Vol. 210: 255–283, 2001 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published January 26

REVIEW

Sampling design begets conclusions: the statistical basis for detection of injury to and recovery of shoreline communities after the 'Exxon Valdez' oil spill

Charles H. Peterson[1,*], Lyman L. McDonald[2], Roger H. Green[3], Wallace P. Erickson[2]

[1]University of North Carolina at Chapel Hill, Institute of Marine Sciences, Morehead City, North Carolina 28557, USA
[2]WEST, Inc., 2003 Central Avenue, Cheyenne, Wyoming 82001, USA
[3]Department of Zoology, University of Western Ontario, London, Ontario N6A 5B7, Canada

ABSTRACT: The joint effect of multiple initial decisions made about sampling design in evaluation of environmental impacts using observational field assessments influences the ability to detect and accurately estimate responses. The design can dictate in advance whether the study can identify even large impacts that truly exist. Following the 'Exxon Valdez' oil spill in 1989, 4 separate studies of effects of the spill on the intertidal biota were conducted. Studies overlapped sufficiently in geographic area, shoreline habitat, and biological response variables to permit contrasts showing how the aggregate of multiple design decisions led to differences in conclusions. The SEP (Shoreline Ecology Program) supported by Exxon and the CHIA (Coastal Habitat Injury Assessment) funded by the Exxon Valdez Oil Spill Trustee Council shared a core approach of establishing a stratified random design of site selection. The Exxon-supported GOA (Gulf of Alaska) study and the NOAA (National Oceanographic and Atmospheric Administration) Hazmat (Hazardous Materials) study both chose to employ subjective choices of fixed sites. Despite many common goals, these 4 studies differed greatly in: (1) sampling effort (area covered per sample quadrat, sample replication within sites, numbers of study sites per category, numbers of samplings, and total areas sampled) and sampling design (philosophy of targeting sampling effort, complete randomization versus matched pair designs, sampling frame, treatment of habitat heterogeneity within sites, interspersion of sites, and control of shoreline treatment and oiling intensity); (2) analytical methodology (analysis of covariance versus paired designs, treatment of subsamples as replicates in *F*-ratios, logic of inferring recovery, and power calculations); and (3) choice of biological response variables (taxonomic level of analysis, aggregating versus splitting separate communities, and scope of communities and habitats examined). The CHIA and NOAA Hazmat studies of epibiotic responses in sheltered rocky shores of Prince William Sound made several decisions to enhance detection power and produced similar conclusions about large reductions in total biotic cover of intertidal space, *Fucus* cover, mussel abundance, abundance of the limpet *Tectura persona* and a balanoid barnacle, and increases in open space and abundance of an opportunistic barnacle, *Chthamalus dalli*. In contrast, the SEP study of this same habitat and geographic region adopted design choices resulting in lower power of detection in 12 (vs CHIA and vs NOAA Hazmat) of 15 separate decisions (with one tie in each contrast). Accordingly, the SEP study was able to detect declines only in *Fucus* cover and occasionally in total limpet abundance but not in total epifaunal or mussel or balanoid barnacle abundance and, unlike the results of the other 2 studies, most of the taxa analyzed showed apparent increases rather than decreases from oiling and shoreline treatment. The more powerful GOA and CHIA studies of impacts of oiling in the Gulf of Alaska, where oil grounded 1 to 8 wk later and in more weathered condition than in Prince William Sound, showed more consistent and larger reductions in intertidal biota in the sheltered rocky habitat than did the SEP study of Prince William Sound. Thus, the combined effects of many design decisions that reduced power to detect impacts in the SEP study led to failure to demonstrate large impacts of the spill documented by other studies of the same habitat in the same and the more remote region.

KEY WORDS: Analytical methodology · Biological response variables · Environmental impact assessment · 'Exxon Valdez' oil spill · Intertidal biota · Sampling design · Statistical power · Type II error

[*] E-mail: cpeters@email.unc.edu

INTRODUCTION

In the 2 decades that have passed since publication of Green's (1979) book on environmental statistics with its 10 simple rules for how to test effectively for environmental responses, the literature on assessment designs for environmental impact studies has grown dramatically in abundance and sophistication. By 1986, the National Research Council (1986) in its comprehensive report on oil in the sea wrote that '[a] discussion of statistical techniques used in studies of the fate and effects of oil in the environment would need to cover most of the areas in modern statistics.' More recently, new analytical tools (e.g., Field et al. 1982, Clarke & Ainsworth 1993, Manly 1997) and novel statistical test designs (e.g., Underwood 1981, 1994, 1997, Stewart-Oaten et al. 1986, 1992, Wiens & Parker 1995, Schmitt & Osenberg 1996) have been developed for the explicit task of assessing environmental impacts. Despite this growth in technical sophistication and subtlety of assessment designs and analyses, relatively simple decisions, assumptions, and conditions of the study design can still be extracted to explain how conclusions may be dictated by choice of study design.

Approaches to inferring the degree of injury to natural resources following an environmental perturbation differ dramatically, with major implications for the types of statistical support required. In ecology and environmental sciences, the practice of applied statistics has been slowly moving away from testing a null hypothesis as the sole means of evaluating responses and towards estimation statistics (Stewart-Oaten et al. 1992, Johnson 1999). Testing the null hypothesis of no effect on various species populations is trivial when deaths have been observed. The more compelling goal is to estimate the magnitude of the loss or the time course to recovery, and in such cases statistics are critical in calculating the confidence intervals or the Bayesian credibility intervals around those estimates. This estimation approach is not yet engrained into the fabric of environmental assessment. Some acute injuries resulting from toxic exposure are currently estimated, for example those based upon counts of oiled, dead seabirds (e.g., Piatt & Lensink 1989). For this class of injury, inferential statistics are used to assess sampling efficiency and confidence intervals of the estimates and to model post-mortem fate and transport (Piatt & Ford 1996). A typically small fraction of those seabirds that are killed by acute contact with oil spilled at sea is subsequently recovered, thus requiring a model to adjust mortality estimates for probability of recovery (Ford et al. 1987). Despite such examples of using body counts to estimate mortality, the intellectual framework that has guided most formal natural resource damage assessments that use field observations has concentrated on testing null hypotheses of no effect.

Another fundamental contrast in approach to environmental assessment juxtaposes design/data-based and model-based methods (Gilbert 1987). A design/data-based protocol involves an empirical assessment to collect new data in a design intended to provide estimates of biological parameters or a direct test of an hypothesis (e.g., Cochran 1977). A pure model-based protocol includes no collection of new data, but rather involves construction of a deductive model to reach a conclusion. In practice, even design/data-based inference depends on models because its probabilities are based on the specific randomization procedure used to select sites, assign treatments, and so forth (Cox 1958, Manly 1997). Perturbations (oiling) are not assigned at random to sites, so design/data-based inference requires an assumed model intrinsic to observational studies that oiled and unoiled sites would have the same distribution in the absence of the oiling. What one is really interested in is how the specific oiled sites differ from the state that they would have reached in the absence of the perturbation. Inference is further complicated by realization that what is desired is actually a time sequence of how the oiled sites would have behaved had there been no oil spill. Because all 4 studies of oil spill impact contrasted here are design/data-based, relevant issues that arise include comparisons of the model-based rationales that underlie the design protocols chosen in each case. Some have argued that design/data-based studies provide superior quantification of injury and greater reliability (Gilbert 1987, Johnson et al. 1989) because models typically require multiple assumptions wherever their demands for information go beyond available scientific understanding of critical underlying processes. For example, Peterson (2001) shows how use of an ecotoxicity risk assessment approach to model an oil spill as a pulse perturbation of acute mortality fails to include important chronic and indirect delayed impacts, which can best be evaluated empirically in a long-term field assessment program. Yet, the large differences in outcomes among the field studies of 'Exxon Valdez' oil spill impacts on intertidal communities that we consider here illustrate a potential shortcoming in design/data-based approaches, their dependency on critical design decisions.

Many generic components of a study design affect the outcomes of observational assessments and the conclusions derived from them (e.g., Bernstein & Zalinski 1983, Gilbert 1987, Peterson 1993, Mapstone 1995, Underwood 1997). When using the approach of testing null hypotheses to assess injury to species populations through field observations, achieving a defensible balance between type I and type II errors is a challenge.

Because heterogeneity within natural ecological systems creates noisy data, the issue of power and type II error in environmental assessment has attracted substantial attention in recent literature (Green 1989, 1994, Eberhardt & Thomas 1991, Fairweather 1991, Peterman & M'Gonigle 1992, Peterson 1993, Steidl et al. 1997). Prior to the guidance provided by these publications, the historical tradition of natural science had focused almost exclusively on avoiding type I errors, falsely concluding that an effect exists (Toft & Shea 1983). In environmental studies, such type I errors are a special concern when the mechanistic causal model is absent or equivocal. Controlling the type I error rate at a low level has, however, the implication of elevating the likelihood of making the alternative type II error of falsely concluding absence of an effect when there is one. In tests where the null hypothesis of no significant effect cannot be rejected, basic scientists now routinely estimate the power of the test or the magnitude of detectable effect size so as to provide an indication of the probability of type II errors. In environmental assessments using hypothesis testing approaches, where making a type II error may be especially serious or costly to the public interest, there is added incentive for insuring high power in important tests (Fairweather 1991, Peterman & M'Gonigle 1992). Because of this interplay between type I and type II error rates, the issue of power of designs can be abstracted to a question of burden of proof: should managers of public trust resources carry the burden of proving that injury occurred beyond some small doubt or should the party responsible for the environmental incident carry the burden of proving that there was not any injury (Dayton 1998). In the ideal situation, a study design will have low error rates for both types of mistaken conclusions, but in many cases high costs of such powerful designs will require compromising the traditionally fixed type I error rates to insure low type II errors. In either case, attention to power in all aspects of the study design is critical to making correct inferences about impacts under natural conditions of environmental variability.

Here we take advantage of a rare opportunity to compare redundant, or at least broadly overlapping, studies of how the same ecological system responded to a large environmental perturbation. Paine et al. (1996) decried the huge wastage of funds involved in the intensive and redundant assessments of ecological impacts following the 'Exxon Valdez' oil spill of 1989. Here, we exploit this rare duplication to evaluate how simple design decisions within large complex studies combine to influence the outcomes of those studies. We review the design decisions and conclusions of 4 separate, largely independent studies of the impacts of the same environmental incident, the 'Exxon Valdez' oil spill, on the same system, the shoreline community of intertidal plants and animals. Each of these studies required investment of substantial resources and involved rather sophisticated sampling and analytical designs using trained statisticians and biologists, yet they appeared to reach different conclusions. We show by this example how application of relatively simple concepts and principles in sampling design and statistical analysis can explain the reasons for reaching different conclusions. Specifically, we demonstrate how multiple design decisions that affect the ability to remove or avoid bias in parameter estimates from observational studies and influence the power to detect effects can combine to produce inconclusive results.

MATERIALS AND METHODS

The 'Exxon Valdez' oil spill. On 24 March 1989, the tanker 'Exxon Valdez' ran aground on Bligh Reef in the northeastern region of Prince William Sound, Alaska, leading to a release of approximately 11 million gallons (35 000 tonnes) of North Slope crude oil. After 3 d of calm weather, the floating oil was then transported by storm winds and prevailing currents to the southwest, where it first encountered the shorelines of many of the islands in Prince William Sound (see map of oil transport in Babcock et al. 1996). Subsequently, oil was transported out of Montague Strait, contacting several sites along the outer Kenai coast and then inside lower Cook Inlet. Finally, the oil traveled westward to the shores of the Kodiak Island complex and the Alaska Peninsula along the Shelikov Strait. During this 1 to 8 wk journey, the floating oil changed chemically and physically as some more volatile components entered the atmosphere, as others were oxidized photochemically and via bacterial metabolism, and as the floating crude oil became more consolidated into mousse and patties (Wolfe et al. 1994).

