$ comes from a small subset of the observations; see Feller 1966:491.) Its variance is $\sigma_{B\cdot}^2/n_B + \sigma_{A\cdot}^2/n_A$, where $\sigma_{B\cdot}^2 = \Sigma\, \sigma_{Bi}^2/n_B$ and σ_{Bi}^2 is the variance of the i^{th} "Before" difference. The Welch t test is approximately valid in general, because S_w^2 approaches this variance. S_s^2 does so only if $\sigma_{B\cdot}^2 = \sigma_{A\cdot}^2$, i.e., the standard t test is valid for unequal sample sizes only if the average Before and After variances are the same.

For moderate sample sizes, the Welch t test may be "liberal": its true rejection probability may be slightly greater than the nominal value because its degrees of freedom are overestimated. The standard formula divides an estimate of $2\{E[S_w^2]\}^2$ by an estimate of $V[S_w^2]$. With heterogeneous variances, the latter estimate, $s_B^4/n_B^2(n_B - 1) + s_A^4/n_A^2(n_A - 1)$, is biased low: roughly, for Normal variables, s_B^4 approaches $(\sigma_{B\cdot}^2)^2$ instead of the desired $\Sigma\, \sigma_{Bi}^4/n_B = (\sigma_{B\cdot}^2)^2 + V(\sigma_B^2)$, where $V(\sigma_B^2)$ is the variance of the set $\sigma_{B1}^2, \sigma_{B2}^2, \ldots$ (Cressie and Whitford 1986). But, since variances must be positive, $V(\sigma_B^2)$ is unlikely to be significantly larger than $(\sigma_{B\cdot}^2)^2$, which is overestimated by s_B^4, so the correct degrees of freedom are likely to be at least half the nominal value. If the nominal value is 30 or more, this error has little effect. Unfortunately, we know of no simulation studies of this case.

Randomization tests

The assumptions for randomization tests (which are sometimes called permutation tests) are usually satisfied in experiments by the investigator's deliberate random assignment of units to treatments. This is not possible in intervention analysis: one cannot randomly assign sampling times to "Before" and "After." Instead it is assumed that "Nature" does the random assigning: under the null hypothesis, the "Before" and "After" observations are assumed to be independent draws from a common distribution.

Thus all of the assumptions listed in the *Introduction* are required, except only assumption 3(c), Normality. The user of RIA must show either that these assumptions hold or that RIA remains valid when they fail.

The randomization test is *not* valid for unequal variances. For large sample sizes, it is invalid in the same way, and to the same extent, as the standard t test discussed above. The limiting level and power of the randomization test are the same as those of the standard t test. For equal variances, this result was proved by Hoeffding (1952), with the restriction that the original distributions have finite third absolute moments, in our notation, $E|D_{Bi}|^3 < \infty$ and $E|D_{Ai}|^3 < \infty$, which is satisfied in almost all realistic cases. Romano (1990) proves it without requiring either equal variances or the third moment restriction. Our moderate sample (20 and 40) simulations with Normal variables agreed closely with these asymptotic results. One of us has also extended Hoeffding's proof to the case where variances change within periods (A. Stewart-Oaten, *unpublished manuscript*): Romano's work suggests the third moment restriction is unnecessary here, too. Romano also shows that the randomization test based on

medians is invalid for non-identical distributions, even for equal sample sizes, unless the Before and After probability densities at their medians are equal or satisfy an unlikely condition.

For small sample sizes, it is easy to construct examples for which the randomization t test is invalid for non-identical distributions, even when the variances are the same.

Independence

Standard two-sample tests, including t tests and randomization tests (when these are based on randomization by "Nature" rather than by an experimenter), assume that the D_{Bi}'s and D_{Ai}'s are independent.

In the assessment problem, the most likely violation is positive serial correlation: observations (D_{Bi}'s and/or D_{Aj}'s) close in time may tend to be close in value. In this case, the variance of the average of the differences, e.g., $V(D_{B\cdot})$, is no longer the variance of a single observation divided by the sample size, e.g., $V(D_{Bi})/n_B$, but is larger. If this is not allowed for, all these tests will reject true null hypotheses more frequently than advertised, because observed averages will be less precise than they are assumed to be.

The observed D_{Bi}'s and D_{Aj}'s vary for two reasons. One is sampling error: the estimated Impact–Control difference at a given sampling time will not exactly equal the true difference at that time. But our concern is not with this "true difference," which itself varies naturally over time: any particular Before and After values are almost certain to be different even if there is no perturbation effect. Our concern is with the mean of the "true difference," i.e., the mean of the stochastic process of which the entire set of true differences over a period is a single realization (see Stewart-Oaten et al. 1986).

Correlation can arise from the second source of variation: the deviation between the true difference and its mean. This potential problem has been termed "pseudoreplication in time" (Hurlbert 1984). Two deviations will be correlated if the time between them is short enough that the same random events (births, deaths, movements, etc.) play significant roles in both. The variation in the true difference would then be underrepresented in the sample, leading to underestimation of the variance of $D_{B\cdot}$ or $D_{A\cdot}$.

Whether serial correlation in the observed differences is sufficient to invalidate the test for an effect must be assessed by formal tests and by *a priori* arguments and models based on knowledge of the populations under study. Stewart-Oaten et al. (1986) present arguments and a simple (though easily extended) model suggesting that, provided the additivity assumption holds, only large, local events (occurring at one site but not the other) should introduce serial correlation. Non-local events (e.g., storms) should have similar population consequences at both Impact and Control, and thus cancel (at least approximately) when we take differences. Small events (e.g., individual births and deaths) should not affect the populations for far into the future, and are likely to be swamped by the sampling errors (which are independent).

Correlation should be insignificant if sampling times are sufficiently separated so that a single event is unlikely to have a large local effect for more than one time. Arguments and models indicating how large a separation is needed should depend on the organism. For some populations, e.g., those which are short lived, highly mobile or strongly density dependent, local changes, even if large, will have only a brief effect.

For others, observations a year or more apart may be significantly correlated. A sedentary species whose larvae or seeds disperse unevenly in space over a short annual recruitment/settlement period is likely to have much the same local population within a year (between one recruitment period and the next) but quite different populations between years: variation in recruitment might be a long-lasting large local effect. Another case occurs when dispersion between Impact and Control sites is rare, as for lakes. For example, Osenberg et al. (1988) analyzed size-specific growth rates of sunfish in eight lakes over a 10-yr period and found that half of the interpretable variation arose from lake X year interactions. Since the growth of these fishes is closely tied to the availability of their resources (Mittelbach 1988, Mittelbach et al. 1988, Osenberg et al. 1988, Osenberg and Mittelbach 1989), these data suggest that the abundances of the invertebrate prey also exhibit lake X year effects. Some fish populations are also known to exhibit dramatic population cycles that may result from strong age class interactions (e.g., Aass 1972, Hamrin and Persson 1986, Townsend 1989), and the timing of these cycles may well vary from lake to lake. If variation in fish density cascades to lower trophic levels (Carpenter et al. 1987), then this could introduce local year (or even longer period) effects in a number of biological variables measured in a BACIP study.

There is no guaranteed resolution of these uncertainties. Whatever testing procedure is used should be derived from a model that is plausible and survives diagnostic checking against the data, both formal tests and informal inspection, especially plots. The plausibility is important. For example, a single year of data would be insufficient for a test of serial correlation in the examples just given, since the main source of variation, between years, is never observed. An implausible model might survive diagnostic checking in these cases, and could then be used to indicate a "perturbation effect" that was really natural year-to-year variation. In most cases, we would expect several years of Before and After data to be needed, with serial correlation of the Before differences checked by the Durbin-Watson (Durbin and Watson 1971) and Ljung-Box (Ljung and Box 1978) tests, and one-way ANOVA, using years as "treatments."

If serial correlation appears significant, either *a priori*

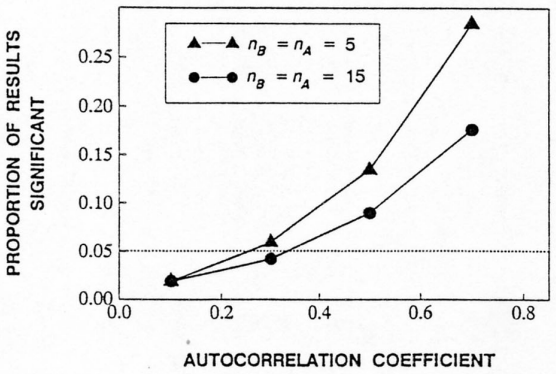

FIG. 2. An example of the effect of serial correlation on the randomization t test. The proportion of results that were significant (when H_0 is true and testing is done at the .01 level) is plotted against the value of the autocorrelation coefficient. Results are shown for sample sizes of 5 and 15 in each period. In all cases results are based on 5000 simulated trials, and randomization tests were based on the random selection of 5000 permutations. Data were generated from the same Gaussian autoregressive model of order one separately for each period.

or as a result of tests, the test for a change needs to be based on a model that includes plausible representations of the non-ignorable types of correlation (e.g., Box 1954, Box and Tiao 1965, 1975, Tiao et al. 1975, Jones 1980, 1981, McDowall et al. 1980), and is itself subjected to diagnostic checks (Box 1980).

Carpenter et al. (1989) recognize that serial correlation can inflate Type I error rates in randomization tests, but suggest that the rule "reject if the nominal P value is $< .01$" gives a conservative .05-level test. This rule lacks generality and seems to us undesirable. First, if the correlation is weak, this test is too conservative and is inefficient. Second, if the correlation is strong enough, the test is invalid. Fig. 2 shows that, for samples of 15 from a first-order autoregressive model with equal variances, the test is invalid if $r > 0.3$.

ADDITIVITY

Suppose the two populations vary but tend to track one another so that the density in the Impact area is typically 50% of that in the Control area. Then the difference between the raw Impact and Control densities will also vary. The effects of location and time on the means at the Impact and Control sites are not additive: the time effect does not cancel when we take the differences. Such non-additivity has three consequences.

Two arise when there is systematic (e.g., seasonal) variation in the overall density. The mean Impact–Control difference (the mean of the stochastic process mentioned in the previous section) then varies over time. The correct model for the data will not be the one the test is based on, i.e., $D_{Bi} = \mu_B + \epsilon_{Bi}$ and $D_{Ai} = \mu_A + \epsilon_{Ai}$, where the errors, ϵ_i, have mean zero, but $D_{Bi} = \mu_B + N_{Bi} + \epsilon_{Bi}$ and $D_{Ai} = \mu_A + N_{Ai} + \epsilon_{Ai}$ where the N_i's are non-random.

One consequence is that the test for an effect is not comparing μ_B with μ_A but comparing $\mu_B + N_{B\cdot\cdot}$ with $\mu_A + N_{A\cdot\cdot}$. These could differ solely because of the choice of sampling times, e.g., if the fraction of summer samples is higher in the Before period than in the After. Of course, we can balance the samples with respect to seasons, but there may be other cycles, perhaps unknown, that are not balanced.

Second, if we have balanced cycles, the N_i's will not bias the estimates of the means, but they will add to the estimated variances: the test will be more conservative (and less efficient) than it should be.

The third consequence arises from random natural variation, such as major storms or long spells of unusual weather. This changes densities in both areas; without additivity, it also changes their difference. Thus region-wide, long-lasting random variation may not tend to be cancelled when we take differences. The assumption that the observed differences are independent is then less plausible.

For hypothesis testing, the obvious way to satisfy the additivity assumption is to transform the data. If the data are multiplicative, as in the example above, we would expect to transform to logs. In practice, the "right" transformation may not be known, and various methods have been suggested for choosing a transformation in this situation (Tukey 1949, Box and Cox 1964, Andrews 1971, Carroll and Ruppert 1981, 1984, Hinkley and Runger 1984).

It may be that there is no monotone transformation for which the data (or the underlying process that produced them) are additive. For example, it may be that Impact densities are higher than Control in winter, but are lower in summer. In such cases a different analysis may be better. We return to this below. The main message is that the problem of non-additivity cannot be ignored, regardless of whether the final test is a t test, a randomization test, or something else.

EFFICIENCY

Validity is not the only important consideration in the choice of tests. We also want a test that is efficient, i.e., which has good power.

All three of the tests discussed here can be inefficient, because they are based on the Before and After sample averages. The average is, in some non-Normal cases, an inefficient estimator: for a given sample size, there are other unbiased estimators with much smaller variances for non-Normal distributions and only slightly larger variances for Normal distributions (Andrews et al. 1972). These "efficiency robust" estimators maintain small variances against a range of distributions by reducing the influence of the extreme observations.