About half the spilled oil was estimated to have come ashore on beaches, 40 % within Prince William Sound and 7 to 11 % on shorelines of the Kenai and Alaska Peninsulas and on the Kodiak Island complex (Spies et al. 1996). Aerial surveys by the Alaska Department of Environmental Conservation reported that by the end of summer 1989: (1) 280 of the 1182 miles observed in Prince William Sound revealed light-to-heavy oiling; (2) 168 of 1039 miles of Kenai Peninsula-Cook Inlet shoreline observed revealed light-to-heavy oiling; and (3) 590 out of 1850 miles observed on the Kodiak-Alaska Peninsula region showed light-to-heavy oiling (ADEC 1989). Heavy oiling was much more concentrated in Prince William Sound. The shoreline assessment conducted by Exxon using aerial videotape suggested even greater lengths of oiled shoreline, including over 500 miles in Prince William Sound (Neff et al. 1995). Intensive treatment of oiled shorelines applied to displace and remove the oil from the intertidal zone took place in summers of 1989 to 1991 with some further work in spring 1992. Shoreline treatments included mechanical excavation of shoreline rocks, pressurized application of hot or cool water, hand cleaning of rocks, and bioremediation via nutrient application (Mearns 1996). Injuries to the biota of the oiled shorelines thus included both the effects of oiling and the effects of shoreline treatment.

Studies of impact of the 'Exxon Valdez' oil spill on intertidal communities. Four major studies were conducted to assess the responses of intertidal biota to the 'Exxon Valdez' oil spill and subsequent shoreline treatment: CHIA (Coastal Habitat Injury Assessment) (McDonald et al. 1995, Highsmith et al. 1996, Stekoll et al. 1996, Sundberg et al. 1996), conducted by scientists funded by the Exxon Valdez Oil Spill Trustee Council (a consortium of federal and state government agencies with responsibilities for managing public trust resources); SEP (Shoreline Ecology Program) (Page et al. 1995, Gilfillan et al. 1995a), conducted by Exxon contractors to evaluate impacts in Prince William Sound; GOA, also funded by Exxon but targeting the Gulf of Alaska spill area (Gilfillan et al. 1995b); and a fourth study conducted by the NOAA (National Oceanographic and Atmospheric Administration) Hazmat (Hazardous Materials) program (Driskell et al. 1996, Houghton et al. 1996, Lees et al. 1996). A fifth study initiated by the US National Park Service was terminated when oil failed to come ashore at any of the sites where pre-spill surveys were conducted. Design and analysis were complicated by: (1) the large extent of oiled shoreline; (2) the extreme heterogeneity of affected shoreline types and habitats; (3) the varying degrees and types of oiling and shoreline treatment; (4) inaccuracies in maps and databases on shoreline types and locations of oil and shoreline treatments; and (5) the challenge of making statistical inferences on the full geographic extent of the injuries, magnitude of injuries, and extent of recovery. These difficulties in designing intertidal assessments ultimately forced both the CHIA and the SEP studies to be completely redesigned after the first field season, resulting in loss of critical information about injury in 1989.

Two fundamentally different approaches were embodied within these 4 damage assessment studies, an approach involving random selection of sites in some fashion designed to permit extrapolation to a larger spill-affected area and an alternative approach of choosing fixed sites, selected to cover particular habitats and/or to reflect known shoreline oiling and treatments. The CHIA study and the SEP (specifically the SRS [stratified random sampling] portion; Page et al. 1995) study for Prince William Sound followed the first of these approaches. The NOAA Hazmat study and the GOA assessment adopted the second approach. (There was also a small portion of both the Trustee Council- and the Exxon-funded programs in Prince William Sound that employed fixed sites.) The fundamental dichotomy in these 2 approaches is so great that it should not be surprising that different conclusions might arise from them. In practice, the fixed-site results are probably most useful in providing separation of otherwise confounded effects of oiling and shoreline treatment and in assessing the process of recovery at fixed sites of known history, whereas the programs involving random sampling of shorelines are more appropriate for extrapolating to estimate the full extent of injury. We devote most of our effort in this paper to comparing the 2 stratified random studies because they held an important geographic region (Prince William Sound) and a year (1990) in common and ostensibly had similar goals with similar methodologies. A close examination of the 2 protocols readily reveals fundamental differences.

The stratified random designs. CHIA and the stratified random portion of SEP, despite their common approach of random site selection, possessed many differences in design that caused them to generate different conclusions. The first decision made in designing both studies was to stratify the shorelines of both the oiled and reference areas into distinct subregions that were more homogeneous with respect to factors that influence the intertidal biological communities (e.g., Cochran 1977, Thompson 1992). These choices of strata were made in somewhat different ways in the 2 studies. CHIA first stratified by geographic area into 3 spill regions (Prince William Sound, Kenai Peninsula-lower Cook Inlet, and Kodiak-Alaska Peninsula) because of likely pre-existing biological differences among these environmentally and geographically different regions and because of changes in the quality of the oil over time during its sequential transport to the different regions. The Exxon-supported assessment studies stratified geographically by conducting 2 independent studies in the 2 separate regions, Prince William Sound (SEP; Page et al. 1995, Gilfillan et al. 1995a) and the Gulf of Alaska (GOA; Gilfillan et al. 1995b), which included both the Kenai coast and the Kodiak archipelago-Alaska Peninsula area.

A second decision held in common, although done with operational differences by each study, was to stratify by habitat type within each geographic region. The SEP design defined and sampled 4 intertidal habitats in Prince William Sound: exposed bedrock, sheltered bedrock, boulder/cobble, and pebble/gravel (Page et al. 1995). All sites close to eagle nests were eliminated in SEP to avoid any unintentional impacts on a charismatic species. Fine-sediment shores were deemed too rare to sample. CHIA chose to ignore and not sample the steep and dangerous wave-exposed rocky shores where worker safety would be imperiled. The CHIA design identified 5 different intertidal habitats: exposed rocky shores, sheltered rocky shores, coarse-textured beaches, fine-textured beaches, and sheltered estuarine shores (Sundberg et al. 1996). Sampling in CHIA was conducted in all habitats, although fine-textured beaches were dropped after the 1990 field season and analyses of this habitat remain incomplete. Not all habitats were sampled in every geographic area: sheltered estuarine shores and exposed rocky shores were not included in the Kodiak-Alaska Peninsula design, and exposed rocky shores were also excluded from sampling in the Kenai Peninsula-Cook Inlet region (Stekoll et al. 1996). Thus, despite some general similarity in habitat identification (especially for exposed and sheltered rocky [= bedrock]), there were also differences between these 2 studies in their choices of habitat strata. Contrasts of results between studies are most readily achieved for the sheltered rocky habitat, which was sampled in all regions by each study. Although the boulder/cobble and pebble/gravel habitats of SEP appear to match the coarse-textured beaches of CHIA, sampling differences between studies in this environment and the lack of parallelism in habitat definitions inhibit definitive comparisons.

A third decision made in establishing the design of these 2 stratified random shoreline assessments was to stratify by elevation in the intertidal zone, recognizing that differences in aerial exposure have critical impacts in structuring intertidal communities (e.g., Connell 1972). SEP identified and sampled biological communities at 4 elevation zones: the upper intertidal at mean high water; the middle intertidal at mean tide level; the lower intertidal at mean lower low water; and the subtidal at 3 m below mean lower low water (Page et al. 1995). CHIA stratified initially into 4 zones, each of the first 4 m of vertical drop starting from mean high water (Highsmith et al. 1996). Because the mean low water level is 3.4 m below mean high in Prince William Sound, the fourth meter of vertical drop was not always accessible to sampling, so sampling at this level was discontinued after the 1990 field season. Thus, the elevation zones sampled in the 2 stratified random designs are similar but not identical.

In both stratified random studies, a GIS-based map of shoreline type and oiling was used to help identify the sampling frame of oiled and references sites for each habitat stratum (McDonald et al. 1995, Page et al. 1995, Sundberg et al. 1996). For CHIA, shoreline segments 100 to 600 m long within each stratum were defined and selected by a random procedure with probability proportional to length, while eliminating all sites less than 100 m long. SEP defined study sites of uniform length of 100 m, eliminating study sites less than 100 m for all strata except 1 habitat type where sites of 60 m were used. SEP then similarly selected a simple random sample of equal-sized sites from each stratum. Both CHIA and SEP carved up any long stretches of identical habitat into contiguous segments equal to the maximum length. Consequently, the sampling frame for both studies eliminated from consideration some types of sites that truly exist, the short segments of habitat, thereby requiring the assumption that conclusions from the longer beach segments can be extrapolated to shorter sites. In both CHIA and SEP, if the field assessment team found that a study site was misclassified by any criterion, then it was dropped and could not be included among the possible sampling sites for the stratum to which it properly belonged. During analysis of the SEP, data were combined across strata in multivariate analyses without regard for the unequal weights created by the actual practice of stratified random sampling. This shortcoming in the SEP design violates the principle of equal (or, more generally, known) probability of sampling every site within a stratum. CHIA planned for unequal probabilities of sampling from the start and modified the weights assigned to the data by knowledge of the imperfections in classification and how sites were correctly and incorrectly assigned to strata. The method used by CHIA (see McDonald et al. 1995) to apply unequal weights was the Stouffer-Liptak meta-analysis procedure (Folks 1984), also known as the inverse normal method (Hedges & Olkin 1985) or a consensus test (Rice 1990). Unpublished analyses conducted during SEP indicate, however, that this problem of unequal weights had minor effects on conclusions of SEP (J. Harner pers. comm.).

The 2 stratified random assessment studies differed fundamentally in the protocol for site selection and the basic design for contrasting sites. CHIA, as it was conducted from 1990 onwards, used a control-treatment paired design (Skalski & Robson 1992), in which the oiled sites were randomly sampled from a frame of all moderately and heavily oiled sites. Then, each oiled site was compared against a matched reference site, with selection of the match done on the basis of geographic proximity, beach slope, wave exposure, substrate composition, nearshore bathymetry, and proximity to sources of freshwater (McDonald et al. 1995, Highsmith et al. 1996). Reference sites included lightly oiled as well as unoiled sites, potentially making inferences on the impacts of oiling conservative and reducing power to detect oil spill effects if light oiling actually caused any mortality. SEP selected both oiled and reference sites from a Prince William Sound sampling frame without pairing. By adhering to a strict randomization procedure, reference sites were not interspersed with oiled sites and the frame of selection for reference sites was broadly enough defined that it included sites near glacier inputs that seasonally lowered salinities and enhanced turbidity relative to the oiled sites. Site properties of sediment size, total organic carbon, and wave energy were measured and used as covariates to analyze effects of oiling (Page et al. 1995). Four oiling categories were used in SEP: heavy, moderate, light to very light, and none. Neither of these approaches necessarily addresses a critical underlying problem in such a design. The investigator in such environmental assessments does not control the oiling and does not assign oiling at random; thus the possibility that oiled sites differ systematically from unoiled sites because of differential exposure to current flux or some other factor is not tested, and the logic of both designs is deficient (Peterson 1993). This concern illustrates how some sort of conceptual model underlies even a design/data-based assessment.

Subsampling techniques at each site differed between the 2 stratified random assessment studies, as did methods of using the data for statistical inference. Both CHIA and SEP used quadrats on rocks and cores in soft sediments located along replicate transects to provide the subsamples at each site, but the size, number, and shape of quadrat samples all differed greatly. CHIA involved 5 different sampling dates: 1 pilot sampling in 1989, then 2 samplings (early and late summer) in both 1990 and 1991 to provide an indication of rate and extent of recovery (McDonald et al. 1995). The SEP study of biological response was a 1-time sampling of Prince William Sound in summer 1990, although some repeated sampling of the fixed study sites occurred with minimal reporting of the results (Gilfillan et al. 1995a, Page et al. 1995). The choices of injury indicators differed between the 2 studies, with CHIA measuring the abundance and biomass by species of intertidal plants and invertebrates, while SEP generally pooled species into coarser taxonomic groupings and community-level parameters (such as species richness and Shannon-Wiener diversity) for their analyses. CHIA included a sampling of intertidal fishes, which were not a part of the SEP sampling (Barber et al. 1995). Decisions about pooling samples and about using subsamples as replicates differed in the analyses.