A major virtue of randomization tests is the possibility of greater efficiency, from the use of robust estimates whose distributions are hard to determine, e.g.,

the median. However, as we have seen, these tests are likely to be invalid when the null distributions are not identical, as in the assessment problem.

Fortunately, there are robust estimators whose variances can be estimated. These can be used as the basis for "t-like" tests (both standard and Welch). Examples include trimmed means (Yuen 1974), biweight estimators (Kafadar 1982), modified maximum likelihood estimators (Tiku and Singh 1982), and many others (Andrews et al. 1972). Perhaps the easiest to use are the trimmed means, although the biweight may be the most efficient overall (Gross 1976).

In many cases a reasonable approach is to use both a Welch t test and an efficient Welch t-like test. Only if they disagree is there a problem requiring a closer look at the data. Then the focus might well be on any skewness that might cause the tests to be testing different things: if so, the investigator needs to decide what kind of change is of concern.

Discussion

A main point of this paper is that there is no panacea for the difficult and messy technical problems in the analysis of data from assessments or unreplicated experiments using the BACIP design. Statistical analyses must be based on plausible models, themselves based on *a priori* empirical and theoretical arguments and checked by formal and informal methods.

In particular, randomization tests are likely to be invalid in assessment if sample sizes are unequal, because a crucial assumption, equal variances of the Before and After deviations, is likely to be violated. The Welch t test is more likely to be valid, because it does not require this assumption, and violation of its Normality assumption is not likely to be important to its validity. However, both tests also require the assumptions of independence and additivity.

We have concentrated on randomization t tests, but similar comments apply to virtually all "distribution free" and nonparametric tests. They require the additivity and independence assumptions and, contrary to frequent suggestions (e.g., Carpenter 1990, Jassby and Powell 1990), are less likely to be valid than are modifications of classical parametric tests, when distributions vary over time and sample sizes are unequal, as must be expected for assessment data.

For the remainder of this paper, we turn from the validity and efficiency of tests of "no effect" to the more important, if less technical, question of their proper role.

The "P value" is the probability that data indicating an effect as strongly as our data do, or more so, would arise by chance if in fact there was no effect. Reckhow (1990) asserts that it is often misinterpreted as the probability that there is no effect, and advocates direct calculation of this probability by Bayesian methods. We disagree.

First, the prudent solution to misinterpretation of classical P values is improved explication rather than dumping the methods.

Second, Bayesian conclusions depend on subjective prior probabilities, which are likely to vary widely, especially in adversarial situations; there is a risk that debates about effects will focus less on the data and more on the credentials of the "experts" whose priors are invoked. For example, Reckhow (1990) claims that P values are misleading because Bayesian calculations of the probability of no effect by Berger and Sellke (1987) are usually much larger. But these calculations are based on a prior probability of ≈ 0.5 that there is indeed no effect. In most assessment problems we would regard this prior probability as quite unrealistic: there is almost certainly *some* effect, so the prior probability of no effect should be close to 0; the Bayesian posterior probability of no effect could then easily be smaller than the P value.

Third, and most important: neither a P value nor a Bayesian posterior probability, for a null hypothesis that is inherently implausible, is adequate for such purposes as making decisions about ending or mitigating the impact, resolving legal disputes, designing future power plants or sewage outfalls, managing ecosystems, or studying the biological mechanisms involved (e.g., National Research Council 1990:76). The important questions are how large the effects are, and whether they matter. The main statistical tasks are estimating effect sizes and estimating the precision of these estimates, not hypothesis testing.

For this, there is a standard classical format, confidence intervals. There are Bayesian alternatives, but the disagreement between the two is usually minor for large or moderate sample sizes (Pratt 1965), provided that the prior distribution does not have a sharp peak. In assessments, where there are usually many interacting species, environmental parameters and physiological processes, many of them poorly understood, we would expect honest prior distributions to be quite diffuse.

Any test can be used to form a confidence interval for the size of the effect: the confidence interval is the set of values, δ, for which the null hypothesis "the change in the difference of the means is δ" is accepted. For many parametric tests, this interval is as easily calculated as the test itself. Randomization tests are much harder to convert, although efficient algorithms exist for some special cases (Pagano and Tritchler 1983, Tritchler 1984).

But not all parametric tests will lead to useful estimates in the assessment problem. If a transformation is needed for additivity, the test will concern the mean difference of transformed data; the ecological significance of a change in this mean may be obscure. In some cases there may be no suitable transformation, e.g., if the "Control" population density is greater than the Impact density in winter but smaller in summer, no monotone transformation can achieve additivity.

Perhaps most important, real perturbation effects might not be constant, even if we have the correct transformation. They may vary seasonally or in response to other conditions. For example, the cooling water system of the San Onofre Nuclear Generating Station may reduce irradiance (and gametophyte survival) over the San Onofre Kelp bed when the current flows South, but increase it when the current flows North (Murdoch et al. 1989).

One way to deal with these problems is to think of the "Control" density and other variables (e.g., season, current direction, water temperature, etc.) as predictors. Using regression methods on the Before data, we could estimate the function that best predicts the Impact area density from these predictors. The perturbation effect could be estimated as the difference between this function and the corresponding function obtained from the After data. This approach allows for effects that vary with environmental conditions, includes quantitative estimates of uncertainty (via confidence bands), and is conducive to graphical presentation, which many audiences may find easier to understand. At least one successful example of this approach already exists (Mathur et al. 1980).

This would not usually be a "clean" approach. It would involve regressions based on guessed functional forms, which would be checked, in part, by formal statistical tests, themselves often approximate. Few, if any, of the confidence intervals could be regarded as exact. Increasing exactness, e.g., by incorporating the uncertainty over functional form into the confidence interval, would be a difficult task, requiring some arbitrary judgments, and probably of little use to readers.

But restricting consideration to questions that allow formally exact answers (or appear to), such as overall tests for an effect, risks losing the information of most value: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tukey 1962).

ACKNOWLEDGMENTS

We thank Bill Murdoch, Bill Lenarz, Alec MacCall, Russ Schmitt, and Keith Parker for helpful discussion and/or comments on earlier versions of this manuscript. This work was supported in part by NSF grant BSR8905867, the Minerals Management Service, United States Department of the Interior, under MMS Agreement Number 14-35-0001-30471 (The Southern California Educational Initiative); the University of California Coastal Toxicology Program; the Ecology Research Division Office of Health and Environmental Research, United States Department of Energy, grant DE-FG03-89-ER60885, and the Marine Review Committee. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the United States Government or of the Marine Review Committee.

LITERATURE CITED

Aass, P. 1972. Age determination and year-class fluctuations of cisco, *Coregonus albula* L., in the Mjosa hydroelectric reservoir, Norway. Report of the Institute of Freshwater Research Drottningholm **52**:5–22.

Andrews, D. F. 1971. A note on the selection of data transformations. Biometrika **58**:249–254.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. 1972. Robust estimates of location. Princeton University Press, Princeton, New Jersey, USA.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: the irreconcilability of P-values and evidence. Journal of the American Statistical Association **82**:112–122.

Box, G. E. P. 1954. Some theorems on quadratic forms applied to the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics **25**:484–498.

———. 1980. Sampling and Bayes inference in scientific modelling and robustness (with discussion). Journal of the Royal Statistical Society **A143**:383–430.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society Series **B26**:211–252.

Box, G. E. P., and G. C. Tiao. 1965. A change in level of a non-stationary series. Biometrika **52**:1509–1526.

Box, G. E. P., and G. C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association **70**: 70–79.

Carpenter, S. R. 1989. Replication and treatment strength in whole-lake experiments. Ecology **70**:453–463.

———. 1990. Large-scale perturbations: opportunities for innovation. Ecology **71**:2038–2043.

Carpenter, S. R., T. M. Frost, D. Heisey, and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology **70**: 1142–1152.

Carpenter, S. R., J. F. Kitchell, and J. R. Hodgson. 1987. Cascading trophic interactions and lake productivity. BioScience **35**:634–639.

Carroll, R. J., and D. Ruppert. 1981. On prediction and the power transformation family. Biometrika **68**:609–616.

Carroll, R. J., and D. Ruppert. 1984. Comment on Hinkley and Runger 1984. Journal of the American Statistical Association **79**:312–313.

Cressie, N. A. C. 1982. Playing safe with misweighted means. Journal of the American Statistical Association **77**:754–759.

Cressie, N. A. C., and H. J. Whitford. 1986. How to use the two sample *t* test. Biometrical Journal **28**:131–148.

Durbin, J., and G. S. Watson. 1971. Testing for serial correlation in least squares regression. III. Biometrika **58**:1–19.

Efron, B. 1969. Student's *t* test under symmetry conditions. Journal of the American Statistical Association **64**:1278–1302.

Feller, W. 1966. An introduction to probability theory and its applications. Volume II. Wiley, New York, New York, USA.

Gans, D. J. 1981. Use of a preliminary test in comparing two sample means. Communications in Statistics, Simulation and Computation **B10**:163–174.

Glass, G. V., P. D. Peckham, and J. R. Saunders. 1972. Consequences of failure to meet the assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research **42**:237–298.

Gross, A. M. 1976. Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association **71**:409–416.

Hamrin, S. F., and L. Persson. 1986. Asymmetrical competition between age classes as a factor causing population

oscillations in an obligate planktivorous fish species. Oikos **47**:223–232.

Hinkley, D. V., and G. Runger. 1984. The analysis of transformed data (with discussion). Journal of the American Statistical Association **79**:302–320.

Hoeffding, W. 1952. The large sample power of tests based on permutations of observations. Annals of Mathematical Statistics **23**:169–192.

Hurlbert, S. J. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**: 187–211.

Jassby, A. D., and T. M. Powell. 1990. Detecting change in ecological time series. Ecology **71**:2044–2052.

Jones, R. H. 1980. Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics **22**:389–395.

———. 1981. Fitting a continuous time autoregression to discrete data: applied time series analysis II. Pages 651–682 *in* D. F. Finley, editor. Academic Press, New York, New York, USA.

Kafadar, K. 1982. Using biweight m-estimates in the two sample problem. Part 1: symmetric populations. Communication in Statistics, Theoretical Methods **11**:1883–1901.

Ljung, G. M., and G. E. P. Box. 1978. On a measure of lack of fit in time series models. Biometrika **65**:297–304.

Mathur, D., T. W. Robbins, and E. J. Purdy. 1980. Assessment of thermal discharges on zooplankton in Conowingo Pond, Pennsylvania. Canadian Journal of Fisheries and Aquatic Science **37**:937–944.

McDowall, S. P., R. McCleary, E. E. Meidinger, and R. A. Hay. 1980. Interrupted time series analysis. Sage Publications, Beverly Hills, California, USA.

Mittelbach, G. G. 1988. Competition among refuging sunfishes and effects of fish density on littoral zone invertebrates. Ecology **69**:614–623.

Mittelbach, G. G., C. W. Osenberg, and M. A. Leibold. 1988. Trophic relations and ontogenetic niche shifts in aquatic ecosystems. Pages 219–235 *in* B. Ebenman and L. Persson, editors. Size-structured populations. Springer-Verlag, Berlin, Germany.

Murdoch, W. W., B. Mechalas, and R. C. Fay. 1989. Final report of the Marine Review Committee to the California Coastal Commission on the effects of the San Onofre Nuclear Generating Station on the Marine Environment. Available from the California Coastal Commission, San Francisco, California, USA.

Murphy, B. P. 1976. Comparison of some two sample means tests by simulation. Communications in Statistics, Simulation and Computation **B5**:23–32.

National Research Council. 1990. Managing troubled waters. National Academy Press, Washington, D.C., USA.

Osenberg, C. W., and G. G. Mittelbach. 1989. Effects of body size on the predator–prey interaction between pumpkinseed sunfish and gastropods. Ecological Monographs **59**: 405–432.

Osenberg, C. W., E. E. Werner, G. G. Mittelbach, and D. J. Hall. 1988. Growth patterns in bluegill (*Lepomis macrochirus*) and pumpkinseed (*L. gibbosus*) sunfish: environmental variation and the importance of ontogenetic niche shifts. Canadian Journal of Fisheries and Aquatic Sciences **45**:17–26.

Pagano, M., and D. Tritchler. 1983. On obtaining permutation distributions in polynomial time. Journal of the American Statistical Association **78**:435–440.

Posten, H. 1978. The robustness of the two-sample t-test over the Pearson system. Journal of Statistical Computation and Simulation **6**:295–311.

———. 1979. The robustness of the one-sample t-test over the Pearson system. Journal of Statistical Computation and Simulation **9**:133–149.

Pratt, J. W. 1965. Bayesian interpretation of standard inference statements (with Discussion). Journal of the Royal Statistical Society **B27**:169–203.

Pratt, J. W., and J. D. Gibbons. 1981. Concepts of nonparametric theory. Springer-Verlag, New York.

Reckhow, K. H. 1990. Bayesian inference in non-replicated ecological studies. Ecology **71**:2053–2059.

Romano, J. P. 1990. On the behavior of randomization tests without a group invariance assumption. Journal of the American Statistical Association **85**:686–692.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical methods. Seventh edition. Iowa State University Press, Ames, Iowa, USA.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology **67**:929–940.

Stigler, S. M. 1976. The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. Journal of the American Statistical Association **71**:956–960.

Tan, W. Y. 1982. Sampling distributions and robustness of t, F and variance ratio in two samples and ANOVA models with respect to departure from Normality. Communications in Statistics **A11**:2485–2511.

Tiao, G. C., G. E. P. Box, and W. J. Hamming. 1975. Analysis of Los Angeles photochemical smog data: a statistical overview. Journal of the Air Pollution Control Association **25**:260–268.

Tiku, M. L. 1980. Robustness of MML estimators based on censored samples and robust test statistics. Journal of Statistical Planning and Inference **4**:123–143.

Tiku, M. L., and M. Singh. 1982. Robust tests for means when population variances are unequal. Communications in Statistics **A10**:2057–2071.

Townsend, C. R. 1989. Population cycles in freshwater fish. Journal of Fish Biology **35** (supplement A):125–131.

Tritchler, D. 1984. On inverting permutation tests. Journal of the American Statistical Association **79**:200–207.

Tukey, J. W. 1949. One degree of freedom for non-additivity. Biometrics **5**:232–242.

———. 1962. The future of data analysis. Annals of Mathematical Statistics **33**:1–67.

Yuen, K. K. 1974. The two-sample trimmed t for unequal population variances. Biometrika **61**:165–170.

Yuen, K. K., and W. J. Dixon. 1973. Approximate behavior and performance of the two sample trimmed t. Biometrika **60**:369–374.

V. Statistical power: the design and application of BACIPS studies.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. *Ecological Applications* **4**:16-30.

DETECTION OF ENVIRONMENTAL IMPACTS: NATURAL VARIABILITY, EFFECT SIZE, AND POWER ANALYSIS[1]

CRAIG W. OSENBERG
Department of Integrative Biology, University of California, Berkeley, California 94720 USA, and Coastal Research Center, Marine Science Institute, University of California, Santa Barbara, California 93106 USA

RUSSELL J. SCHMITT AND SALLY J. HOLBROOK
Coastal Research Center, Marine Science Institute and Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

KHALIL E. ABU-SABA AND A. RUSSELL FLEGAL
Institute of Marine Sciences, University of California, Santa Cruz, California 95064 USA

Abstract. The power of any test of an environmental impact is simultaneously constrained by (1) the variability of the data, (2) the magnitude of the putative impact, and (3) the number of independent sampling events. In the context of the Before–After-Control–Impact design with Paired sampling (BACIP), the variability of interest is the temporal variation in the estimated differences in a parameter (e.g., population density) between two unperturbed sites. The challenges in designing a BACIP study are to choose appropriate parameters to measure and to determine the adequate number and timing of sampling events. Two types of studies that are commonly conducted can provide useful information in designing a BACIP study. These are (1) long-term studies that provide estimates of the natural temporal and spatial variability of environmental parameters and (2) spatial surveys around already-perturbed areas ("After-only" studies) that can suggest the magnitude of impacts.

Here we use data from a long-term study and an After-only study to illustrate their potential contributions to the design of BACIP studies. The long-term study of parameters sampled at two undisturbed sites yielded estimates of natural temporal variability. Between-site differences in chemical–physical parameters (e.g., elemental concentration) and in individual-based biological parameters (e.g., body size) were quite consistent through time, while differences in population-based parameters (e.g., density) were more variable. Serial correlation in the time series of differences was relatively small and did not appear to vary among the parameter groups. The After-only study yielded estimates of the magnitude of impacts through comparison of sites near and distant from a point-source discharge. The estimated magnitude of effects was greatest for population-based parameters and least for chemical–physical parameters, which tended to balance the statistical power associated with these two parameter groups. Individual-based parameters were intermediate in estimates of effect size. Thus, the ratio of effect size to variability was greatest for individual-based parameters and least for population and chemical–physical parameters.

The results suggest that relatively few of the population and chemical–physical parameters could provide adequate power given the time constraints of most studies. This indicates that greater emphasis on individual-based parameters is needed in field assessments of environmental impacts. It will be critical to develop and test predictive models that link these impacts with effects on populations.

Key words: Before–After-Control–Impact design; environmental impact; environmental monitoring; impact assessment; individual vs. population parameters; pollution; produced water; serial correlation; spatial variability; statistical power; temporal variability.

INTRODUCTION

A principal challenge posed in field assessments of environmental impacts is to isolate the effect of interest from noise introduced by natural spatial and temporal variability. If the size of an impact from a human disturbance is small relative to natural variability, it will be difficult to detect with any degree of confidence. Therefore, it is critical to consider statistical power in planning and interpreting environmental impact assessment studies (Green 1989, Fairweather 1991, Faith et al. 1991, Osenberg et al. 1992*a*, Mapstone, *in press*; see also Peterman 1990, Cooper and Barmuta 1993). Consideration of power can also guide the selection of environmental parameters and sampling intensity. These are important design criteria because time and

[1] Manuscript received 12 June 1992; revised 19 April 1993; accepted 7 May 1993.

financial constraints typically limit the number of parameters that can be measured and the number of samples that can be collected.

Calculation of statistical power, which is the probability of rejecting the null hypothesis of "no effect" when it is false, requires specification of the number of replicates as well as the ratio between the size of an effect and the variability among the replicates (Cohen 1977). Because there are many assessment designs, each of which makes different assumptions about the meaning of "effect," "variability," and "replicate" (Green 1979, Stewart-Oaten et al. 1986, Eberhardt and Thomas 1991, Underwood 1991, 1994, Osenberg et al. 1992*a*), the general assessment design must be specified before power can be discussed unambiguously. In assessing the environmental impacts of a particular anthropogenic activity, we typically require a design that explicitly deals with the lack of spatial replication and randomization (e.g., nuclear power plants are not replicated and placed at random sites along the United States coastline: Stewart-Oaten et al. 1986). The Before–After-Control–Impact design with Paired sampling (BACIP: Stewart-Oaten et al. 1986, Schroeter et al. 1993, Stewart-Oaten, *in press*; see also Campbell and Stanley 1966, Eberhardt 1976, Skalski and McKenzie 1982, Bernstein and Zalinksi 1983, Carpenter et al. 1989) meets this criterion, and is the focus of our analyses and discussion.

In its simplest formulation, BACIP requires simultaneous (Paired) sampling several times Before and After the perturbation at a Control and an Impact site. The measure of interest is the difference (hereafter referred to as "delta," Δ) in a parameter value (in its raw or transformed state) between the Control and Impact sites as assessed on each sampling date (e.g., $\Delta_{Pi} = \log(C_{Pi}) - \log(I_{Pi})$, where C_{Pi} and I_{Pi} are estimates of the parameter at the Control and Impact sites on the i^{th} date of the period P: i.e., Before or After). The average delta in the Before period is an estimate of the average spatial difference between the two sites, which provides an estimate of the expected delta that should exist in the After period in the absence of an environmental impact (i.e., the null hypothesis). The difference between the average Before and After deltas ($\Delta_{B\cdot} - \Delta_{A\cdot}$) provides a measure of the magnitude of the environmental impact. Confidence in this estimate is determined by the variation in deltas (among sampling dates within a period, S_Δ), as well as the number of sampling dates (i.e., replicates) in each of the Before and After periods ($n_B + n_A = n$). For the purposes of this study, we define

$$\text{Effect size} = \Delta_{B\cdot} - \Delta_{A\cdot}, \quad (1)$$

$$\text{Variability} = S_\Delta = [\Sigma (\Delta_{Pi} - \Delta_{P\cdot})^2]^{1/2}/(n_P - 1)^{1/2}, \quad (2)$$

$$\text{Standardized effect size} = |\Delta_{B\cdot} - \Delta_{A\cdot}|/(2 \times S_\Delta). \quad (3)$$

We assume for convenience that variability (S_Δ), as well as sample size (n_P), are equal in the Before and After periods (but see Stewart-Oaten et al. 1992). Note that the standardized effect size (Eq. 3), which consists of two components (defined by Eqs. 1 and 2) expresses the effect size in standard deviation units and enters directly into conventional calculations of power (Cohen 1977). We double the standard deviation of deltas (S_Δ) in the denominator of Eq. 3 based on the assumption that the resulting test will be two-tailed (Gill 1978).

Unlike other designs, the variability of interest, S_Δ, is not a simple measure of within-site sampling variability. Rather, it is a measure of the actual temporal variation in deltas, as well as within-site sampling error (which contributes to error in estimating the actual delta on any date). Fig. 1 illustrates how this variability of deltas can be altered without any change in the average temporal variability of a parameter (e.g., density), or in the amount of within-site sampling error. The critical feature in determining the variability among deltas is the extent to which estimates of parameters at the two sites track one another though time; Magnuson et al. (1990) refer to this as "temporal coherence."

To aid in the planning of a BACIP study, it would be helpful to find previous BACIP studies conducted in a comparable situation (e.g., similar perturbation in a similar environment) and review the results for variability and effect size. This would permit estimation of the number of sampling dates needed to achieve a given level of power (e.g., Bernstein and Zalinski 1983) or a given amount of confidence in estimates of the effect size (e.g., Bence et al., *in press*, Stewart-Oaten, *in press*). For example, parameters with large standardized effect size (i.e., relatively large effect size and small variability) will yield more powerful assessments with fewer sampling events than parameters with low standardized effect size. Obtaining an adequate number of sampling events in the Before period is crucial in a BACIP assessment, since once the perturbation begins it is no longer possible to obtain additional Before samples. Unfortunately, there are few existing BACIP studies that permit this type of analysis.

In the absence of this information, other data could be used to guide the design of BACIP studies. Two types of non-BACIP studies are more common and can offer insight. The first are long-term studies that document natural spatial and temporal variability, and therefore can provide estimates of S_Δ (Eq. 2). The second are "After-only" studies that assess impacts using a post-impact survey of sites that vary in proximity to the perturbation. After-only studies are a common type of field assessment approach, but they confound effects

TABLE 1. List of the types of parameters used to explore natural temporal variability in deltas—the differences in parameter values between the Control and Impact sites (from the long-term study)—and to obtain estimates of effect size from an existing perturbation (from the "After-only" study).

Parameter type*	Source: Long-term study (Variability)	Source: After-only study (Effect size)
Chemical–Physical		
Water temperature (no. depths)	2	2
Seston characteristics	3	0
Sediment quality	2	2
Sediment elements	11	9
Water column elements	12	8
Individual-based		
Field collections		
Urchin size and condition	5	0
Cumacean body size	2	0
Transplants		
Mussel performance	(10)†	12
Abalone performance	0	4
Population-based (no. of taxa)		
Band transects	6	0
Infaunal cores	11	10
Quadrats	1	0
Emergence traps	4	0
Re-entry traps	3	0

* For each parameter type, we give the number of parameters quantified at each site (e.g., for infaunal density, 11 taxonomic groups yielded sufficient data for analysis in the long-term study). Details on parameters are given in the Methods section.

† The 10 estimates of variability for mussel performance, in parentheses, were collected as part of the After-only study but analyzed in the same manner as data from the long-term study.

of the perturbation with natural spatial variability. Still, After-only studies can suggest the size of effects that might occur in response to a particular perturbation (Eq. 1).

In this paper we illustrate how information from long-term studies and After-only studies can be combined to help plan BACIP studies. We show how this information can be used to guide the selection of parameters and determine sampling schedules given constraints of time and funding. Our presentation consists of four analytical steps: (1) estimation of temporal variability of deltas using results from a long-term study; (2) estimation of the likely magnitude of impacts using results from an After-only study; (3) determination of the number of sampling dates required to detect the estimated impact given the background variability (at a specified level of power); and (4) exploration of serial correlation, using the long-term data set, to assess the time necessary to achieve the required number of independent sampling dates. We contrast results for chemical–physical (e.g., chemical concentrations, sediment characteristics), individual-based biological (e.g., body size, growth), and population-based biological (e.g., density) parameters, and conclude there is a critical need to increase the use of individual-based parameters in field studies of environmental impacts.