The fixed-site designs. The NOAA Hazmat program began as Exxon's project in 1989 and was then taken over by NOAA for subsequent years. One component study was designed as a short-term (3 to 10 d) evaluation of the biological impacts of several alternative beach treatments: low-pressure warm-water wash, the dispersant Corexit 7664, the beach cleaner Corexit 9580 M2, and high-pressure hot-water wash (Lees et al. 1996). The evaluation of each treatment was done separately at a single site, chosen to facilitate testing, in oiled, protected boulder/cobble habitat. Plots were selected at random within sites to receive treatment and the test of effects was a contrast of pre- to post-treatment abundances of epibiotic species. Lacking control plots, this design assumes that changes over the 3 to 10 d period were all due to treatment: this assumption can be only partially relaxed by contrasts with other sites nearby that were similarly oiled but untreated (Lees et al. 1996). Randomly selected 0.25 m^2 quadrats were used as the sampling units within each treatment plot. The design of this study lends itself to direct observation of process and inference on mechanism, but inference to the entire frame of oiled and treated sites, as defined by the SEP and CHIA protocols, requires some (perhaps plausible) assumptions. For example, to extrapolate, one would need to assume that variance in oiling response among sites of this habitat type is so low that the single study site represents the distribution of responses well and that the observed treatment effects remain nearly constant across different habitats.

The longer-term components of this NOAA Hazmat study of fixed sites involved use of replicate sites not chosen by a formal randomization process (Driskell et al. 1996, Houghton et al. 1996). Sites were rapidly selected during the spill event to represent oiled and unoiled sites in each major habitat type (rocky, boulder/cobble, and mixed-soft). Sites were then post-categorized by beach treatment into oiled but not treated (except perhaps by cool-water flushes and/or bioremediation fertilizers), oiled + treated with pressurized hot water (plus also fertilizers), and unoiled. Additional sites were added in 1990, 1991, and 1992 to enhance replication. This study had 2 components with different methodologies: an infaunal study using 10.7 cm diameter cores, whose invertebrates were retained on 1 mm mesh (Lees et al. 1996), and an epibiota study using 0.25 m^2 quadrats (Houghton et al. 1996). Two elevation strata were sampled in the infaunal component and 3 in the epibiota component. Thus, while assumptions are required to extrapolate results to the entire sampling frame of oiled and treated shores, the study serves directly to separate out the 2 sources of impact to shoreline biota, while providing a long-term monitoring to document recovery processes at the fixed sites in both hard- and soft-bottom communities.

The fixed-site portion of SEP included 12 sites, 8 of which were selected and sampled in 1989 (Page et al. 1995). Sites were chosen to represent certain conditions of special interest or concern, such as soft-sediment habitats or especially heavily oiled sites, so the results of this sampling are not intended to be representative of a broad type of oiled shores. Some fixed sites were revisited in 1990 and 1991, but sampling was not conducted in a comparable fashion that would allow contrast of biological parameters (Page et al. 1995). Two types of samples were taken in the fixed-site component, 0.25 × 0.125 m scrapes of rock surfaces and 10 cm diameter cores to 10 cm depth in sediments. Each type of sample was sieved in the field through a 1 mm mesh to yield abundance data for targeted taxa. A sample of each type was taken, wherever possible, at each of 3 intertidal elevations on each of 3 transects, established in perpendicular orientation to the shoreline.

GOA also did not choose sites at random and thereby employed a fixed-site design (Gilfillan et al. 1995b). Sites were chosen to cover the 2 geographic regions of the Gulf, the Kenai coast and the Kodiak-Alaska Peninsula area. Sites were also selected to provide coverage of 3 different habitats: bedrock, boulder/cobble, and pebble/gravel. Five sites characterized as sand or mud were also sampled, but sand sites lacked controls and mud sites lacked oiled counterparts, thereby preventing contrasts that might allow inference of impacts of the oil spill. Within each habitat and geographic region, replicate sites were selected in each of 3 levels of oiling, moderate-to-heavy, light, and unoiled references. For all sites, biological data were collected in 1989 by estimating areal coverage of dominant species in 0.5 × 0.5 m quadrats, by collecting 0.063 m^2 scrape samples from rock surfaces in the first 2 types of habitats, and by collecting 0.0078 m^2 cores to 15 cm depth from sediments in the latter 2 types of habitats (boulder/cobble possessing both hard- and soft-bottom substrate). Organisms collected on a 1 mm sieve comprised the core samples in the 1989 data set. Samples were taken at 5 elevations within the intertidal zone on each of 3 replicate vertical transects. In 1990 only core samples were taken but at a subset of sites mostly where cores had not been taken in 1989, preventing meaningful inferences about the status of the communities in 1990 and the recovery process in the intervening year. Thus, the target populations to which results of this study were intended to apply are the moderately-to-heavily oiled sites and the lightly oiled sites in each of 3 rocky habitat types. Such extrapolation requires several (perhaps plausible) assumptions, including especially the representative nature of the sets of subjectively chosen, fixed sites.

RESULTS OF COMPARISONS OF STUDY DESIGNS

Issues of sampling intensity and effort

(1) Area covered per sample

The 4 independent studies of effects of the 'Exxon Valdez' oil spill on intertidal epibiota used samples that encompassed substantially different surface areas (Table 1). SEP used scrape samples covering 0.031 m^2 (results of 0.25 m^2 photographs do not appear in the publications), whereas the CHIA study took samples of 0.10 (destructive scrape collection) and 0.40 (areal cover) m^2 (results of 1.5 to 1.7 m^2 samples for nearest neighbor measures do not appear in the publications) and NOAA Hazmat used areal cover samples of 0.25 m^2. Thus, the area of intertidal rock surface assayed in a single sample was 3 (scrape) to 13 (areal cover) times larger in CHIA and 8 (areal cover) times larger in NOAA Hazmat than in SEP. The corresponding GOA study (Gilfillan et al. 1995b) used samples of 0.25 m^2 for areal coverage and 0.063 m^2 for scrapes of rock surfaces: areal cover samples of a size equal to that of NOAA Hazmat and 8 times the size of SEP, and scrape samples each of an area about the same as CHIA and 8 times that of SEP (Table 1).

The areal size of a sample is an important design consideration for several reasons. The optimal size of each sample depends on the variable to be estimated. In estimating individual species densities, small sample sizes may be preferable provided that the sampling effort saved is simply redistributed by utilizing numerous small samples. For example, point estimates are a common standard in studies of spatial cover on rocky shores. However, the substantially smaller samples used in SEP were not accompanied by higher replication of sampling so total sampling effort was much lower than in the other studies (see points 2 to 5 in Table 1). A larger area of coverage by a sample can achieve better representation by spreading the sample out over a larger range of any natural gradient or across spatial heterogeneity. In this case, the CHIA scrape samples were rectangular with 0.2 m in the horizontal axis and 0.5 m in the vertical axis, along which tidal exposure differences directly and indirectly induce great changes in biota. As stated by Krebs (1989), '...nearly everyone has found that long thin quadrats are better than circular or square ones of the same area. The reason for this is habitat heterogeneity. Long quadrats cover more patches.' Thus, the larger CHIA scrape samples covered double the vertical distance of the 0.125 × 0.25 m scrape samples in SEP and of the 0.25 × 0.25 m scrape samples of GOA. The rocky intertidal community of the Pacific Northwest is well known for its patchy mosaic nature as a consequence of varying duration of succession since disturbance (Paine & Levin 1981). SEP data on distributions of abundances of species in scrape samples support the conclusion that its samples were small relative to patch sizes in that over half the distributions were best described by the negative binomial (Gilfillan et al. 1995a). Thus, CHIA scrape samples are likely to cover more of the vertical environmental gradient and aver-

Table 1. Contrasts of sampling intensity and effort in the 4 main studies to assess injury to intertidal biota following the 'Exxon Valdez' oil spill (the SRS portion of the SEP study by Exxon, the Gulf of Alaska (GOA) study of Exxon, the CHIA study of the Exxon Valdez Oil Spill Trustee Council, and the NOAA Hazmat study). PWS: Prince William Sound

Issue	SEP[a]	Exxon GOA[b]	CHIA[c]	NOAA[d]
(1) Area covered by single sample	Scrape - 0.031 m^2 Core - 0.0078 m^2	% cover- 0.25 m^2 Scrape - 0.063 m^2 Core - 0.0078 m^2	% cover - 0.40 m^2 Scrape - 0.10 m^2 Core - 0.10 m^2	% cover - 0.25 m^2 Core - 0.0090 m^2
(2) Numbers of replicate subsamples per site at a given elevation	3 transects	3 transects	6 transects	5, 10, 10 (by level) for % cover; 10 for cores
(3) Numbers of study sites per category	2–6 randomly selected sites in each of 4 habitat types and 4 oiling levels (64 total sites)	In 1989 2–6 sites from each of 3 habitat types and 3 oiling levels (43 total sites)	3, 5, 4 and 2 pairs in PWS in exposed rocky, sheltered rocky, coarse textured, and estuarine habitat: 6½ sheltered rocky, 5 coarse textured, 2½ estuarine pairs in GOA (56 total sites)	10–18 sheltered rocky, 10–14 mixed soft and additional boulder/ cobble unanalyzed (22–30 total sites for the 2 habitats)
(4) Numbers of sampling dates	1	1	4	5[e]
(5) Total area sampled in a given stratum in a single year (e.g., sheltered rocky at mid elevation)	~1.36 m^2 (in 1990)	~9.6 m^2 (in 1989: includes scrape (1.1 m^3) % cover (8.5 m^2) on sheltered and exposed rocky)	~ 60 m^2 (in 1990: includes scrape (12 m^2) and % cover (48 m^2) and 2 dates)	~ 35 m^2 (in 1992)

[a] The stratified-random component (SRS) of Exxon's study in Prince William Sound (Gilfillan et al. 1995a, Page et al. 1995)

[b] The Exxon-sponsored study of the Gulf of Alaska shores (Gilfillan et al. 1995b)

[c] The study sponsored by the Exxon Valdez Oil Spill Trustees (McDonald et al. 1995, Highsmith et al. 1996, Stekoll et al. 1996)

[d] The study conducted by the NOAA Hazmat Program (Driskell et al. 1996, Houghton et al. 1996)

[e] This number refers only to the sampling dates of rocky shores from 1990–1992: soft-sediment sampling was reported for 4 dates, 1989–1992, and sampling of each has continued annually through 1999 (A. J. Mearns pers. comm.)

age more of the heterogeneity than the other scrape samples, thereby reducing error variance and enhancing power. The larger samples used for estimating percent cover by CHIA, NOAA Hazmat, and GOA studies would be expected to do an even better job of averaging across gradients and patches.

If sampling is being done to estimate a community property such as species richness or species diversity from each sample, then small sample size can also be problematic by leading to greater variance among samples (Pielou 1966). Such community response variables were calculated and reported in SEP (Gilfillan et al. 1995a), where areal coverage of both epifaunal (scrapes) and infaunal (cores) samples was the smallest of the 4 studies, and in GOA (Gilfillan et al. 1995b), where areal coverage of scrape samples was intermediate among studies and core samples as small as in SEP (Table 1). CHIA did not report species richness and diversity, but NOAA Hazmat, which used larger areal cover samples than the scrape samples used in SEP and GOA, did analyze and report such community responses (Houghton et al. 1996). The NOAA Hazmat samples for epibiotic cover would be expected to have yielded lower error variances for species richness and diversity than the smaller scrape samples of SEP and GOA, but because of reduced detail in visual estimates the NOAA Hazmat epibiotic samples probably missed some of the rarer species. The differences among studies in the surface areas covered by the infaunal cores (Table 1) are small and probably do not affect the power of statistical inferences on soft-sediment biota greatly.