METHODS

Background

To help guide the planning of a BACIP study of a particular planned intervention, it would be best to examine results of several preexisting BACIP studies that examined impacts on many parameters in response to the same intervention in identical environments. Of course, such studies do not (and cannot) exist, but the congruence between this ideal and the realized match serves as a guide to the potential accuracy of the general guidelines that emerge.

The first step in this process is to define the intervention. To illustrate our approach, we focus on the nearshore discharge of an aqueous waste called "produced water." Produced water is a complex wastewater generated from the production of oil and contains a variety of petroleum hydrocarbons, heavy metals, and other potential pollutants (Middleditch 1984, Higashi et al. 1992). Although concerns have been raised about possible environmental effects of produced water in marine environments (Neff 1987, Neff et al. 1987, Osenberg et al. 1992*b*, Raimondi and Schmitt 1992), there have been no field assessments with sufficient Before data to allow separation of impacts from other sources of spatial and temporal variability (Carney 1987; also see Underwood 1991, Osenberg et al. 1992*a*).

We explore results from a long-term study of natural spatial and temporal variability and an "After-only" study to substitute for the absence of existing BACIP studies. The two studies were both conducted in nearshore habitats along the coast of Santa Barbara County in southern California. The benthic environments are both dominated by soft-bottom habitats, and the studies used many of the same methods and quantified many of the same parameters. In each study, parameters had been selected based upon their perceived relevance to the impacts of produced water (e.g., Boesch and Rabalais 1987). (Because the long-term study is actually part of the "Before" sampling of a BACIP study of produced-water impacts, even the parameters examined in this study were selected with respect to produced-water discharge.) However, these parameters, which include chemical, physical, and biological characteristics (Table 1), are commonly measured in field assessments of other impacts in marine environments. We next review the two studies, the methods that were used, and the parameters that were measured.

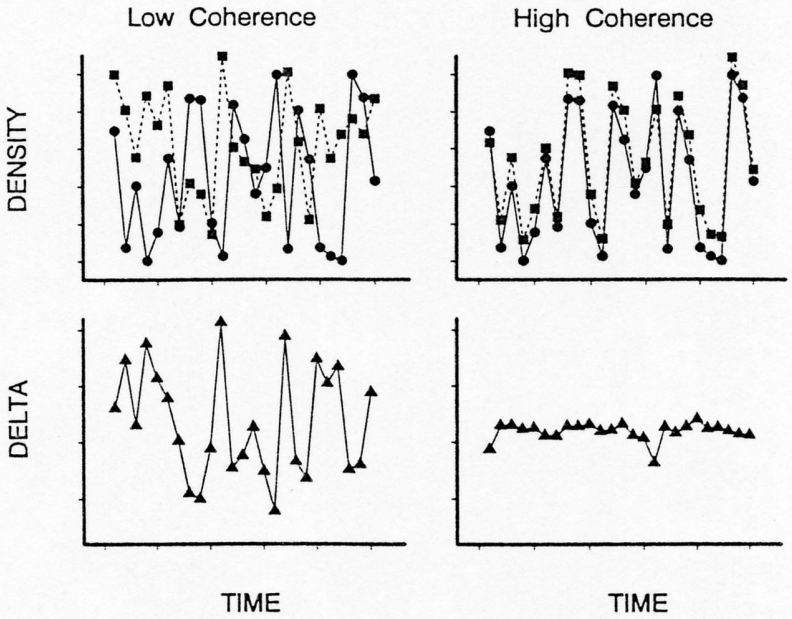

FIG. 1. Patterns of spatial and temporal variation in population densities that lead to high and low variation in deltas. (Δ = difference in parameter values between the Control and Impact sites.) Simulated data (top panels) are from two pairs of sites. In both panels temporal variation in density (at a site) and the average difference between the sites are similar. The panels differ in the degree to which the estimated densities at the paired sites track one another through time. On the left, poor tracking (i.e., low coherence: Magnuson et al. 1990) leads to a low correlation between densities at the two sites ($r = -0.25$), while on the right, good tracking (i.e., high coherence) leads to a stronger correlation in densities ($r = 0.98$). The bottom graphs show the resulting differences in density (deltas). Low temporal coherence in densities (or any other parameter of interest) leads to high variability in deltas, while high coherence leads to low variability in deltas.

Natural variability assessed from long-term study

The two sites that comprise the long-term study are located ≈1.6 km apart offshore of Gaviota, California (≈34°27′29″ N, 120°12′43″ W) at a water depth of ≈27 m. Various biological and chemical–physical parameters (Table 1) were sampled at the sites for periods ranging from 1.5 to just over 3 yr beginning in February 1988. For a given sampling date a single value was obtained for each parameter at each site, and a delta was calculated as the difference between the log-transformed values at the two sites for that i^{th} date:

$$\Delta_i = \log(X_{1i}) - \log(X_{2i}), \quad (4)$$

where X_{1i} and X_{2i} are the values of parameter X at each of the two sites (1 and 2) on the i^{th} date. Original parameter values were log-transformed to better satisfy assumptions of additivity required by BACIP (Stewart-Oaten et al. 1986) and to facilitate comparison of deltas for parameters measured in different units (the transformed deltas are unitless). For each parameter, variability was quantified as the standard deviation of the deltas (S_Δ) calculated over all available sampling dates (Eq. 2).

Population-based parameters.—Densities of infaunal organisms were estimated ≈8 times per year. On each sampling date 12 cores (each 78 cm² × 10 cm deep) were collected. Samples were preserved in a 10% buffered formalin solution and sieved through a 0.5-mm mesh sieve. Organisms were identified and counted from at least four of these cores per site per sampling date. Because this community is extremely speciose, with many species represented by only a few organisms or by zero counts on particular dates, and because zeros can cause difficulties in BACIP analyses (Stewart-Oaten et al. 1986), infaunal organisms were grouped into broad taxonomic units, such as families and classes (see discussions on aggregation in Herman and Heip [1988], Warwick [1988], and Frost et al. [1992]).

Numbers of infaunal organisms that migrated from the sediments into the overlying water (i.e., demersal zooplankton) were estimated using two emergence funnel traps (each covering a bottom area of 0.23 m²) and three reentry traps (each 0.05 m² in area), which were deployed at both sites ≈8 times per year (for more detail on trap designs and function, see Alldredge and King 1980, Stretch 1983). Traps were set out for a 24-h period. Following retrieval, contents were preserved, sieved through a 0.5-mm mesh sieve, and organisms

were identified and counted as with the infaunal cores (Table 1).

Densities of larger epifaunal and demersal organisms (e.g., fish, sea stars, tube anemones) were estimated visually along band transects by divers. Two band transects (each 40 m × 1 m) were established along the 27-m isobath at both sites on each sampling date, and all large organisms within the transect were counted. Most were identified to species, although we grouped many of them into larger taxonomic units for these analyses. Due to their greater maximum density, white sea urchins (*Lytechinus anamesus*) were counted in five non-permanent quadrats, each 1 m^2 in area, at both sites on all dates. Densities of urchins and other epifaunal and demersal organisms were estimated 8–12 times per year.

Individual-based parameters.—The size (length of metasome) of two cumacean species was measured from samples obtained from the emergence traps. Other individual-based parameters (Table 1), including average test diameter, gonad mass, somatic tissue mass, and gonadal/somatic index, were calculated from samples of the white sea urchin, *Lytechinus anamesus*. The average condition of individual urchins of a given size was estimated by calculating adjusted means for each site and date based on ANCOVA using each collection as a group, log(test diameter) as the covariate, and log(tissue mass) as the response parameter. Urchins were sampled for these analyses 11 times during the study. As part of the After-only study, we also obtained estimates of variability for several other individual-based parameters derived from study of the mussel *Mytilus californianus* (see below: *Combining results on effect size and natural variability*).

Chemical–physical parameters.—Chemical and physical parameters were examined that were thought to be indicative of the future plume's chemistry (e.g., elevated levels of certain heavy metals) or of the discharge's physical effects (e.g., altered sediment traits due to scouring of substrate or altered sedimentation rates and temperature due to local oceanographic effects) (Table 1). Seston flux was estimated by particulate accumulation in two sediment traps (5.1 cm in diameter) that were filled with a mixture of seawater, formalin, and salt; the dense preservative remained in the sediment traps during the deployment and had an initial salinity of ≈65 g/L and a formalin concentration of 5%. Sediment traps were deployed ≈3 m above the sediments and retrieved by divers after 3–7 d. Traps were deployed ≈8 times per year. Prior to analysis, large invertebrates were removed (aided by a dissecting microscope), following which the dry mass and ash free dry mass (AFDM) of the particles were determined. Sedimentation rate was calculated as the mass of material (on a dry-mass or AFDM basis) per square centimetre per day. The percentage of organic matter in the seston was estimated as the ratio of AFDM to dry mass.

Sediment grain size and percentage of organic matter were characterized from two sediment cores (20.3 cm^2/core, 5 cm deep) collected from both sites ≈8 times per year. Sediment organic matter (SOM) was estimated based on combustion (for 4 h at 450°C) of subsamples from one core. The fine sediment fraction (percentage) was estimated from the other core as the percentage (by dry mass) of the sample that passed through a 0.063-mm mesh sieve.

Water temperature was recorded approximately monthly at 3 m depth intervals. Here we use data for the 6 m and 21 m depths.

Surficial sediments (approximately the top 1 cm) were collected 4 times per year for analyses of trace and bulk elements. Three samples were collected at each site in acid-cleaned polyethylene containers by divers using trace metal clean-sampling techniques. Any overlying water was decanted and samples were frozen. Sediments were later thawed and extractions performed by leaching 2 g sediment in 20 mL of 0.5 mol/L HCl for 24 h. The leachate was then filtered through a 0.45-μm mesh teflon filter using procedures reported previously (Oakden et al. 1984). This extraction is considered to be relatively selective for the biologically available concentrations of many metals, such as Pb, Cu, and Ag (Luoma et al. 1991). Leachates were analyzed for bulk elements (Al, Ca, Fe, Mg, Mn, P) and trace elements (Ba and Zn) by inductively coupled plasma–atomic emission spectrometry (ICP–AES). Other trace element (Cr, Cd, and Pb) concentrations were determined by graphite furnace atomic absorption spectrometry (GFAAS). Environment Canada reference sediments (BCSS-1, MESS-1, PACS-1) were analyzed concurrently to quantify the extraction efficiency for each element. All analyses were normalized to sediment dry mass.

Unfiltered water samples were collected 2 times per year from each site at two depths (surface and 21 m). The samples were extracted using the ammonium 1-pyrrolidinedithiocarbamate/diethylammonium diethyldithiocarbamate (APDC/DDC) extraction method described by Bruland et al. (1985). Trace element concentrations (Ag, Cd, Co, Cu, Fe, Ni, Pb, Zn) were measured by GFAAS. Procedural blanks were measured in each sample set. Each set of samples was analyzed in duplicate after a series of intercalibrations with Environment Canada reference seawater (CASS-1). These analyses were conducted concurrently with analyses of sea water from San Francisco Bay, and details of the procedural blanks and intercalibrations are provided in a report on those data (Flegal et al. 1991).

Effect size estimated from an After-only study

The After-only study was conducted at a produced-water outfall located near Carpinteria, California (34°23′10″ N, 119°30′31″ W) that was the subject of recent investigations of potential environmental im-

pacts (Higashi et al. 1992, Krause et al. 1992, Osenberg et al. 1992*b*, Raimondi and Schmitt 1992). The Carpinteria sites are ≈50 km from the Gaviota sites. Although the two locations (Carpinteria and Gaviota) are both open-coast, soft-bottom environments in the Santa Barbara Channel and have many species in common, the bottom depths sampled differed between the Carpinteria (11 m) and Gaviota (27 m) sites.

An intensive spatial survey of infauna was conducted along the 11 m isobath at the Carpinteria study area in 1990, ≈12 yr after produced water was first discharged at this location (Osenberg et al. 1992*b*). In a single survey, 20 sites were sampled along a spatial gradient from 2 to 1000 m up coast (West) and down coast (East) of the diffusers. Infaunal densities were estimated at each site by collecting eight cores (78 cm^2 per core to a depth of 10 cm). These were processed as described for the long-term study, and a mean density was calculated for each taxon at each of the 20 Carpinteria sites.

All chemical–physical parameters examined as part of the long-term study at Gaviota were also estimated at the Carpinteria sites, except those related to seston quality and deposition and several elements. Methods were identical to those used at Gaviota (described above, see *Natural variability assessed . . .*).

Individual-based biological data were obtained by transplanting individuals of known size and/or age to several of the sites. Mussels (*Mytilus californianus* and *M. edulis*) were transplanted to six sites to determine if proximity to the outfall influenced their individual growth and condition (Osenberg et al. 1992*b*). Forty individuals from a uniform size distribution (range: 20–60 mm shell length) of a mussel species were put into a bag of 1.25-mm mesh oyster netting, and one bag of *M. californianus* and one of *M. edulis* were attached to buoy lines ≈3 m above the sediments. Mussels were retrieved and frozen after 3–4 mo in the field. Final shell length, initial shell length, dry gonadal tissue mass, and somatic tissue mass were then measured for each mussel. Site-specific estimates of average gonadal condition (gonad mass at a given size), somatic condition, total condition, and gonadal–somatic index were obtained by running analyses of covariance (ANCOVA) for each parameter for each mussel species using log(final shell length) as the covariate. Average shell growth and tissue production were estimated using log(initial shell length) as the covariate. Adjusted means were obtained for each parameter at each of the six sites.

Abalone larvae were raised in the laboratory and transplanted in small flow-through cages to 6–8 sites located 5–1000 m from the diffuser (Raimondi and Schmitt 1992). Three measures of per-capita settlement and metamorphosis were derived from transplants that lasted ≈4 d: (1) the proportion of late-stage larvae that successfully settled in the field, (2) the proportion of late-stage larvae that successfully metamorphosed in the field, and (3) the proportion of early-stage larvae that subsequently settled in the laboratory after addition of a chemical inducer (for details, see Raimondi and Schmitt [1992]). An additional measure of individual performance was obtained from a short-term transplant: the proportion of early-stage larvae still swimming after 6 h in the field.

To obtain estimates of the magnitude of impacts due to produced water we calculated means (e.g., of density or performance) for three distance categories: Near (sites <25 m of the diffuser), Far (25–200 m), and Control (>200 m). We then calculated near-field and far-field effect size as the difference between log(Mean Near or Mean Far) and log(Mean Control). This is equivalent to the impact size (expressed in log units) of a BACIP study (Eq. 1) assuming no natural spatial variation between the sites (i.e., $E(\Delta_S) = 0$). While this assumption cannot be tested without Before data, available evidence suggests that natural spatial gradients are small relative to the impacts of produced water (Osenberg et al. 1992*b*, Raimondi and Schmitt 1992).

Combining results on effect size and natural variability

For parameters that were common to both the After-only study and the long-term study, the standardized effect size was calculated as the ratio between the absolute value of the effect size, which was obtained from the long-term study, and twice the standard deviation of deltas, which was obtained from the After-only study (Eq. 3). In some cases, however, the same parameters were not measured in both studies, and other steps were required before proceeding with the power analyses.

For example, there were four chemical–physical parameters that provided estimates of effect size but not variability. All four parameters were elemental concentrations (i.e., Cu in sediments and Co, Ag, and Pb in the water column), so we used the average standard deviation for other elements (in either the sediments or water column) in the calculation of the standardized effect size.

Conversely, there were chemical–physical and population-based parameters that provided estimates of variability but not effect size (i.e., parameters estimated from sediment traps, band transects, emergence traps, reentry traps, and quadrats in addition to several elemental concentrations: Table 1). For these parameters we calculated standardized effect sizes using the average effect size for similar parameters that were measured as part of the After-only study.

Estimating standardized effect sizes for individual-based parameters posed a more difficult analytical problem because the individual-based data from the long-term study were derived from field collections of organisms, whereas the transplants conducted in the After-only study used organisms of known size, or cohorts of known number and age. Therefore, the trans-

plants removed several sources of potential variability present in estimates from the long-term study. Because mussels had been transplanted during four different periods (spread over a total of 14 mo), we were able to obtain estimates of variability for the mussel parameters. The standard deviation of differences between log-transformed parameters measured at the 1000-m and 100-m sites was calculated for ten of the mussel parameters over the four periods. Because the 100-m site is probably influenced slightly by the discharge of produced water (Osenberg et al. 1992*b*, Raimondi and Schmitt 1992), this approach will overestimate S_Δ if there is temporal variation in the effects of produced water.

Standardized effect size was then calculated as explained above using these new estimates of variability for all mussel parameters except tissue production (for which we had only one survey and therefore could not estimate S_Δ, the variation among sampling dates within a period). The mean standard deviation of deltas for the mussel parameters was used to estimate the standardized effect sizes for mussel tissue production and abalone performance parameters, which lacked estimates of S_Δ. The standardized effect sizes for the individual-based parameters derived from the long-term study were calculated using the mean effect sizes based on the mussel and abalone transplants.

For each parameter we estimated the sample size (total number of sampling dates in the Before and After periods) needed to have an 80% chance of detecting ($\alpha = .05$) an impact characterized by the parameter's standardized effect size. All power analyses were based on two-tailed t tests as provided in Gill (1978). The number of sampling dates in the Before and After periods was assumed to be equal.

Serial correlation

The power analyses yield the number of independent sampling events (i.e, dates) needed for a given level of power (e.g., 80%). The time scale over which those samples must be collected will depend on the amount of serial correlation in the time series of deltas for each parameter (Stewart-Oaten et al. 1986). Serial correlation can be directly incorporated into the analyses of BACIP data (Stewart-Oaten et al. 1992), but power is greatest when serial correlation is absent. Therefore, we tried to determine the most intensive sampling schedule that would avoid substantial amounts of serial correlation. By doing so, we could roughly translate the number of independent sampling events into an estimate of the minimum amount of time required by the BACIP study.

Because rigorous analyses of serial correlation require long time series of data, and because the approach we outline here is imprecise to begin with (i.e., extrapolating from two different studies to the design of a future one), we used a simpler approach to provide a general guide to sampling frequency. For each parameter sampled as part of the long-term study, we examined the correlation between the delta measured on one sampling date (Δ_i) and the delta measured on the next date on which sampling for that parameter was conducted (Δ_{i+1}). Only parameters with data from ≥ 8 dates were included in the analyses.

Results

Natural variability assessed from a long-term study

Data from the long-term study revealed that the variation in deltas (i.e., in the difference in parameter values between sites) was lowest for chemical–physical parameters, intermediate for individual-based parameters, and greatest for population-based parameters (Fig. 2). Most (28 of 30) of the chemical–physical parameters exhibited less variation in deltas than did the least variable population-based parameter. Almost all of the population-based parameters (24 of 25) were more variable than the most variable of the 7 individual-based parameters. Within a parameter group, no systematic differences were apparent among data collected using different techniques (e.g., densities based on infaunal cores vs. band transects, or water column elements vs. sediment elements), and there were no apparent trends among the population-based parameters related to the level of taxonomic aggregation (see Frost et al. 1992). All else being equal, these data suggest that chemical–physical parameters will provide more reliable indicators of environmental impacts than population-based parameters due to their smaller variability.

Effect size estimated from After-only study

The After-only study provided estimates of effect sizes, which varied with proximity of the sampled sites to the produced-water diffuser. In general, sizes of effects were correlated ($r = 0.62$, $n = 47$) for sites near to and far from the diffuser (Fig. 3), and the magnitudes of effects consistently were greatest nearer the diffuser. This pattern suggests that impacts diminished with distance away from the disturbance.

Both positive and negative changes in parameter values with distance from the diffuser were observed, and the sign depended on the particular parameter or parameter group examined. For example, concentrations of water-column metals were higher nearer the diffuser, whereas most measures of individual performance were lower. Similarly, some taxa were more abundant closer to the diffuser, while others were less abundant. These two patterns in density probably reflect positive responses to organic enrichment (from oil constituents) and negative responses to toxicants present in pro-

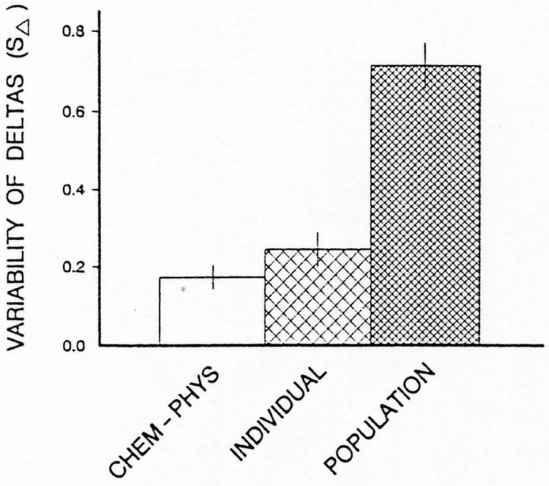

FIG. 2. Temporal variability in estimates of the deltas (S_Δ) for chemical–physical, individual-based, and population-based parameters. Data were derived from the long-term study. For each parameter on each sampling date, a delta was estimated based on the difference between the log-transformed means at two sites (e.g., Log(mean density at Site 1 on date *i*) − Log(mean density at Site 2 on date *i*)). Shown are the standard deviations of deltas (mean ± 1 SE) for parameters in each of the three groups. Means are based on 30, 7, and 25 different parameters for chemical–physical, individual, and population groups, respectively. Here all individual-based data are derived from field collections.

duced water (e.g., Spies and DesMarais 1983, Osenberg et al. 1992*b*; see also Pearson and Rosenberg 1978, Ferris and Ferris 1979).

In evaluating power the crucial factor is the absolute size of the change and not the sign (i.e., a positive or negative response). Although quite variable, the population-based parameters had absolute values of effect sizes that were about twice those for individual-based parameters, and four times larger than effect sizes for chemical–physical parameters (Fig. 4). This pattern was similar for both Near and Far sites ($r = 0.72$, $n = 47$), although the overall magnitude of effects was lower at the Far sites (Fig. 4). For simplicity, we focus on results from the Near sites in the following sections.

Combining results on effect size and natural variability

Estimates of natural variability in individual-based parameters were derived from field collections, whereas those for effect size were obtained from transplants. To make the estimates more comparable, we calculated variability of deltas for individual performance of mussels from four separate transplants in the After-only study. The results show that all 10 indices of mussel performance were relatively invariable over time (S_Δ mean ± 1 SE = 0.080 ± 0.20, range: 0.007–0.220). Indeed, most (70%) of these estimates of mussel performance were less variable than almost all (94%) of the parameters measured in the long-term study.

The results from the long-term study and the After-only study yielded the opposite conclusions about the power associated with different parameter groups. On one hand, the population-based (and individual-based) parameters should be the most powerful due to their larger average effect sizes (Fig. 4), while the chemical–physical (and individual-based) parameters should be more powerful due to their smaller average variability (Fig. 2). Ultimately, the more powerful parameters will be those with the greatest standardized effect size (i.e., signal to noise ratio: Eq. 3). Due to their relatively large effect sizes but low variability, individual-based parameters (particularly those derived from transplants) had larger standardized effect sizes than either the chemical–physical or population-based parameters (Fig. 5). With respect to the individual-based parameters, the transplants yielded standardized effect sizes that were >3 times larger than those derived from field collections.

The standardized effect sizes for both chemical–physical and population-based parameters were low and quite similar (Fig. 5), due to the lower variability associated with chemical–physical parameters (Fig. 2) and the greater effect sizes associated with population-based parameters (Fig. 4). The standardized effect sizes for these two groups of parameters were one-half and one-seventh the magnitude of those for individual-based parameters derived from field collections and transplants, respectively (Fig. 5).

These results indicate that power to detect changes from exposure to produced water should be greatest

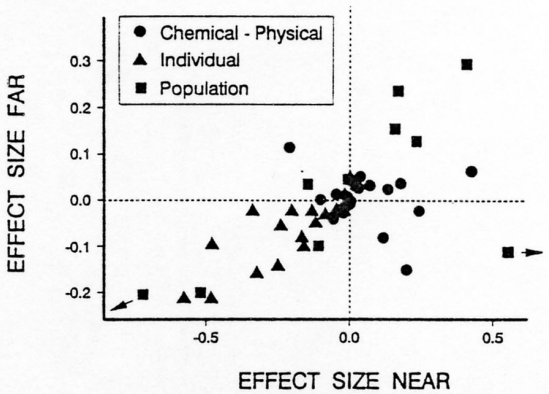

FIG. 3. Effect sizes estimated from sites near and far from an operating produced-water diffuser (an "After-only" study). Positive values indicate larger parameter values near (or far from) the diffuser relative to control sites, while negative values indicate the opposite. The two population-based parameters next to the arrows have effect sizes that are off the scale: (−0.92, −0.85) and (0.917, −0.13).