(2) Sample replication

Replication within a site was achieved by collecting subsamples along replicate vertical transects for each of the 4 intertidal resource assessment studies. The numbers of these replicate transects (and thus the numbers of replicate subsamples per site at any given elevation) varied among studies by a factor of 2 to 3 (Table 1). SEP and GOA had the fewest replicates (3), while CHIA used 6 and NOAA Hazmat 5 to 10, depending on elevation. Especially as it approaches 1, the number of replicates has important consequences on the standard error around estimates of mean densities of individual species and on the adequacy with which the community is characterized. By this criterion alone, power to detect effects of the oil spill on densities of species at specific sites would be lowest for SEP and GOA, greater for CHIA, and greatest for NOAA Hazmat. To evaluate the effect of replication within sites on estimating species richness, we used the data from actual samples to compare the SEP results to the NOAA Hazmat results. Using all the 1990 sheltered rocky mid-intertidal sites, on average 3 replicates of the 0.031 m^2 scrapes from SEP contained 10.75 taxa, with new taxa added in the third replicate averaging 21.5% of the total. In contrast, using the NOAA Hazmat sample results for this same habitat but employing 10 replicate 0.25 m^2 quadrats, an average of 17.3 taxa was obtained and new taxa added in the last (the tenth) replicate averaged only 3.0% of the total. Thus, the NOAA Hazmat sampling was achieving a much more complete representation of the community composition than the SEP study. Because area covered per sample and replication both differed between the 2 studies, this contrast confounds the effects of size of the sampling unit and replication of samples (number of transects). To separate these effects, we resampled the NOAA Hazmat sample results by randomly selecting 3 quadrats (the same number as in SEP) in 20 separate randomization runs. On average, 12.0 taxa were obtained, with new taxa added in the last replicate averaging 10.6% of the total. This comparison demonstrates that both small size of the sampling unit and low replication contributed to the poor representation of the community membership, but that replication had the greater impact on this response variable. By missing more (generally rarer) taxa in the SEP sampling, the power to detect oiling effects on species richness would probably be lower except in the unlikely case where oiling affected species richness by eliminating common (readily detected) but not rare species.

(3) Numbers of study sites per category

The 4 studies of impact to intertidal biota used different numbers of study sites in their assessment designs, although within a given habitat type and oiling treatment category, numbers of replicate sites were surprisingly similar (Table 1). The SEP study of Prince William Sound used random numbers to select 64 study sites, more than any of the other studies (Page et al. 1995). These study sites represented a mean of 4 (range of 2 to 6) replicate sites in each of 16 cells, defined by habitat (4 types) and oiling level (4 types). CHIA assessed species densities at 56 study sites (Stekoll et al. 1996), of which 28 were located within Prince William Sound and the other 28 within the other 2 geographic areas (Kenai Peninsula - lower Cook Inlet and Kodiak archipelago-Alaska Peninsula). Within Prince William Sound, the CHIA design involved an average of 3.5 (range of 2 to 5) replicate pairs of study sites for each of 4 habitat types, where each pair involved contrast of 2 oiling levels (moderate–heavy vs light–no oil). In the Gulf of Alaska areas outside the sound, CHIA consisted of 3 to 7 pairs of study sites within each of 3 habitat types, again using the pairing to contrast moderately-to-heavily oiled shores with lightly oiled or unoiled shores. GOA included 43 replicate study sites in 1989, when biological data were collected (Gilfillan et al. 1995b). These study sites represented a mean of 4.8 (range of 2 to 6) replicates for each cell of a matrix of 3 habitat types and 3 oiling levels. The NOAA Hazmat study in Prince William Sound employed the smallest total number of study sites (22 to 30, depending on year); however, these covered only 2 habitat types for which results have been reported. Because the design within each habitat involved a contrast of 3 treatments (oiled-untreated vs oiled + treated vs unoiled controls), the average number of replicate study sites for this NOAA Hazmat program was 3.3 to 6 for the sheltered rocky habitat and 3.3 to 4.7 for the mixed soft-sediment habitat (Driskell et al. 1996, Houghton et al. 1996).

On the sole basis of this criterion of replication at the level of study site, it is essentially impossible to order the different studies by power to detect impacts. Within a given jointly sampled habitat type and oiling treatment level, replication hardly varied among studies. However, the numbers of oiling treatments and habitats fully sampled did differ. Thus, GOA sampled more total sites (43) than the CHIA component in the Gulf of Alaska region (28). Furthermore, SEP sampled more sites (64) than the CHIA Prince William Sound component (28), and within the sheltered rocky and mixed cobble soft-sediment habitats, the NOAA Hazmat study had site replication equal to or greater than that of SEP (Table 1). In comparing study designs, the intrinsically interesting issue of how many oiling treatment levels to establish arises. In cases where the shape of the possible relationship between treatment level and response is unknown, Cox (1958) recommended using none, medium, and high as the generally most efficient design. None of the studies followed that advice. Although SEP used 4 oiling categories in its design, one can show that CHIA with only 2 oiling categories and half the total number of sites in a given habitat would be nearly as efficient in detecting the effect of oiling and with equal total numbers of study sites could have been more efficient, assuming that light oiling and unoiled controls truly did not differ and that medium and heavy oiling did not differ. (On the other hand, if *a priori* reasons suggested importance for distinguishing the magnitude of impacts at different levels of oiling, a design comparing only medium-to-heavy oiling with light-to-no oiling fails even to pose the required contrast.)

To demonstrate that inclusion of treatment categories that do not explain any of the variance can actually lower power to detect the effect of the treatment, assume 4 oiling categories and consider the following hypothetical example. The no oiling and light oiling both have a true mean (u) of 20 for the response variable, while the medium and heavy oiling categories have the identical u of 15. Assume that each category has a standard deviation (sigma) of 5 and site replication of 16. Then, under an alpha of 0.05, simple computation of the power to reject the hypothesis of no oiling effect for alternative 1-factor ANOVA designs that do or do not subdivide those oiling categories that truly do not differ reveals that: (1) for the design that distributes its 64 sites among the 4 oiling categories, power = 0.92 (expected $F(3,60) = 5.33$ with a p-value, assuming no oil effect, of 0.0025); while (2) for the design that uses only half as many sites (32) but specifies and contrasts only the 2 (truly different) oiling categories, power = 0.87 (expected $F(1,30) = 8$; $p = 0.0083$). If in this latter design, 64 sites were employed and assigned to the 2 oiling categories, power = 0.99 (expected $F(1,62) = 16$ with a p-value, assuming no oil effect, of 0.00017). Thus, allocating sampling effort among categories that do not differ incurs a cost in efficiency of the resulting assessment design. Although samples of petroleum hydrocarbons in sediments have demonstrated that many control sites were actually exposed to some oiling (Jewett et al. 1999), implying that lightly oiled and reference sites were not likely to differ in general, and responses to heavy and medium oil rarely differed (Gilfillan et al. 1995a), we do not know that the assumptions used in this demonstration approximate the realty of the spill responses. Consequently, we cannot conclude that the CHIA and SEP designs for Prince William Sound necessarily achieved nearly equal efficiency through their site allocation, but we similarly cannot conclude that replication at the site level created substantial differences in detection power among studies. Nevertheless, the calculations of relative efficiency of alternative designs emphasize the wisdom of Cox's (1958) design advice.

(4) Numbers of sampling dates

The number of dates on which sampling is conducted and the time period encompassed by the sampling design have important consequences for the ability to detect and quantify oiling effects and to infer recovery trajectories (Stewart-Oaten et al. 1986). The 4 large studies of oil spill impacts on intertidal biota differed greatly in numbers of samplings (Table 1). SEP provided biological information for only a single 1-time (summer 1990) sampling, whereas CHIA employed 4 sampling dates over a period of 2 summers (1990 and 1991) and NOAA Hazmat sampled rocky shore sites on 5 occasions over 3 yr (1990 to 1992). Limited sampling of sites for CHIA occurred for 3 more years through 1994 (Stekoll & Deysher 1996) and the NOAA Hazmat study has been continued annually through summer 1999 (A. J. Mearns pers. comm.). GOA sampled biota in each of 2 years (1989 and 1990) but did not resample the same sites with the same methods in 1990, preventing meaningful temporal contrasts and resulting in availability of biological data for only the single summer of 1989. Sampling for all studies was conducted during a single season, summer, so that no study was able to evaluate the possibility of seasonal variation of oil spill effects (which may be important for such dynamic variables as algal cover). By sampling during both early and late summer in 1990 and 1991, CHIA could make some inferences about seasonal dynamics. In 1989, NOAA Hazmat sampled 1 set of sites in April, May, July, and September (Houghton et al. 1996), providing the best (but limited) seasonal sequence of any study.

Such differences in numbers of samplings have several important consequences. First, repeated sampling has the implication of reducing the degree of uncertainty in inferences about impacts of the oil spill simply because of the likely additional power from increasing replication that is provided by the additional sampling effort. At a minimum, multiple tests of effects could be conducted (1 for each sampling date, assuming no pooling of similar dates), with the power of each test essentially identical to the power of the single test available for a 1-time sampling. Moreover, with repeated observations on the same sites over time, estimation of time trends will be possible and site differences more precisely estimated with repeated measures ANOVA (Kuehl 2000).

The potential for an interaction with date represents a second grounds for preferring multiple sampling dates. This allows treatment × time interactions to be tested in statistical analyses and, even more importantly, provides estimates of how the differences between oiled and controls sites change through time. Such interactions provide insight into many processes that cannot be evaluated by a single sampling, including most importantly an evaluation of whether recovery has been initiated in the biota and what the time course to complete recovery is estimated to be. A detrimental effect of an environmental perturbation is much more serious biologically if it lasts a longer period of time, so repeated sampling to document recovery with confidence is a critical component of a good assessment design. For example, Barber et al. (1995) demonstrated that the abundance of intertidal fish was still significantly reduced by about 50% on oiled shores in 1990: by sampling again in 1991, they showed by exploring the significant time × oiling interaction that recovery in total intertidal fish abundance was nearly complete by 2.5 yr after the oil spill.

Third, repeated sampling allows detection of any delayed effects: delays in effects are an expected outcome of indirect effects operating within a community, such as trophic cascades (Schoener 1993, Menge 1995). NOAA Hazmat detected a delayed effect of the oil spill on the rockweed *Fucus gardneri*, the major provider of biogenic habitat in the upper and mid intertidal of this system: after years of convergence in *F. gardneri* abundances on oiled and unoiled shores, a large fraction of the population died at oiled sites but not on control shores, perhaps because the plants on the oiled shores were dominated by individuals from a single post-spill age class which all senesced together (Paine et al. 1996, Houghton et al. 1997).

Finally, repeated sampling allows testing of the necessary assumption of an observational damage assessment design that lacks pre-spill data, namely that the oiled and reference sites would possess similar biotic communities with similar species abundances in the absence of the intervention caused by the spill. In its strictest form, this assumption is almost certainly false: the oiled sites are not likely to be identical to the unoiled sites or else the physical processes transporting the oil would not have treated them differently. It is possible that the oiled sites over some period of time, decades perhaps, would resemble the control sites in important biotic parameters, but a single 1-time sampling is missing an entire component of variance to assess this and to correct estimates of oil spill effects for the degree of violation of this assumption. The magnitude of difference and variance over time in the differences between oiled sites and unoiled sites, absent the effects of oiling, are needed to assess whether such natural differences are not as big as the presumptive effect of oiling estimated from a single sampling. By having a time series of assessments at the oiled and reference sites, this weak assumption of identity of undisturbed communities is testable by evaluating the degree of ultimate convergence of oiled and reference sites (Skalski & Robson 1992, Barber et al. 1995, Highsmith et al. 1996) and by using differences and their variances between oiled and reference sites in a time period absent of oil effects to assess the significance of the presumptive oil effects and make necessary adjustments in its magnitude. This is analogous to the BACI (Before-After-Control-Impact) designs of Stewart-Oaten et al. (1986) except that the natural variation between those sites used for the treatment and those used for controls is evaluated not before the perturbation but instead after recovery has occurred. Because the oil seemed to preferentially strike sites that had greater current flows, and thus greater probabilities of encounter with oil, and because greater current flows imply higher fluxes and greater potential settlement of invertebrate larvae and algal spores, it is possible that all 3 assessment studies of injury to intertidal biota underestimated the effects of the spill because control sites were not chosen to represent locations of high current flow (Peterson 1993, Highsmith et al. 1996). Further temporal sampling can evaluate this possibility and other presently uncontemplated reasons for oiled and control sites to differ naturally. In summary, the differences in numbers of samplings among the assessment studies imply further disparity in power to detect and characterize oiling effects, with SEP and GOA lacking the power that temporal replication likely brings to the other studies, lacking the ability to estimate the time course of recovery, unable to assess any delayed effects of oiling, and unable to evaluate the degree to which oiled and reference sites represent a good match.

(5) Total area sampled

Although the importance of total area sampled within a habitat cannot be fully appreciated without knowledge of how effort was distributed within and among samples, transects, sites, and dates, computation of this simple metric helps illuminate intrinsically large differences among studies in their effort and in the consequent reliability of their conclusions. For any given elevation on shore and jointly sampled habitat type, the sampling effort conducted in Prince William Sound covered dramatically less shoreline area in SEP than in CHIA or the NOAA Hazmat studies (Table 1). For example, the total area sampled in mid-elevation on sheltered rocky shores was 1.36 m^2 in SEP in 1990 (Gilfillan et al. 1995a, Page et al. 1995). This represents contents of 44 scrape samples, each of a size of 0.031 m^2. In contrast, NOAA Hazmat sampled a total area of about 35 m^2 of this habitat in a single sampling of 1992 (ten 0.25 m^2 quadrats at 14 sites; Houghton et al. 1996) and CHIA sampled about 60 m^2 (not including the nearest neighbor samples) of this habitat in Prince

William Sound in 1990 alone (sixty 0.10 m^2 scrape and sixty 0.40 m^2 areal cover samples on each of 2 dates; Highsmith et al. 1994, 1996). If analogous calculations were made pooling over all sampling dates, disparities among studies would grow much larger because the SEP study of biotic response variables was conducted on only a single date, while the others sampled on multiple occasions. GOA sampled about 9.6 m^2 of combined sheltered (4.0 m^2) and exposed (5.6 m^2) rocky habitat at the mid-tidal level (Gilfillan et al. 1995b), 3.5 times the area sampled by SEP in these 2 habitats pooled but far less than the area of these habitats sampled by CHIA or NOAA Hazmat. The GOA sampling consisted of eighteen 0.063 m^2 scrape samples plus thirty-four 0.25 m^2 areal cover quadrats. The most important consequence of such differences in total sampling effort is a large disparity in reliability and confidence in the estimates of injuries derived from the separate studies. This contrast of studies in areal coverage of epibiotic sampling involves 2 different types of samples, scrapes and cover estimates. Scrapes provide more detailed information on organism abundance, including small organisms, and biomass, if weights are taken. Areal cover estimates are cruder estimates that can be feasibly made over wider areas. But even within each of these 2 types of samples, GOA and SEP used substantially less effort (total areal cover) than CHIA for scrapes and GOA used less effort than CHIA and NOAA Hazmat for areal cover samples, even in a single sampling (Table 1).

Issues of sampling design

(6) Philosophical support for targeting putative affected areas

The initial (often trivial) objective of an environmental impact study is often to test the null hypothesis that there was no impact of the event in question, here the oil spill and associated shoreline treatments. Failure to reject this null hypothesis should be done with adequate power to have detected any biologically significant impact. The subsequent (more meaningful) objective is to estimate the magnitude of impacts and the time course of recovery. Thus, evaluating the biological significance (in one sense) as opposed to the statistical significance of demonstrated differences is a necessary part of the philosophical basis of injury assessment. The 2 stratified random sampling designs possessed contrasting underlying philosophies despite their general similarity (Table 2). SEP was designed to permit extrapolation to the entire 'affected area' of Prince William Sound (Page et al. 1995). The sampling frame included lightly and very lightly oiled shores, thereby diverting effort and potentially compromising power (see above) by including a category in the categorical ANOVA or ANCOVA design that may not be expected to differ much or at all from unoiled controls. In addition, much of the sampling effort was expended in the dominant high-energy habitat, where oil is more likely to be removed by wave action, where vertebrate consumers are less able to forage, and where environmental sensitivity is presumed to be lower (Teal & Howarth 1984, NRC 1986). Nevertheless, the presumption of low sensitivity of high-energy environments is arguable, justifying assessment effort in them. Injury from the oil spill to benthic resources may be just as serious in wave-beaten rocks, even if natural, physically forced clean-up occurs more readily, intrinsically high productivity speeds recovery, and higher-level predators have limited access to benthos. The CHIA study, in contrast, incorporated only a single contrast of oiling categories and concentrated efforts in presumably more sensitive habitats with lower physical energy (Sundberg et al. 1996), including the 'estuarine' soft sediment marshes that are relatively rare in the spill region but known to be highly sensitive (Teal & Howarth 1984). These allocations of effort reflect different philosophies of assessment (Gilfillan et al. 1996). Ideally, necessarily limited resources should not be squandered either by devoting extensive effort to sampling even abundant habitats with low sensitivity or by oversampling rare but sensitive habitats in hopes of detecting small but biologically unimportant differences. However, sensitivity is little more than an informed guess and biological importance is often a value judgment. There is another related implication of assessment philosophy that influences detection of injuries. By sampling only the most abundant habitats and assessing the community of species so defined by that sampling choice, SEP failed to focus on response variables chosen *a priori* to be likely to be affected. In contrast, CHIA included a study of subtidal eelgrass habitats and focused on sampling techniques appropriate to assess abundances of amphipods and echinoderms (Dean et al. 1996, Jewett et al. 1999), taxa known to be sensitive to oil toxicity (Warwick & Clarke 1993, Peterson et al. 1996). Thus, fundamental philosophical decisions about the sampling frame and the response variables to measure differed between the 2 stratified random sampling studies in ways that made the SEP results more applicable to the most common, but perhaps least sensitive habitat and the most common oiling category (light oiling) and the CHIA study more likely to identify effects of the oil spill by targeting its sampling efforts on more sensitive, but less common habitats and biota (Gilfillan et al. 1996).

Table 2. Differences among studies of injury to the intertidal biota after the 'Exxon Valdez' oil spill in sampling design. PAH: polycyclic aromatic hydrocarbon

Issue	SEP	Exxon GOA	CHIA	NOAA
(6) Philosophy of sampling affected area	Effort devoted equally to 4 selected habitats (exposed bedrock, sheltered bedrock, boulder/cobble, pebble/gravel)	Effort devoted equally to 3 selected habitats (bedrock, boulder/cobble, pebble/gravel)	Effort devoted equally to 4 selected habitats including fine sediments (exposed rocky, sheltered rocky, coarse textured, estuarine)	Effort devoted equally to 3 selected habitats (sheltered rocky, exposed boulder/cobble, mixed soft)
(7) Random site selection vs matched-pair design	Random selection of both oiled and reference sites within habitat	Subjective choice of oiled and reference sites	Random selection of oiled sites then matched to reference sites	Subjective choice of oiled and reference sites
(8) Sampling frame	Extended to include reference sites in SW areas near mainland glacier ice-melt	Sites subjectively chosen to cover shoreline along spill path	Reference sites paired with oiled sites by geographic proximity, freshwater influence, and 4 other factors	Sites subjectively chosen for known treatment history and coverage of the spill area in PWS
(9) Treatment of habitat heterogeneity within sites	Systematic spacing of transects did not exclude secondary habitat	Subjective site selection provided some homogeneity	All secondary habitat excluded	Subjective site selection provided some homogeneity
(10) Interspersion of sites	Poor interspersion in some cases because low replication (2–6) and random selection makes accidental overweighting of areas probable	Sites chosen in part to achieve even coverage and dispersion	Pairing (blocking) helps guarantee good interspersion of oiled and reference sites	Sites chosen in part to achieve interspersion
(11) Controls for shoreline treatment and oiling intensity	Shoreline treatment uncontrolled but design employed 4 levels of oiling and a measured sediment PAH covariate	Shoreline treatment uncontrolled but design employed 3 levels of oiling	Shoreline treatment uncontrolled and only 2 levels of oiling in design (with very light oiling included in reference sites)	Designed specifically to address shoreline treatment effects vs oil effects

(7) Random site selection versus matched-pair design

The 2 stratified random sampling designs differed fundamentally in their underlying approach to removing the intrinsic bias associated with an observational study (Table 2). SEP chose for each habitat both oiled (3 levels) and reference sites at random from a GIS (Geographical Information System) frame of all possible sites of sufficient length within each oiling category, followed by ANCOVAs. CHIA used a design in which only oiled sites were chosen at random from a GIS frame of all heavily and moderately oiled sites of sufficient length within each habitat and geographic area. Then, every oiled site was paired with a matching reference site, with matching done on the basis of geographic proximity, beach slope, wave exposure, substrate composition, nearshore bathymetry, and proximity to sources of freshwater. This control-treatment paired design of Skalski & Robson (1992) attempts to control for multiple important sources of error in estimation of treatment effects. In the SEP design, attempt was made to control for other confounding factors by measuring site covariates (wave energy for epibiota in scrape samples; and sediment size, total organic carbon, and wave energy for infauna in cores) and using them in ANCOVAs (Page et al. 1995). If the relationships of the response variable to each of the covariates is linear and parallel, and the distributions of the covariates are symmetrical, this regression approach is generally a superior means of controlling bias (Cochran & Rubin 1973). On the other hand, pairing (blocking) methods are generally more efficient, thus enhance power more, and are effective even in the case of complex, non-linear relationships between dependent and independent variables (Cox 1957).

Neither of these alternative approaches to controlling error is free of assumptions, most of which were probably not fully satisfied. ANCOVA assumes that the effects of the covariates and resulting residual errors are correctly modeled. The blocking approach makes assumptions about which variables are important to control. Levels of matching variables are not known without error and uncertainty. No 2 sites are identical in the full suite of matching variables, so trade-offs are necessary in choosing the paired control sites. However, the use in CHIA of multiple matching variables, including spatial proximity, which is a proxy for other unmeasured and uncontemplated confounding factors, to achieve a blocked design represents application of a widely applied and powerful method for controlling natural heterogeneity and enhancing power in environmental assessments (Skalski & Robson 1992). The philosophical approach of SEP and GOA explicitly contends that natural variation among reference sites should be a part of error against which to gauge and test environmental impact. In analyses of epibiota on rock surfaces, SEP used only 1 covariate, wave energy, instead of the 6 matching factors considered in CHIA, probably removing less of the natural site variation from the error variance and thus not enhancing power to the same degree. In analyses of infaunal invertebrates in cores, SEP employed 4 (3 independent) covariates but these were measured after the oiling treatment and 1 of those (organic content) was likely itself influenced by the treatment. Under that condition, the adjustment for the covariate removes part of the treatment effect (Fairfield Smith 1957, Winer 1971). The GOA and NOAA Hazmat studies, in which core samples were also analyzed, used neither pairing nor ANCOVA to control for natural variation. The best general means of controlling error variance is blocking combined with regression adjustment for covariates (Cochran & Rubin 1973), but none of the studies adopted this mixed strategy. For example, the CHIA design could have been improved by using covariates measured in each quadrat to further control for natural variability (Sundberg et al. 1996).

(8) Sampling frame

To employ a fully randomized observational design like that of SEP for selecting both oiled and control sites, one must clearly and appropriately define the sampling frame from which those site selections are to be made. Both the SEP and the CHIA designs used a frame that included only those sites where shoreline segments equaled or exceeded 100 m (or 60 m in 1 habitat) and eliminated other sites from consideration (namely certain dangerous sites in CHIA and those near eagle nests in SEP), which limits the ability of both studies to extrapolate to the complete universe of oiled shores. However, the use of random selection procedures to identify the control sites, while sounding unbiased in principle, poses a serious problem. The oil did not strike shorelines at random but instead struck certain geographic areas and certain environments preferentially. Specifically, within Prince William Sound the oil beached largely on the islands rather than the mainland and probably more frequently encountered shores with greater current flux. Consequently, if the oil did not strike shores at random, the selection of reference sites should ideally mimic the selectivity of the oiling process (Peterson 1993).