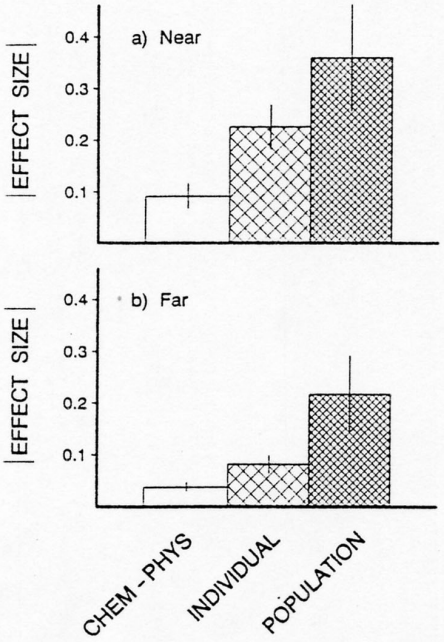

FIG. 4. Absolute effect sizes (mean ± 1 SE) for chemical–physical, individual-based, and population-based parameters based on sites (a) near and (b) far from the diffuser. Sample sizes (number of parameters) were 21, 16, and 10 for the chemical–physical, individual, and population groups, respectively.

for individual-based parameters derived from transplants, and next greatest for individual-based parameters obtained from field collections. For an equivalent number of estimates (i.e., sampling dates), power should be considerably lower for chemical–physical and for population-based parameters. For example, based upon average standardized effect sizes (Fig. 5) and a Type I error rate of .05, the numbers of independent sampling dates needed to achieve power of 80% are ≈4 for individual-based parameters from transplants, 24 for individual-based parameters from field collections, 90 for chemical–physical parameters, and 95 for population-based parameters.

Most individual-based parameters required <20 (and typically <10) sampling dates to achieve 80% power (Fig. 6). Over half of the chemical–physical and population-based parameters required 100 or more sampling dates to reach 80% power (Fig. 6). To provide an idea of how many parameters would have high power for a logistically reasonable number of surveys that would also permit model development and testing (Stewart-Oaten, *in press*), we determined the fraction of parameters in each group with a sufficiently large standardized effect size (>0.52) to yield power of at least 80% with 30 sampling dates ($n_B = n_A = 15$). Using this guideline, 81% (13/16) of individual-based parameters from transplants and 43% (3/7) of those from field collections had power that exceeded 80%. By contrast, only 18% (6/34) of the chemical–physical and 4% (1/26) of the population-based parameters achieved this level of power after 30 surveys.

The preceding analyses were based on effect sizes estimated from sites Near the produced-water diffuser. Repeating the analyses using data from the Far sites yielded similar patterns, although, as expected, the overall power was much lower or the number of sampling dates needed for a given level of power was much higher. For example, the smaller effect sizes (estimated from sites far from the diffuser) resulted in more than half of the parameters in each of the three groups requiring >100 sampling dates to achieve 80% power. Only 26% of the individual-based parameters (all from transplants) required <30 sampling dates, while none of the chemical–physical or population-based parameters achieved the same power with 30 dates.

Serial correlation

Our analyses suggested that impacts on individual-based parameters are the most likely to be detected with a limited number of sampling dates. The analyses assumed that each sampling date provided an independent estimate of the true deltas (i.e., the underlying difference in parameter values between the Control and Impact sites). We examined patterns of serial correlation from the long-term study to gain insight into the

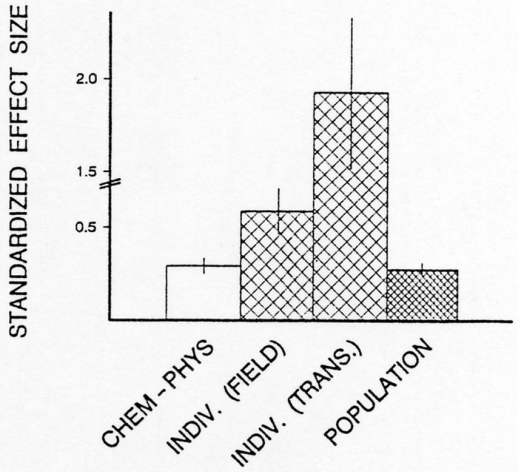

FIG. 5. Standardized effect size (|Effect size|/[2 × S_Δ]) for each parameter group; the measure is the ratio of effect size to twice the standard deviation of delta. Shown are means ± 1 SE, based on 34, 7, 16, and 26 parameters (from left to right). Individual-based parameters are divided into estimates derived from field collections and those derived from transplants of marked individuals or caged cohorts. Note the break in the vertical scale between 0.7 and 1.5.

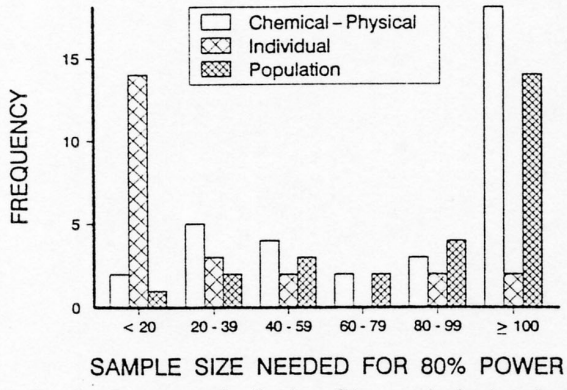

FIG. 6. Frequency distribution of the sample size (number of independent sampling dates) for parameters in each group that is required for 80% power. Power analyses are based on standardized effect sizes (Fig. 5).

frequency with which samples could be collected without grossly violating the assumption of temporal independence. This provided information on the time frame needed to collect series of independent samples.

There were no cases of significant ($P < .05$) negative serial correlation, and only 8% (4 of 50) of the parameters exhibited significant positive serial correlation (e.g., Fig. 7). Of the four parameters with positive serial correlation, two were chemical–physical parameters (seston sedimentation rate and seston percentage organic matter), and two were population-based parameters (densities of sea pens and sea urchins: Fig. 7c and d). None of the individual-based parameters exhibited significant serial correlation.

Serial correlation appeared to arise in the population-based parameters as a result of long-term trends in the deltas (Fig. 7c and d). For example, the white sea urchin (*Lytechinus anamesus*) exhibited strong seasonal migrations, and was present during the winter and spring but absent during the summer and fall. The relative density at the two sites appeared to be set when urchins reappeared in winter; the ranking of the two sites was consistent within a year, but varied greatly among years (Fig. 7c). This suggests that replicates should be collected only once per year, or a yearly average obtained from more frequent collections.

Density of sea pens (*Acanthoptilum* sp. and *Stylatula* sp.) exhibited an even longer term trend (Fig. 7d). One site tended to have a greater density than the other site prior to October 1989, but the reverse was true for all samples collected after this date (Fig. 7d). This could have arisen, for example, by a strong recruitment event in the fall of 1989 at only one of the sites.

Despite these two examples, serial correlation was not a general problem for the various parameters estimated in our long-term study (e.g., Fig. 7a and b). On average, the serial correlation for each of the three parameter groups was only 0.1–0.2 (Fig. 8). Simulations suggest that serial correlation of this order intro-

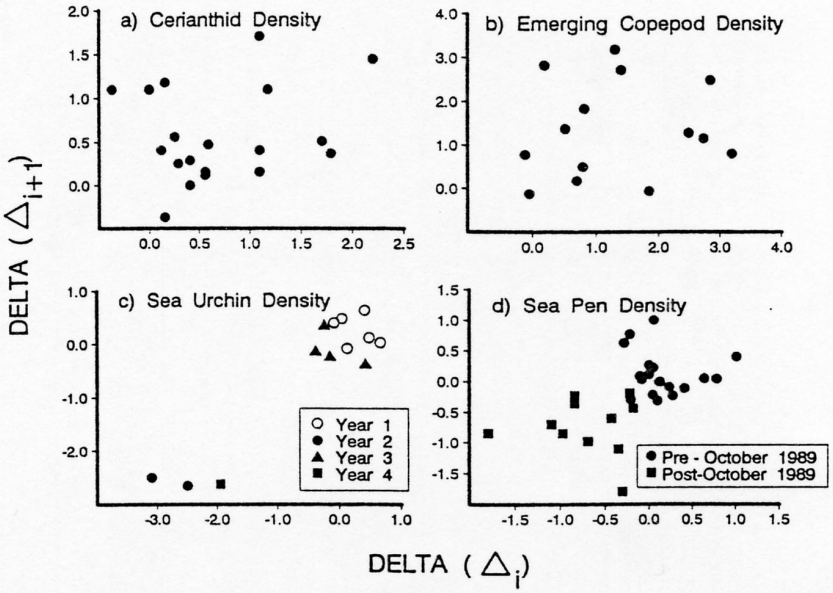

FIG. 7. Patterns of serial correlation in deltas for four population-based parameters. These are the difference in density of: (a) cerianthid (burrowing) anemones (from band-transect estimates); (b) copepods (from emergence traps); (c) white sea urchins (*Lytechinus anamesus*) (from quadrat samples); and (d) sea pen density (from band transects). There is significant serial correlation in (c) and (d), and data are separated into temporal groups to help distinguish the long-term patterns.

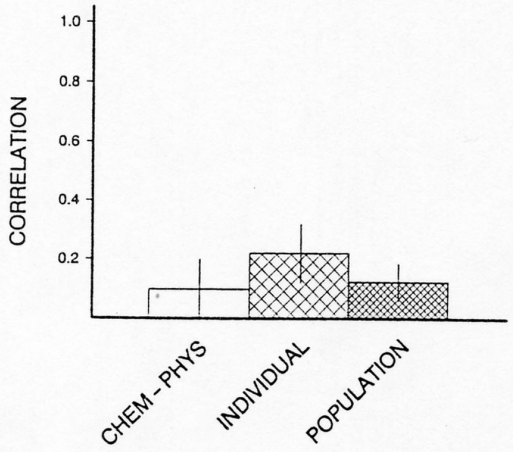

FIG. 8. Degree of serial correlation in deltas for each parameter group. Shown are means ± 1 SE, based on 18, 7, and 25 parameters for chemical–physical, individual, and population groups respectively.

duce only small error into tests of impacts (Carpenter et al. 1989, Stewart-Oaten et al. 1992).

Based on these results, we assumed that sampling could occur every 60 d without yielding substantial amounts of serial correlation. Assuming that six samples are collected per year and the Before and After periods are of equal duration, the estimates of sample size (number of independent sampling events) can be translated into the number of years the assessment study must be conducted. Achieving 80% power would require 16 yr for population-based parameters, 15 yr for chemical–physical parameters, 4 yr for individual-based parameters from field collections, and 1 yr for individual-based parameters from transplants. To achieve 80% power for only a quarter of the parameters in each group, the required study duration is reduced to 11 yr for population-based parameters, 7 yr for chemical–physical parameters, 3 yr for individual-based parameters from field collections, and <1 yr for individual-based parameters from transplants.

DISCUSSION

Because relatively few well-designed studies of planned perturbations have been completed, there is a sparse empirical base to guide the design of future assessment programs (e.g., Carney 1987, Spies 1987, Underwood 1991, Stewart-Oaten, *in press*). Recent discussions have highlighted general design considerations that should be incorporated in Before–After-Control–Impact approaches (e.g., Stewart-Oaten et al. 1986, 1992, Underwood 1994, Stewart-Oaten, *in press*), but these say little about specific considerations regarding sampling frequency and parameter selection. Often a study must be planned in the absence of sufficient preliminary data to properly guide sampling decisions (Stewart-Oaten, *in press*). It is crucial to obtain good estimates of sampling variability and the size of impacts that might arise (or that are deemed ecologically important: Underwood and Peterson 1987, Yoccoz 1991), but this information typically is lacking. In the absence of a BACIP (or analogous) study conducted previously on a similar perturbation in a similar habitat, it is vital that other existing data be used to guide specific design considerations.

Given limitations on time and funding, the selection of parameters and frequency of sampling are especially crucial features of the design process. One of the most acute constraints is the time available to collect data prior to the perturbation. In many situations the Before period probably will be rather abbreviated for a variety of reasons beyond scientific control. Therefore, parameter selection and sampling design should take into account the low numbers of temporal replicates that likely can be collected prior to the commencement of the disturbance (see Stewart-Oaten [*in press*] for discussion of model development based on these data). Key considerations in this regard are the likely variability in the parameter estimate (e.g., delta) and the probable magnitude of response to the disturbance, both of which influence statistical power to detect an effect. Constraints on the number of temporal replicates in the Before period are most likely to hamper detection of impacts on population density and chemical–physical characteristics, and least likely to affect detection of effects on individual performance. Unfortunately these results suggest that many field monitoring programs might be compromised because individual-based parameters rarely are examined (e.g., Carney 1987).

There are, however, compelling reasons to examine population and chemical–physical parameters despite the expected low power. First, chemical and physical properties describe the direct effect of many perturbations, and in many cases impacts could be ameliorated by subsequent intervention (e.g., source reduction, reduced discharge limits). Second, population attributes, such as density, reflect the ecological consequences of the disturbance, and are features of fundamental concern to resource managers and regulatory agencies. In addition, some species receive special regulatory consideration. Another reason is that, while the average power for population or chemical–physical parameters is low, some species or chemical–physical parameters will have greater power than others. The approach described here is equally useful in identifying promising candidates within a parameter group as it is in guiding allocation of effort among groups. Finally, the actual impacts of the new disturbance, of course, cannot be known a priori, and effects on populations and chemical–physical parameters certainly can be much larger (or variation much smaller) than anticipated based on extrapolations from other data sets.

It is useful to consider why the population-based and chemical–physical parameters had low and similar power, because low power arose for different reasons. Population parameters were highly responsive to produced water (i.e., larger impact), but exhibited much greater natural variability. In contrast, the chemical–physical parameters had much lower variability in deltas, but were not greatly altered by the discharge of produced water. It appears that these results generally will hold for other types of point-source disturbances in the marine environment. Many chemical–physical parameters probably are influenced largely by large-scale oceanographic processes that similarly affect nearby sites. For example, certain chemical–physical attributes (e.g., sedimentation rate, water temperature, nutrient flux) are strongly associated with upwelling conditions, which is a region-wide phenomenon (e.g., Landry and Hickey 1989). In these situations, differences in these parameter values between Control and Impact sites (i.e., the deltas) will be similar through time (see also a related discussion in Magnuson et al. [1990], which discussed temporal coherence of chemical–physical and biological parameters in freshwater lakes).

The relatively small response that we observed of chemical–physical parameters to the discharge of produced water is also consistent with recent analyses of the general effect of waste discharges on the distribution of trace elements in coastal waters. For example, massive discharges (10^9 L/d) of wastewaters in the Southern California Bight have had a negligible (<1%) impact on concentration of cadmium in those waters (Sañudo-Wilhelmy and Flegal 1991). Similarly, Schmidt and Reimers (1991) found that, in the Santa Barbara Basin, the fraction of certain metals (Cd, Cu, Ni, Pb) from human sources that is deposited in sediments near municipal outfalls is quite small (<1%) compared to the amount released. In both cases, natural inputs and physical mixing processes appeared to have reduced the contribution from human inputs to a small fraction of the background level. So for chemical–physical parameters the large spatial scale of events that drive natural variation can lead to low variability in deltas, while other natural processes can greatly diminish the signal provided by anthropogenic perturbations.

Population density, by comparison, is known to be highly responsive to local conditions, and can exhibit considerably different temporal patterns among neighboring sites (e.g., Holbrook et al. 1990, Magnuson et al. 1990, Schmitt and Holbrook 1990). The high sensitivity to local conditions potentially can translate into strong local responses to natural phenomena (thus increasing S_λ) as well as anthropogenic perturbations such as wastewater discharges (thus increasing effect size). Within-site sampling error also can contribute to the high variability as benthic populations are notoriously difficult to sample (Vezina 1988, Thrush et al. 1994).

It is important to note that the variability reported here (e.g., Fig. 2) is a measure of the variability (over time) in *estimates* of the *differences* between sites. This variability includes both the true temporal variation in deltas and variation due to sampling error within a site (which adds error to the estimation of delta on any date). The contribution of sampling error will be a function of spatial variability within a site and sampling intensity, and therefore will vary with the within-site sampling design. This suggests that the variation in deltas (S_λ) for population-based parameters could be reduced by more intensive sampling on each date, rather than increasing the number of dates. However, partitioning of observed variation for the long-term data set revealed that the deltas for population-based parameters were more variable due both to sampling error (i.e., high within-site spatial variation) and site-specific temporal variability (i.e., high variation in the actual deltas through time) (C. W. Osenberg, *personal observation*); increasing the sampling intensity within a date would reduce the observed variation (S_λ) by only $\approx$50%. Therefore, even if sampling error were removed (e.g., through more exhaustive sampling), population-based parameters still would be more variable than the chemical–physical or individual-based parameters (see Fig. 2).

Our estimates of S_λ probably are typical because the within-site sampling design of our long-term study is similar to that used in many assessment studies (see Thrush et al. 1994). The costs and benefits of adjusting within-site sampling intensity to achieve greater power can be analyzed (e.g., the importance of within-site accuracy vs. more sampling dates), although with limited resources greater precision ultimately would be accomplished at the cost of fewer sampling dates (which is the unit of replication in a BACIP design).

Difficulty in sampling populations or other parameters within sites not only can affect the variance of the estimate, it also might lead to overestimation of effect sizes from After-only studies (Figs. 3 and 4). This would be true especially for a parameter that is not affected by the perturbation, and thus should have an effect size of zero. Our approach would overestimate this effect by confounding sampling error and any underlying spatial gradient as an effect of the perturbation. If so, the calculated number of surveys (sample size) needed for a given level of power would be underestimated. While this bias will exist for any parameter, our data suggest that, on average, it will be most acute for population-based parameters. Hence, limitations on detecting impacts at the population level may be even more difficult than our analyses suggest.

In contrast to population and chemical–physical parameters, individual-based parameters had relatively high power owing to relatively low levels of variability (Fig. 2) and intermediate effect sizes (Fig. 4). Although 80% power could be achieved for many of the param-

eters we examined with fewer than 10 sampling dates, it is unwise to reduce the sampling intensity below a level at which model development and testing can be performed (Stewart-Oaten, *in press*). Our data also indicate that variability in the deltas for individual-based parameters can be reduced by use of transplants (Fig. 4), which results in increased power (Fig. 5). This presumably occurs because, compared with estimates from field collections, transplants remove noise introduced by individual variation as well as variability between sites over time. For example, size-specific growth rates can be assessed accurately using marked individuals of known size; because size can influence growth and size distributions can vary among sites (e.g., Osenberg et al. 1988), an analysis based on marked individuals is likely to be more powerful than one based on field collections.

It should be noted that several of the transplant-derived parameters for which we had relatively high power are closely related to population-based parameters, which had much lower power. For example, transplants of abalone larvae provided estimates of per capita settlement rates. In the field, natural rates of per capita settlement can be estimated from observed settlement rates and/or larval supply, both of which require estimation of density (e.g., Olson 1985, Keough 1986, Victor 1986, Raimondi 1990). Therefore, these field estimates would have considerable error for the same reasons that population parameters were highly variable. The use of transplants surmounted much of this problem by using cohorts of known size, thereby eliminating much of the variability that plagues the population parameters.

The observation that individual-based parameters may yield more powerful assessments is troubling given the rarity with which they are measured in field assessments. Care must be taken to guard against only considering parameters that yield low probabilities of demonstrable results (e.g., chemical–physical and population attributes), and inclusion of individual-based parameters could greatly increase the sensitivity of assessment studies (Carney 1987, Osenberg et al. 1992*a*; see also Jones et al. 1991). However, the need to investigate individual-based parameters goes far beyond power considerations; it is the individual-based (and demographic) parameters that provide the mechanisms that underlie changes at the population (and therefore community) level. Furthermore, these individual-based parameters provide an explicit connection with detailed laboratory studies that focus on individuals and mechanisms of toxicity. What is needed are more realistic studies of individual-based effects under field conditions combined with both mechanistic laboratory studies and field assessments of population-level consequences. Recent advances with individual-based models (DeAngelis and Gross 1992) provide an explicit framework for making these fundamental linkages among environmental chemistry, physiology, and population ecology (e.g., Hallam et al. 1990). Such models provide a powerful, mechanistic approach to assessing impacts on natural populations and complement the traditional approach of monitoring environmental impacts.

ACKNOWLEDGMENTS

The assistance of Don Canestro is gratefully acknowledged. Also assisting in the field or laboratory were S. Anderson, M. Carr, M. Edwards, D. Forcucci, D. Heilprin, T. Herrlinger, B. Hoffman, T. Kaltenberg, P. Krause, Shi-Yong Lin, A. Martinez, M. Perez, P. Raimondi, D. Reed, D. Steichen, K. Sydel, and V. Vrendenburg. Genine Scelfo performed trace metal analysis of seawater samples, and Bonnie Williamson provided technical and logistical assistance. P. Raimondi provided the data on abalone larvae and A. Stewart-Oaten provided helpful discussions. We also appreciate the critical comments of A. Stewart-Oaten, J. Levinton, and an anonymous reviewer on an earlier draft. This research was funded by the Minerals Management Service, U.S. Department of Interior under MMS Agreement No. 14-35-001-3071, and by the U.C. Coastal Toxicology Program. The views and conclusions in this paper are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

LITERATURE CITED

Alldredge, A. L., and J. M. King. 1980. Effects of moonlight on the temporal migration patterns of demersal zooplankton. Journal of Experimental Marine Biology Ecology **44**: 133–156.

Bence, J. R., A. Stewart-Oaten, and S. C. Schroeter. *In press.* Estimating the size of an effect from a Before-After-Control-Impact-Pairs design: the predictive approach applied to a power plant study. *In* R. J. Schmitt and C. W. Osenberg, editors. Ecological impact assessment: conceptual issues and application in coastal marine habitats. University of California Press, Berkeley, California, USA.

Bernstein, B. B., and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. Journal of Environmental Management **16**:35–43.

Boesch, D. F., and N. N. Rabalais, editors. 1987. Long-term environmental effects of offshore oil and gas development. Elsevier Applied Science, New York, New York, USA.

Bruland, K. W., K. H. Coale, and L. Mart. 1985. Analysis of seawater for dissolved cadmium, copper and lead: an intercomparison of voltametric and atomic absorption methods. Marine Chemistry **17**:285–300.

Campbell, D. T., and J. C. Stanley. 1966. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, Illinois, USA.

Carney, R. S. 1987. A review of study designs for the detection of long-term environmental effects of offshore petroleum activities. Pages 651–696 *in* D. F. Boesch and N. N. Rabalais, editors. Long-term effects of offshore oil and gas development. Elsevier Applied Science, New York, New York, USA.

Carpenter, S. R., T. M. Frost, D. Heisey, and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology **70**: 1142–1152.

Cohen, J. 1977. Statistical power analysis for the behavioral sciences. Academic Press, New York, New York, USA.

Cooper, S. D., and L. A. Barmuta. 1993. Field experiments in biomonitoring. Pages 399–441 *in* D. M. Rosenberg and V. H. Resh, editors. Freshwater biomonitoring and benthic

macroinvertebrates. Chapman and Hall, New York, New York, USA.

DeAngelis, D. L., and L. J. Gross, editors. 1992. Individual-based models and approaches in ecology: populations, communities, and ecosystems. Chapman and Hall, New York, New York, USA.

Eberhardt, L. L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management **4**:27–70.

Eberhardt, L. L., and J. M. Thomas. 1991. Designing environmental field studies. Ecological Monographs **61**:53–73.

Fairweather, P. G. 1991. Statistical power and design requirements for environmental monitoring. Australian Journal of Marine and Freshwater Research **42**:555–567.

Faith, D. P., C. L. Humphrey, and P. L. Dostine. 1991. Statistical power and BACI designs in biological monitoring: comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate communities in Rockhole Mine Creek, Northern Territory, Australia. Australian Journal of Marine and Freshwater Research **42**: 589–602.

Ferris, V. R., and J. M. Ferris. 1979. Thread worms (Nematoda). Pages 1–33 *in* C. W. Hart, Jr., and S. L. H. Fuller, editors. Pollution ecology of estuarine invertebrates. Academic Press, New York, New York, USA.

Flegal, A. R., G. J. Smith, G. A. Gill, S. Sañudo-Wilhelmy, and L. C. D. Anderson. 1991. Dissolved trace element cycles in the San Francisco Bay estuary. Marine Chemistry **36**:329–363.

Frost, T. M., S. R. Carpenter, and T. K. Kratz. 1992. Choosing ecological indicators: effects of taxonomic aggregation on sensitivity to stress and natural variability. Pages 215–227 *in* D. H. McKenzie, D. E. Hyatt, and V. J. McDonald, editors. Ecological indicators. Volume 1. Elsevier Applied Science, New York, New York, USA.

Gill, J. L. 1978. Design and analysis of experiments in the animal and medical sciences. Volumes 1–3. Iowa State University Press, Ames, Iowa, USA.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York, New York, USA.

———. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research **50**: 195–205.