This selectivity in the oiling process represents a substantial challenge to any assessment design, but recognizing the issue allows some of the most serious errors to be avoided. The fully randomized SEP design did not adequately limit selection of reference sites to the island shores towards the eastern side of the sound to correspond to the selectivity of the oiling process. The most serious consequence of this inclusion of sites near the mainland was to allow sites that were already impacted by mainland run-off from low-salinity, high-turbidity ice melt to be included, when such shores had low probability of inclusion in the oiled universe. For example, in SEP 3 of 4 reference sites for the pebble/gravel habitat fell on the relatively unproductive southwest shores of Prince William Sound in close proximity to glaciers (the Bainbridge area and Whale Bay; Page et al. 1995). The field data from surveys of 2 of these imply strong evidence of impacts of ice melt on the intertidal biota in the form of gross biotic impoverishment. Inclusion of such inappropriate controls arising from an unjustified sampling frame can lead to biased estimates of the impacts of the oil spill: in this case, effects are biased downwards, helping to explain the contradiction between the SEP conclusion that the oil spill enhanced diversity and abundance of the intertidal biota in Prince William Sound (Gilfillan et al. 1995a) and the opposite results of the other 3 studies. The CHIA study made this same error in its initial pilot year of 1989, when it failed to recognize the bias for oiling of island shores and had not yet collected comparative field data to show how seriously impoverished shorelines exposed to glacial runoff are. That study was subsequently redesigned with the paired approach to overcome the huge problem of identifying the precise selectivity of the oiling process so as to define properly the sampling frame. Matching by geographic proximity, proximity to freshwater sources, and local bathymetry was intended in CHIA to control for the most serious departures from randomness in the process of oil beaching. Nevertheless, the protocol for matching sites must be unbiased in the sense that the suite of other factors besides oiling that influence biological response variables cannot systematically favor the oiled or the reference sites. Thus, the optimal design is one that combines blocking (stratification) with randomization in some fashion, such as the CHIA protocol. With more resources available for damage assessment, this basic study design could have been further enhanced by random selection of more than 1 matching site from a pool of candidate reference sites (e.g., Underwood 1994). NOAA Hazmat avoided choosing reference sites in those eastern areas of the sound where glacier inputs would confound interpretation, yet the failure to use random selection for any site categories limits the ability to extrapolate results. Similarly, GOA did not use any explicit randomization process to select sites, although those chosen do not reflect any obvious bias relative to geographically based confounding factors.

(9) Treatment of habitat heterogeneity within sites

Although misclassified sites were excluded from sampling in the stratified random studies of SEP (Page et al. 1995) and CHIA (McDonald et al. 1995, Sundberg et al. 1996), many sites that were selected were heterogeneous to such a degree that the site contained a mix of multiple geomorphological habitats. This was a consequence of the original site categorization being done at a coarse spatial scale. CHIA handled this problem by excluding from consideration any stretches of different habitat types, sampling only in the primary habitat type for that site (Highsmith et al. 1996). This decision represents another way in which the CHIA sampling protocol failed to live up to its purported goal of pure random sampling that would allow extrapolation to the entire spill region. SEP apparently included such secondary habitat types within their sampling frame if a systematically located transect fell on a geomorphologically different habitat (Page et al. 1995). This decision permits more rigorous extrapolation of results to the entire spill area. However, the CHIA procedure has the consequence of reducing variances among replicate transects and samples, thereby decreasing error variance and enhancing power to detect impacts of the oil spill. The level of control of within-site heterogeneity among samples was presumably intermediate, somewhere in between SEP and CHIA, for both GOA and NOAA Hazmat. GOA and NOAA Hazmat selected sites subjectively, thereby achieving some control over heterogeneity within sites but GOA located replicate samples by a systematic algorithm (Gilfillan et al. 1995b) and NOAA Hazmat randomly (Houghton et al. 1996). Neither of these fixed-site studies identified a procedure for excluding samples from secondary habitat types within sites.

(10) Interspersion of sites

Another problem that can arise in the selection of sites by randomization is a failure to achieve adequate interspersion of sites, thus creating an unbalanced and confounded representation of the relevant spill region if site replication is low (Hurlbert 1984). Interspersion is actually just one aspect of a broader problem of the potential for extreme outcomes created by fully randomized sampling designs (Cox 1958). Because error distributions include the possibilities of extreme outcomes, a failure to stratify over all possible important parameters (geographic area here) leads to a more inefficient design in which relatively high error variance incorporates variation among strata that could be factored out in a stratified design. It is for this reason that Box et al. (1978) recommend experimental designs that 'block what you can: randomize what you cannot'.

Initial stratification by broad geographic region induced interspersion on a large spatial scale in all the 'Exxon Valdez' oil spill assessments. But on the smaller scale within Prince William Sound, fully randomized site selection created strongly unbalanced site choices on occasion in SEP (Table 2). For example, 3 of 5 exposed bedrock reference sites in SEP fell along the same shore of a single small island (Perry Island; Page et al. 1995). This problem in poor interspersion (confounding) by overweighting the Perry Island shores in the controls for the exposed bedrock habitat implies a need to reassess the statistical test results conditional on the extreme outcome produced by the randomization. Specifically, this observed spatial confounding implies that the actual probability of a biologically extreme outcome is higher than the nominally computed probability. This creates greater uncertainty in estimation of the oil spill impacts. A possible solution to this problem would have been re-randomization, especially if what is considered an extreme order was established *a priori* (Cox 1958). This procedure would have reduced bias at the expense of inaccurate estimation of error, which is poorly estimated by small numbers of sampling sites anyway. The pairing of oiled and reference sites in CHIA achieved interspersion of oiled and reference sites within geographic region, but here too the use of randomly selected oiled sites has the potential to over- or underweight certain areas of the sound. Systematic selection (Thompson 1992) of study sites within the basic strata would have yielded better spatial dispersion in all studies, but there are serious potential problems with systematic sampling if a periodic pattern of some sort is present (Fisher 1971). Much finer stratification would be the preferable means of increasing efficiency of the sampling design and avoiding spatial confounding of study sites; however, if the range of site choices is narrowed too much by fine spatial stratification, the resultant design approaches a systematic design (Cox 1952). Outside Prince William Sound, the CHIA study stratified by treating the 2 geographic areas separately, Kenai-lower Cook Inlet and Kodiak-Alaska Peninsula, whereas GOA pooled results from these 2 geographic areas. Given the substantial differences in response between these areas demonstrated in CHIA (Highsmith et al. 1996, Stekoll et al. 1996), this additional level of stratification in CHIA controlled for a large source of geographic variability, thereby producing greater power to detect (regionally different) responses to the oil spill. Despite the absence of formal stratification in the design, some degree of interspersion of study sites was achieved intentionally in NOAA Hazmat and especially in GOA by a desire to cover the geographic area of the spill.

(11) Controls for shoreline treatment and oiling intensity

In both the CHIA and SEP studies, important variation existed in the level of oiling and the subsequent shoreline treatments that is not controlled for in the statistical design (Table 2). The short- and long-term evaluations of impacts of various shoreline treatments by NOAA Hazmat (Houghton et al. 1996, Lees et al. 1996) show convincingly that shoreline treatment greatly reduced abundances of the rockweed *Fucus gardneri*, the primary provider of shelter and biogenic habitat in the intertidal zone, and of most of the common hard- and soft-substrate invertebrates of the intertidal shores. Yet because of incomplete record keeping, neither CHIA nor SEP nor GOA was able to use type and intensity of shoreline treatment as a covariate or as a stratification factor in the design of the assessment sampling. To the degree that post-spill shoreline treatments augmented injuries caused by oiling alone (Houghton et al. 1996, Lees et al. 1996), confounding of the 2 processes may have enhanced ability to detect impacts of the oil spill, although causing uncertainty in attribution to cause. The ability to separate the effects of oiling from subsequent treatment is the main strength of the assessment design used in NOAA Hazmat, made possible by selecting shorelines before treatment and where reliable treatment records could be maintained. A random selection of study sites would not have afforded this capability.

The intensity of oiling represents a variable that was not well controlled in any of the studies. Sites were categorized by the width of the oil layer initially observed on the intertidal shore, but within a shoreline segment, and especially among elevations on a shore, substantial heterogeneity in oiling existed. A heavy layer of oil deposited high in the intertidal does not imply that oiling was also heavy lower on shore, thereby inducing additional unexplained error variance. In addition, the oiling intensity classification was based on initial oiling even though the oil was remobilized and redeposited in many sites over time. While older oil was probably less toxic chemically, its ability to harm intertidal invertebrates by smothering and other physical mechanisms would remain intact. The failure to control for this variation in oiling intensity renders the assessment designs less able to detect impacts of the spill. SEP did the better job of stratifying by 4 levels of oiling in the selection of the sites for the stratified random component of the study. That study also measured a PAH (polycyclic aromatic hydrocarbons) covariate to help adjust for varying levels of oil in the sediments at its infaunal coring sites (Page et al. 1995). GOA employed 3 levels of oiling in its stratification of sites (Gilfillan et al. 1995b). CHIA, in contrast, selected oiled sites from those that were either heavily or moderately oiled and used some lightly oiled sites as controls (Sundberg et al. 1996). The consequence of this failure to stratify by oiling intensity in CHIA is unclear. If finer oiling classifications reduce otherwise unexplained error variance, then the CHIA pooling of oil categories would have reduced power to detect oiling effects. On the other hand, if both differences between reference sites and lightly oiled sites and differences between heavily and moderately oiled sites are slight (as implied by results of SEP; Gilfillan et al. 1995a), then the simpler CHIA design would be more powerful (see above).

Issues of analytical methodology

(12) ANCOVA with covariate affected by treatment

The fundamentally different sampling approaches used by the suite of studies of the oil spill impacts on intertidal biota implied different abilities and methods to control for environmental variability. Because random selection of both oiled and reference sites for SEP retained substantial uncontrolled variation among sites within treatment categories, covariates were used in ANCOVA to remove some natural environmental variation. For scrape samples, wave energy served as a single covariate and for core samples of infauna, sediment particle size (interdependent % sand and % silt/clay measures), wave energy, and total organic carbon were measured and used as covariates (Page et al. 1995). CHIA used pairing of reference sites with randomly chosen oiled sites to control natural among-site differences in important forcing factors (Sundberg et al. 1996). The covariates used in SEP were undoubtedly measured with error, which in itself reduces power. For example, in the limit, if the covariate has sufficiently high error as to be random, then ANCOVA does no more than reduce the effective replication. Additionally, there existed correlation between both total organic carbon and oiling level and also between wave exposure and oiling level (Page et al. 1995). When a covariate is measured after imposition of the treatment, is affected by the treatment, and represents one of the mechanisms by which the treatment influences the dependent variable, then a basic postulate of ANCOVA is violated (Cox 1958). In that case, some of the effect of the oiling treatment is captured by the covariate term and factored out as if it were merely part of the background variation among sites (Fairfield Smith 1957). Organic content of sediments is likely increased by oiling and is also likely to affect sediment invertebrates (Spies et al. 1988) indirectly by augmenting food resources (a positive influence) and by enhancing sediment oxygen demand and sulfide production (a negative influence). Consequently, while use of covariates to remove uncontrolled variation represents a laudable goal, in practice, the properties of the SEP data set, with error in the measures of the covariate and at least 1 covariate responding to treatment, have the effect of masking impacts of oiling and making impact detection more difficult (Table 3).

CHIA controlled for variation due to background environmental variation to the degree that the pairing process matched sites with similar levels of the 6 factors that were used in the matching. As discussed above, perfect matching against 6 factors simultaneously is impossible and clearly occurred imperfectly. Any interaction among factors that is not captured in the matching process retains unexplained variability in the error variance, analogous to failure to include interactions in regression models describing effects of covariates. An important factor that probably varies systematically between oiled and reference sites in the CHIA design was the current flux, which appears to have been generally greater at oiled sites (Highsmith et al. 1996). None of the injury assessment studies effectively controlled for variation in current flux, although the geographic proximity factor used in matching by CHIA would be expected to have incorporated some of the geographic variance in flow. The failure to control completely for flow differences in all the assessment studies would lead to conservative estimates of injury because the intertidal biota is enhanced in productivity by exposure to increased flows (Leigh et al. 1987), implying that oiled shores should have exhibited naturally richer intertidal communities. The NOAA Hazmat and the GOA studies did not use any formal method to control for environmental covari-

Table 3. Differences among studies of injury to the intertidal biota after the 'Exxon Valdez' oil spill in analytical methodology. DCCA: detrended partial canonical correspondence analysis

Issue	SEP	Exxon GOA	CHIA	NOAA
(12) ANCOVA vs paired design	ANCOVA flawed by use of covariates that were estimated with error, one of which was affected by oiling	Simple ANOVA without control for environmental variation	Pairing by proximity, freshwater exposure, slope, bathymetry, sediments, wave exposure controls for environment	Simple ANOVA without control for environmental variation
(13) Pseudoreplication (in formulation of *F*-ratios)	Subsamples (transects) treated as if they were independent replicates in ANCOVA	Subsamples (transects) in 1 set of analyses treated as if they were independent replicates in ANOVA	Pairs compared by *t*-tests, then meta-analyses used to compute a joint p-value	Randomization ANOVA and *t*-tests using site means as the sampling units
(14) Inferring recovery	Done by comparing oiled points to the spread of unoiled points in a DCCA multi-variate analysis	Not attempted	Used simple convergence of densities between oiled and reference sites over years	Used simple convergence of oiled and reference sites over years
(15) Power analysis	Done by a pilot simulation of power using a philosophy that signals smaller than noise of natural among-site variation need not be detected	Not attempted	Actually computed power for each test based on observed variances and estimated effects magnitude	Not attempted

ates, rendering these studies equally poor at removing variance associated with environmental heterogeneity.