Hallam, T. G., R. R. Lassiter, J. Li, and W. McKinney. 1990. Toxicant-induced mortality in models of *Daphnia* populations. Environmental Toxicology and Chemistry **9**:597–621.

Herman, P. M. J., and C. Heip. 1988. On the use of meiofauna in ecological monitoring: who needs taxonomy? Marine Pollution Bulletin **19**:665–668.

Higashi, R., G. Cherr, C. Bergens, T. Fan, and D. Crosby. 1992. Toxicant isolation from a produced water source in the Santa Barbara Channel. Pages 223–233 *in* J. P. Ray and F. R. Englehardt, editors. Produced water: technological/environmental issues and solutions. Plenum, New York, New York, USA.

Holbrook, S. J., R. J. Schmitt, and R. F. Ambrose. 1990. Biogenic habitat structure and characteristics of temperate reef fish assemblages. Australian Journal of Ecology **15**: 489–503.

Jones, M., C. Folt, and S. Guarda. 1991. Characterizing individual, population and community effects of sublethal levels of aquatic toxicants: an experimental case study using *Daphnia*. Freshwater Biology **26**:35–44.

Keough, M. J. 1986. The distribution of a bryozoan on seagrass blades: settlement, growth, and mortality. Ecology **67**:846–857.

Krause, P. R., C. W. Osenberg, and R. J. Schmitt. 1992. Effects of produced water on early life stages of a sea urchin: stage-specific responses and delayed expression. Pages 431–444 *in* J. P. Ray and F. R. Englehardt, editors. Produced water: technological/environmental issues and solutions. Plenum, New York, New York, USA.

Landry, M. R., and B. M. Hickey, editors. 1989. Coastal oceanography of Washington and Oregon. Elsevier Science, Amsterdam, The Netherlands.

Luoma, S. N., D. J. Cain, C. Brown, and E. V. Axtman. 1991. Trace metals in clams (*Macoma balthica*) and sediments at the Palo Alto mudflat in South San Francisco Bay: April, 1990–April, 1991. United States Geological Survey Report **91-460**.

Magnuson, J. J., B. J. Benson, and T. K. Kratz. 1990. Temporal coherence in the limnology of a suite of lakes in Wisconsin, U.S.A. Freshwater Biology **23**:145–159.

Mapstone, B. D. *In press.* Scalable decision criteria for environmental impact assessment: Type I and Type II errors. *In* R. J. Schmitt and C. W. Osenberg, editors. Ecological impact assessment: conceptual issues and application in coastal marine habitats. University of California Press, Berkeley, California, USA.

Middleditch, B. S. 1984. Ecological effects of produced water effluents from offshore oil and gas production platforms. Ocean Management **9**:191–316.

Neff, J. M. 1987. Biological effects of drilling fluids, drill cuttings and produced waters. Pages 469–538 *in* D. F. Boesch and N. N. Rabalais, editors. Long-term environmental effects of offshore oil and gas development. Elsevier Applied Science, New York, New York, USA.

Neff, J. M., N. N. Rabalais, and D. F. Boesch. 1987. Offshore oil and gas development activities potentially causing long-term environmental effects. Pages 149–173 *in* D. F. Boesch and N. N. Rabalais, editors. Long-term environmental effects of offshore oil and gas development. Elsevier Applied Science, New York, New York, USA.

Oakden, J. M., J. S. Oliver, and A. R. Flegal. 1984. Behavioral responses of a phoxocephalid amphipod to organic enrichment and trace metals in sediments. Marine Ecology Progress Series **14**:253–257.

Olson, R. 1985. The consequences of short-distance larval dispersal in a sessile marine invertebrate. Ecology **66**:30–39.

Osenberg, C. W., S. J. Holbrook, and R. J. Schmitt. 1992*a*. Implications for the design of environmental assessment studies. Pages 75–90 *in* P. M. Grifman and S. E. Yoder, editors. Perspectives on the marine environment. Sea Grant Institutional Program, Hancock Institute for Marine Studies, University of Southern California, Los Angeles, California, USA.

Osenberg, C. W., R. J. Schmitt, S. J. Holbrook, and D. Canestro. 1992*b*. Spatial scale of ecological effects associated with an open coast discharge of produced water. Pages 387–402 *in* J. P. Ray and F. R. Englehardt, editors. Produced water: technological/environmental issues and solutions. Plenum, New York, New York, USA.

Osenberg, C. W., E. E. Werner, G. G. Mittelbach, and D. J. Hall. 1988. Growth patterns in bluegill (*Lepomis macrochirus*) and pumpkinseed (*L. gibbosus*) sunfish: environmental variation and the importance of ontogenetic niche shifts. Canadian Journal of Fisheries and Aquatic Sciences **45**:17–26.

Pearson, T. H., and R. Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology Annual Review **16**:229–311.

Peterman, R. M. 1990. Statistical power analysis can im-

prove fisheries research and management. Canadian Journal of Fisheries and Aquatic Science **47**:2–15.

Raimondi, P. T. 1990. Patterns, mechanisms, consequences of variability in settlement and recruitment of an intertidal barnacle. Ecological Monographs **60**:283–309.

Raimondi, P. T., and R. J. Schmitt. 1992. Effects of produced water on settlement of larvae: field tests using red abalone. Pages 415–430 *in* J. P. Ray and F. R. Englehardt, editors. Produced water: technological/environmental issues and solutions. Plenum, New York, New York, USA.

Sañudo-Wilhemy, S., and A. R. Flegal. 1991. Trace element distributions in coastal waters along the US-Mexican boundary: relative contributions of natural processes vs. anthropogenic inputs. Marine Chemistry **33**:371–392.

Schmidt, H., and C. E. Reimers. 1991. The recent history of trace metal accumulation in the Santa Barbara basin, Southern California Borderland. Estuarine, Coastal and Shelf Science **33**:485–500.

Schmitt, R. J., and S. J. Holbrook. 1990. Contrasting effects of giant kelp on dynamics of surfperch populations. Oecologia **84**:419–429.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, R. O. Smith, and J. R. Bence. 1993. Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. Ecological Applications **3**:331–350.

Skalski, J. R., and D. H. McKenzie. 1982. A design for aquatic monitoring programs. Journal of Environmental Management **14**:237–251.

Spies, R. B. 1987. The biological effects of petroleum hydrocarbons in the sea: assessments from the field and microcosms. Pages 411–467 *in* D. F. Boesch and N. N. Rabalais, editors. Long-term environmental effects of offshore oil and gas development. Elsevier Applied Science, New York, New York, USA.

Spies, R. B., and D. J. DesMarais. 1983. Natural isotope study of trophic enrichment of marine benthic communities by petroleum seepage. Marine Biology **73**:67–71.

Stewart-Oaten, A. *In press.* Problems in the analysis of environmental monitoring data. *In* R. J. Schmitt and C. W. Osenberg, editors. Ecological impact assessment: conceptual issues and application in coastal marine habitats. University of California Press, Berkley, California, USA.

Stewart-Oaten, A., J. R. Bence, and C. W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. Ecology **73**:1396–1404.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology **67**:929–940.

Stretch, J. J. 1983. Habitat selection and vertical migration of sand-dwelling demersal gammarid amphipods. Dissertation. University of California, Santa Barbara, California, USA.

Thrush, S. F., R. D. Pridmore, and J. E. Hewitt. 1994. Impacts on soft-sediment macrofauna: the effects of variation on temporal trends. Ecological Applications **4**:31–41.

Underwood, A. J. 1991. Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. Australian Journal of Marine and Freshwater Research **42**:569–587.

———. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications **4**:3–15.

Underwood, A. J., and C. H. Peterson. 1987. Towards an ecological framework for investigating pollution. Marine Ecology Progress Series **46**:227–234.

Vezina, A. F. 1988. Sampling variance and the design of quantitative surveys of the marine benthos. Marine Biology **97**:151–155.

Victor, B. C. 1986. Larval settlement and juvenile mortality in a recruitment-limited coral reef fish population. Ecological Monographs **56**:145–160.

Warwick, R. M. 1988. The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. Marine Pollution Bulletin **19**:259–268.

Yoccoz, N. G. 1991. Use, overuse and misuse of significance tests in evolution and ecology. Bulletin of the Ecological Society of America **72**:106–111.

REFERENCES

(see also the Literature Cited sections in the reprints in sections III, IV and V)

Boesch, D.F. and N.N. Rabalais, eds. 1987. *Long-term Environmental Effects of Offshore Oil and Gas Development.* Elsevier Applied Science, New York, 708p.

Carney, R. 1987. A review of study designs for the detection of long-term environmental effcts of offshore petroleum activities. Pages 651-696 in: *Long-term environmental effects of offshore oil and -gas development*. D.F. Boesch & N.N. Rabalais, eds., Elsevier Applied Science, New York.

Carpenter S.R., T.M. Frost, D. Heisey, T.K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole ecosystem experiments. *Ecology* **70**:1142-1152.

Fairweather, P.G. 1996. Coping with complexity. *Trends in Ecology and Evolution* **11**:485.

National Research Council. 1990. *Managing Troubled Waters: The role of marine environmental monitoring*. National Academy Press, Washington, D.C. 125 p.

Neff, J.M. 1987 Biological effects of drilling fluids, drill cuttings and produced waters. Pages 469-538 in: *Long-term environmental effects of offshore oil and gas development*. D.F. Boesch & N.N. Rabalais, eds., Elsevier Applied Science, New York.

Osenberg, C.W., S.J. Holbrook and R.J. Schmitt. 1992. Implications for the design of environmental assessment studies. Pages 75-90 in P.M. Griffman and S.E. Yoder, eds, *Perspectives on the Marine Environment*. USC Sea Grant, Los Angeles, California.

Osenberg, C.W. and R.J. Schmitt. 1994. Detecting human impacts in marine habitats. *Ecological Applications* 4:1-2.

Osenberg, C.W. and R.J. Schmitt. 1996. Detecting ecological impacts caused by human activities. Pages 3-16 *in* R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. *Ecological Applications* **4**:16-30.

Piltz, F.M. 1996. Organizational constraints on environmental impact assessment research. Pages 317-328 in R.J. Schmitt and C.W. Osenberg (eds.) *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego.

Rachlin, J.W. 1996. Review of Detecting Ecological Impacts. *Transactions of the American Fisheries Society* **126**:545-547.

Schmitt, R.J. and C.W. Osenberg (editors and contributing authors). 1996. *Detecting ecological impacts: Concepts and applications in coastal habitats*. Academic Press, San Diego. 401 p.

Spies, R.B. 1987. The biological effects of petroleum hydrocarbons in the sea: assessments from the field and microcosms. Pages 411-467 in D.F. Boesch and N.N. Rabalais, editors. *Long-term environmental effects of offshore oil and gas development*. Elsevier Applied Science, New York.

Stewart-Oaten, A., J. Bence and C.W. Osenberg. 1992. Assessing effects of unreplicated perturbations: no simple solutions. *Ecology* **73**:1396-1404.

Stewart-Oaten, A., W.W. Murdoch and K.R. Parker. 1986. Environmental impact assessment: "Pseudoreplication" in time? Ecology 67:929-940.

The Department of the Interior Mission

As the Nation's principal conservation agency, the Department of the Interior has responsibility for most of our nationally owned public lands and natural resources. This includes fostering sound use of our land and water resources; protecting our fish, wildlife, and biological diversity; preserving the environmental and cultural values of our national parks and historical places; and providing for the enjoyment of life through outdoor recreation. The Department assesses our energy and mineral resources and works to ensure that their development is in the best interests of all our people by encouraging stewardship and citizen participation in their care. The Department also has a major responsibility for American Indian reservation communities and for people who live in island territories under U.S. administration.

The Minerals Management Service Mission

As a bureau of the Department of the Interior, the Minerals Management Service's (MMS) primary responsibilities are to manage the mineral resources located on the Nation's Outer Continental Shelf (OCS), collect revenue from the Federal OCS and onshore Federal and Indian lands, and distribute those revenues.

Moreover, in working to meet its responsibilities, the **Offshore Minerals Management Program** administers the OCS competitive leasing program and oversees the safe and environmentally sound exploration and production of our Nation's offshore natural gas, oil and other mineral resources. The MMS **Royalty Management Program** meets its responsibilities by ensuring the efficient, timely and accurate collection and disbursement of revenue from mineral leasing and production due to Indian tribes and allottees, States and the U.S. Treasury.

The MMS strives to fulfill its responsibilities through the general guiding principles of: (1) being responsive to the public's concerns and interests by maintaining a dialogue with all potentially affected parties and (2) carrying out its programs with an emphasis on working to enhance the quality of life for all Americans by lending MMS assistance and expertise to economic development and environmental protection.