(13) Pseudoreplication

Pseudoreplication is defined by Hurlbert (1984) as 'testing for treatment effects with an error term inappropriate to the hypothesis being considered.' This problem most commonly occurs when treatments are not replicated, just subsampled, or when replicates are not statistically independent (Hurlbert 1984). McArdle (1996) has noted that this definition includes 2 issues, replication to eliminate confounding and independence of sampling units. He argued that the first of these concerns is valid but that independence of sampling units is a fallacy because what most statistical methods require is not independence of samples but of errors. Non-independent units can actually be exploited to better describe the true autocorrelated biological patterns and to adjust testing methods appropriately and efficiently (reviewed by Legendre 1993). Non-independent errors can even be accommodated by some statistical solutions (Searle et al. 1992).

In the SEP study, data from the 3 spatially confounded transects within each site were treated as if they represented separate independent sites in determining the significance and approximate power of tests in about 78 % of comparisons (Page et al. 1995). The transects are actually subsamples, not necessarily independent replicates: transects were spaced 20 to 30 m apart. Pooling sources of variation is usually done to increase error degrees of freedom in tests of hypotheses and thereby to increase the power of the tests. That was the intent of the SEP procedure (Table 3). However, the question of when it is appropriate to pool errors is controversial. It is always safest not to pool errors, and that decision seems especially appropriate in this application by SEP because of the proximity and systematic spacing of the transects. To its credit, SEP did employ tests of whether variances among sites differed from variance among transects before pooling, but the rejection criterion was set at 0.05 (Page et al. 1995) instead of the high alpha of 0.20 to 0.30 widely recommended to protect against low power of discrimination (e.g., Winer 1971). One possible reason for relatively high variance among the systematically spaced transects in SEP might be found in its apparent inclusion of transects or sample locations within transects that fell in small stretches of habitat that differed from the nominal habitat type. In contrast, CHIA used the subsamples to compare each matched pair of sites (McDonald et al. 1995) and then developed an overall test of significance across all site pairs using 2 types of meta-analysis (Fisher's procedure of combining p-values from independent tests and Stouffer's procedure [Folks 1984]). By treating the subsamples within sites as if they represented replicate sites, the analyses reported in SEP claim a higher power for their tests than may actually be present because the transects are spatially confounded and likely to be autocorrelated. The *F*-ratio contrasting among-site to within-site variances with degrees of freedom replication of 1 to 5 and 2, respectively, and an alpha of 0.05 as a means of testing for autocorrelation do not meet the burden of proof that Hurlbert's (1984) paper demands. GOA utilized both sites (true replicates) and transects (pseudoreplication) in separate analyses (Gilfillan et al. 1995b). NOAA Hazmat used sites as replicates in most analyses of long-term effects of the oil spill (Driskell et al. 1996, Houghton et al. 1996, 1997).

(14) Inferring degree of recovery

CHIA (Barber et al. 1995, Highsmith et al. 1996, Stekoll et al. 1996, van Tamelen et al. 1997) and NOAA Hazmat (Driskell et al. 1996, Houghton et al. 1996, 1997) used sampling over multiple sampling dates spread over 2 or more years to infer rate of recovery of injured resources in the intertidal habitat. GOA did not attempt to estimate degree of recovery (Gilfillan et al. 1995b). In contrast, SEP (Page et al. 1995, Gilfillan et al. 1995a) devised a method that purported to estimate degree of recovery from a single one-time sampling (Table 3). SEP used a multivariate statistical analysis (the canonical version of DCCA; detrended partial canonical correspondence analysis) to evaluate recovery of the intertidal community. The analyses fitted a DCCA model using all environmental variables, calculated a 95 % probability ellipse for all reference points in an arbitrary 2-dimensional space, and then defined oiled sites to be recovered if points representing those oiled sites fell inside the ellipse (Gilfillan et al. 1995a).

There are several problems both with this approach and its application. While there are useful roles for such ordination analyses (e.g., Field et al. 1982), especially in data exploration, they can often be abused in practice. The precise details of the analyses are commonly obscured within a computer program (Digby & Kempton 1987). In the SEP application, the axes are not explicitly specified and no attempt is made to figure out just what the axes are approximating or to follow up such data exploration with explicit tests of process related to physiological and ecological mechanisms (Gilfillan et al. 1995a). Given in addition that such methods are based on implausible assumptions of normality and represent approximations even then (Digby & Kempton 1987), much more rigorous evaluation of the meaning of the results is required (as in

Clarke & Ainsworth 1993). Even if a defensible multivariate analysis were conducted to form 'recovery ellipses,' there is a philosophical inconsistency. With only 1 point in time over a year after the spill, it is impossible to infer how much of the similarity between oiled and control sites is a consequence of a small initial impact and how much is a consequence of recovery from an initially large impact. In other words, rate of recovery and degree of recovery are impossible to distinguish from a single sampling even using 'recovery ellipses' because one could tell only how similar oiled and reference sites are, not how similarity has changed over time.

Several aspects of the SEP application of DCCA to estimating recovery are troublesome. First, all species occurring in less than 20% of the samples were removed before analysis. This removes a much larger fraction of the information than is the standard practice in conducting DCCA, where a 5% cut is typical. If 20% had to be removed, this suggests instability of the results of the DCCA. Sensitivity to small changes in species inclusions is common in data sets with limited numbers of sites, implying that there may be an intrinsic problem in applying DCCA to this particular data set. The partial justification provided, that a 20% cut produced stable GLIM (General Linear Interactive Model; Numerical Algorithm Group, Inc.) analyses, is unconvincing because the GLIM is a univariate analogue to ANCOVA with a quite different algorithm. (GLIM is unstable when large numbers of zeros occur in the data set.) Regardless of the issue of stability of the DCCA, the use of the 20% cut has the effect of reducing the information in the data and potentially reducing power to detect differences between oiled and reference sites. Second, the inclusion of a second, potentially non-significant axis has the effect of creating a large ellipse in 2-dimensional space that represents an easy standard for oiled points to meet because, as in most ordinations, the higher dimensions would be largely noise. On average, unexplained variance in the SEP DCCA analyses ranged from 65 to 81% across habitats (Gilfillan et al. 1995a), so the degree of (unreported) noise in the second axis must have been high. Third, the results suggest that oiling levels and the physical covariates each account for an average of only about 12% of the total variability (Gilfillan et al 1995a). Partial DCCA performs the ordination after removing effects of the other physical covariates. As discussed above, if oiling acts on the biota in part through its influence on a covariate, then some of the oiling effect is being removed before the analysis is conducted. Furthermore, the calculation of recovery ellipses based on this procedure ignores the possibility that other impacts from oiling may have occurred that were not well correlated with the chosen variables. Retention of unimportant variables that increase the error variance and absence of variables that may be related to oiling impacts have the likely effect of decreasing the power to discriminate differences between oiled and reference sites in this analysis. Fourth, the analysis of recovery ellipses maintained an alpha of 5% without showing that power to detect differences was retained despite this low type I error rate. Finally, this use of multivariate statistics to infer differences between oiled and reference sites requires that the oil spill impacts be detectable over and above the natural variation among sites in a design that did not stratify to remove among-site variation. This is the fallacy discussed earlier that maintains that an impact is insignificant if it is not substantially greater than natural variation (itself dependent on sampling design), a contention that is unjustified (Peterson 1993, Green 1994).

The CHIA and NOAA Hazmat inferences on degree of recovery are themselves far from perfect. In the CHIA program, the full suite of community composition and species abundance data was assessed only through summer 1991 (Highsmith et al. 1996). Only for *Fucus* and some other major occupiers of space on protected rocky shores did some sampling continue beyond that date (e.g., Stekoll & Deysher 1996, van Tamelen et al. 1997). Consequently, the ability to demonstrate convergence of oiled and control communities and the ability to test the assumption that, in the absence of spill influence, oiled and control sites would be biologically identical had not been realized. Sampling in the NOAA Hazmat program has continued annually at many study sites through summer 1999, so the time frame for assessing convergence is much greater. Epibiotic communities on sheltered rocky shores demonstrated convergence by about 1993, but subsequently diverged (Houghton et al. 1997). On mixed sedimentary shores, infaunal abundances had not yet converged by 1998 (Coats et al. 1999). This raises the unanswered question of whether the oiled/ treated sites and control sites possess intrinsic large environmental and biological differences or whether the erosional loss of sediments during application of pressurized hot water has altered the habitat in a way that is slow to recover (Coats et al. 1999). Thus, recovery rate is still not well characterized by CHIA or NOAA Hazmat, but the longer time frames of assessment provide a superior opportunity both to estimate duration of the recovery processes and to test implicit assumptions of intrinsic identity of treatment and reference sites.

(15) Power analysis

The 2 stratified random studies of the impacts of the oil spill used very different approaches to treating questions of power of tests (Table 3). CHIA used observed variances to calculate directly the power of all tests for effects of specified magnitudes (Highsmith et al. 1994). SEP did not calculate power for specific tests for specific magnitudes of effects but instead adopted an approach of estimating power with a pilot power simulation study using a 'signal-to-noise ratio' approach (Page et al. 1995, Gilfillan et al. 1996). Such an approach is justified as a first cut where error variances are unknown but cannot substitute for estimating and controlling error variances in the actual data and basing power analyses on those estimated variances and on specified effect magnitudes for each variable and test. It is worth noting that all these complex and often obfuscating calculations of power would disappear if an estimation approach to the assessment had been adopted. Reporting the standard errors of estimates would replace power calculation with a much more understandable and standard metric (Stewart-Oaten et al. 1992).

The power simulation of SEP is, in addition, based upon the philosophy that a signal-to-noise ratio of 1 is adequate for environmental impact testing. This is a form of the contention that power need only be sufficient to detect a change that is greater than the natural variability. That is a grossly conservative standard (Green 1994). Furthermore, the error that enters into such a contrast is dictated by the study design that yields it, such that all decisions that lead to failure to control for variability in the design affect the magnitude of the standard. For example, a failure to stratify by tidal elevation or by habitat would greatly increase the 'natural variation' as represented by the error variance and thus also the magnitude of an impact that is judged significant. In addition, intrinsic natural variability in space and time is very large in many populations and communities (e.g., Andrew & Mapstone 1987). Just because one population may exhibit higher variability than another does not imply logically that this first population can be hit with a larger impact without loss of value to the ecosystem or to humans (Peterson 1993). In fact, natural environmental variation represents a situation to which species have generally adapted. The effects of a major oil spill do not represent such a situation. They add on top of natural mortality and variation such that a significance standard equal to the magnitude of natural variability does not reflect risk of protracted times of recovery or even extinction at the extreme. The added mortality of an oil spill could reduce abundances to levels well below those from which species are accustomed and adapted to recover rapidly. In addition, the estimates of power provided in SEP are based upon error pooled among sites and among transects, which inflates the degrees of freedom and presumed power, as discussed above. By not providing actual calculations of power of individual tests based upon observed variances, the SEP study fails to show the balance between type I and type II error that is necessary to share the burden of proof in environmental assessments (Dayton 1998). SEP maintained a strict, classical devotion to an alpha of 0.05 (Toft & Shea 1983) without demonstrating that actual power was high for each important test.

Issues of appropriate biological response variables

(16) Taxonomic level used for analysis

CHIA (Highsmith et al. 1996, Stekoll et al. 1996) and NOAA Hazmat (Driskell et al. 1996, Houghton et al. 1996) analyze and report patterns of the abundance of individual species populations, whereas SEP (Gilfillan et al. 1995a) and GOA (Gilfillan et al. 1995b) rarely report results by species and instead tend to pool data for univariate analyses into higher taxonomic categories (Table 4). Such pooling uses less of the available information and can obscure impacts if potentially important compensatory changes occur within higher taxonomic categories. Species that are closely related taxonomically tend to be more ecologically similar and thus represent more likely competitors. Given that sampling of intertidal biota in these 2 stratified random studies of impact to shoreline biota occurred 15 to 17 mo after the spill, and given that competition for space can be an important process in the rocky intertidal habitat (e.g., Connell 1972), it is reasonable to expect that within genera or families compensatory changes may have occurred that could mask the decline of a sensitive member species in the epibiota. For example, Highsmith et al. (1996) showed that, of the species of barnacles on the high rocky shore, *Semibalanus* and *Balanus* spp. tended to exhibit significant declines in density on oiled sites, whereas the opportunistic species *Chthamalus dalli* increased. Analysis of barnacles as a group fails to detect this changing membership and compensatory trends in post-spill dynamics. Consequently, the analysis of biological response of individual taxa at a higher level of taxonomy in systems where competition can be an important factor affecting densities can imply reduced ability to detect significant differences between oiled and reference shores. Through this decision about taxonomic level of biological response variables, SEP and GOA became less likely than CHIA and NOAA Hazmat to detect effects of the oil spill.

There are valid reasons why environmental assessments may not be conducted using species-level discriminations. Species-level identification may not always be possible or financially feasible, although the

Table 4. Differences among studies of injury to the intertidal biota after the 'Exxon Valdez' oil spill in choice of biological response variables

Issue	SEP	Exxon GOA	CHIA	NOAA
(16) Taxonomic level used for analysis	With 4 exceptions at the species level, analyses provided only at higher levels, especially total community	Presents analyses done only at a community level (e.g., total cover, total abundance)	Analyses down to the species level	Analyses down to the species level
(17) Pooling disparate communities	Pooled poorly sampled meiofauna (nematodes) and macrofauna in soft-sediment analyses	Unclear whether all animals retained on 1 mm were used or just macrofauna	Used only macrofauna in the traditional taxonomic separation of soft-sediment communities	Used only macrofauna in the traditional taxonomic separation of soft-sediment communities
(18) Scope of communities and habitats examined	Evaluated the epibiota and infauna where possible on exposed bedrock, protected bedrock, boulder/cobble, and pebble/gravel	Evaluated the epibiota on bedrock and boulder/cobble and the infauna on boulder/cobble and pebble/gravel	Included fine-sediment estuarine habitats, intertidal fishes, and eelgrass amphipods,[a] and echinoderms[b] –all known to be sensitive to toxic contamination	Evaluated the epibiota of rocky shores and equally the infauna of mixed soft-hard shores

[a] Reported in Jewett et al. (1999)
[b] Reported in Dean et al. (1996)

SEP report (Gilfillan et al. 1995a) mentions that identifications were done at the species level. The question of how important the shuffling of species abundances within higher taxa might be to community function and value, of course, is not answered. Most species in a community undergo species-specific interactions with 1 or more members of the community, so it is reasonable to assume that species composition matters. However, that question of the value of biodiversity in communities remains an area of current intense investigation (e.g., Tilman & Downing 1994, Naeem 1997). Several papers have conducted comparisons of parallel multivariate ordination analyses using nonmetric multi-dimensional scaling to assess the consequences of using species-level information versus higher-level data in the same data sets. These studies (e.g., Warwick 1988, Warwick & Clarke 1993) reveal that identifications made at the family level are typically as effective as those made at the species level in detecting patterns in marine infaunal communities. That information critical to discrimination is not lost in changing from species to family levels in infaunal communities may be related to the general ineffectiveness of competition in sedimentary systems (Peterson 1991). Rocky intertidal communities, in contrast, can be strongly organized by competition for space (Connell 1972). Consequently, the compensatory replacement of barnacles evident in the CHIA rocky shore results (Highsmith et al. 1996) that is masked in analysis at the family level may not represent a problem in analysis of infaunal communities. In assessing patterns in infaunal, as opposed to epibiotic, communities, limited resources for study might be more wisely spent on enhancing the statistical rigor and power of the sampling design than on identifications at fine taxonomic levels.

(17) Pooling of disparate communities

The 2 stratified random studies of intertidal impacts of the oil spill on Prince William Sound made different decisions about how to define the response community in sedimentary environments (Table 4). Although both studies employed a mesh size of 1 mm to process core samples, CHIA included only those taxa recognized as macrofauna, those taxa effectively retained on a 1 or 0.5 mm mesh. SEP included meiofauna such as nematodes and copepods in its definition of the infaunal response community. Inclusion of data on meiofaunal taxa such as nematodes means that a species variable is used that is very poorly estimated, since only the largest meiofauna are retained on the mesh and size distributions may vary among sites. Furthermore, retention of even large meiofauna on a 1 mm mesh would be very sensitive to the vigor applied in washing and the amount of sediment retained. Such poorly estimated variables increase error and decrease the ability to detect significant differences in assessments of oil spill effects. The standard protocol for evaluating soft-sediment communities is to treat the meiofauna and macrofauna separately and independently, including use of specialized sampling methodologies to provide stable and repeatable estimates of meiofaunal and macrofaunal abundances (e.g., Peterson et al. 1996). Pooling inadequately estimated meiofaunal counts with macrofauna is simply unjustified and invalid scientifically. The GOA study of infauna shared this problem of merging poorly sampled meiofauna and macrofauna into its estimates of composition, abundance, and species diversity of soft-sediment communities (Gilfillan et al. 1995b), whereas the NOAA Hazmat study generally excluded meiofauna and epifauna from its analyses of oil spill impacts on sedimentary communities (Driskell et al. 1996). This NOAA Hazmat study does the best job of all 4 assessments on infaunal communities because in addition to its proper separation of meio- and macrofauna its results have been published. Results of the CHIA study of infauna have not been made generally accessible.

(18) Scope of communities and habitats examined

The wider the range of systems examined, the greater the chance of finding an impact of an environmental perturbation, if one indeed exists. CHIA examined a wider range of habitats and biotic response systems than did NOAA Hazmat, SEP, or GOA (Table 4). CHIA included estuarine fine-sediment sites, such as the marsh at Bay of Isles in Prince William Sound and the marsh at Tonsina Bay on the Kenai coast (Highsmith et al. 1996, Stekoll et al. 1996). While such fine-sediment shores were rare in the spill path, such systems are highly sensitive to oil, are known to suffer long protracted periods of injury if oiled, and have high ecological value (Teal & Howarth 1984). Consequently, their inclusion in an assessment of injuries resulting from the spill seems reasonable. SEP overlooked this habitat in its randomized design because of its rarity but did follow a fixed marsh site within Bay of Isles that was oiled. CHIA (Barber et al. 1995) also evaluated the response of intertidal fishes to the oil spill (which caused a reduction of about 50 % in fish abundance in summer 1990), whereas no other study included this group of organisms. The benthic life style and close association of these animals with the oiled rocks and sediments for foraging, egg laying, and incubation make them especially likely to be exposed to oil repeatedly and at high concentrations. Furthermore, the importance of small demersal fishes as prey to higher-level consumers such as pigeon guillemots and river otters renders these nearshore fishes important to ecosystem functioning. Consequently, devoting some effort to evaluate this group of potentially sensitive organisms represented a decision that enhanced the ability of CHIA to detect important effects of the spill. Similarly, a subtidal habitat study that began as part of CHIA focused explicit attention on sampling eelgrass beds for amphipods and echinoderms (Dean et al. 1996, Jewett et al. 1999), groups of organisms known to be especially sensitive to toxic chemicals in an environment of high value as a marine nursery habitat. This decision too enhanced the probability of detecting a spill effect as compared to SEP and GOA, which allocated effort to the most common habitats and did not devote attention to developing and applying methodologies for sampling some of the most sensitive taxa and environments. NOAA Hazmat sampled only 2 geomorphological habitats, did not include benthic fish evaluations, but did conduct assessments of eelgrass.

DISCUSSION

The conclusions about the extent of injury to intertidal resources, the types and patterns of injury, and the progress toward recovery differed rather substantially among the 4 assessment studies. The NOAA Hazmat study found that pressurized hot-water wash, a widely applied post-spill shoreline treatment in Prince William Sound, directly killed large numbers of rockweed (*Fucus*) plants, blue mussels and *Protothaca* clams (Lees et al. 1996). Lees et al. (1996) further concluded from short-term (3 to 10 d) experiments that 95 min or more of application of this pressurized hot-water treatment reduced densities of *Fucus*, littorine snails, limpets, and mussels in the mid and upper intertidal elevations by up to 100 %. Losses from this treatment lower on shore were not so great and losses from other beach treatments were more selective of the species that they injured (Lees et al. 1996). Consistent with these short-term observations of mortality and experimental demonstrations of declines following pressurized hot-water washing, the long-term evaluations of the biological impacts of the oil spill and of shoreline treatment also demonstrated large declines in many epibiotic species on sheltered rocky shores and some infaunal species on mixed sedimentary shores. The long-term NOAA Hazmat study revealed that most of the reduction in abundance of intertidal epibiota following the oil spill was caused by shoreline cleanup, with a smaller contribution made by the oiling itself (Houghton et al. 1996). By summer 1991, no significant differences in epibiota remained between oiled-untreated shores and unoiled reference shores, but the epibiota on oiled + treated shores was still significantly depressed (Houghton et al. 1996). Although by summer 1992 few significant differences were detected between oiled + treated shores and unoiled reference shores, in summers of 1994 and 1995 the single-aged stand of *Fucus* exhibited widespread senescence and its demise resulted in declines of associated animals (Houghton et al. 1997). By 1996, recovery had begun anew, suggesting that the oil spill may have induced a cyclic behavior in this system (Paine et al. 1996). In the lower intertidal zone of mixed soft-sediment shores in protected areas, differences in community composition of the infaunal community demonstrated that impacts of oiling and pressurized hot-water treatment were greater than impacts of oiling alone. While oiled-untreated beaches were similar to unoiled ones by summer 1992, the oiled + treated beaches were still impoverished and had not recovered by that date more than 3 yr after the oil spill (Driskell et al. 1996). Direct observations of mortalities of *Protothaca* clams, transport of sediments down-slope and loss of fines, and burial of animals suggest the mechanisms by which much of the impact on treated soft-sediment shores occurred (Driskell et al. 1996).

The conclusions about the effects of the 'Exxon Valdez' oil spill on intertidal biota in Prince William Sound that emerged from CHIA were largely consistent with the conclusions of the NOAA Hazmat study. CHIA was unable to partition effects of oiling and post-spill shoreline treatment, so its conclusions apply to the joint effects of the 2 perturbations. The studies also differed in their duration, with most of the sites examined by CHIA unvisited since 1991 (although some further multi-year data are available for the sheltered rocky habitat in Prince William Sound; van Tamelen et al. 1997). For the 2 years 1990 and 1991, when both CHIA and NOAA Hazmat evaluated the intertidal injuries and dynamics of the recovery process on sheltered rocky shores of Prince William Sound, the results were highly concordant. Specifically, total epibiotic cover, abundance, and biomass were substantially lower at oiled than at reference sites at all 3 tidal levels of sheltered rocky shore in Prince William Sound during the initial visit in 1990 (Highsmith et al. 1996, Houghton et al. 1996, Stekoll et al. 1996). The taxa most responsible for this general pattern were *Fucus*, the barnacle *Balanus glandula*, the limpet *Tectura persona* in the upper intertidal zone, and the mussel *Mytilus trossulus* in the low intertidal (Highsmith et al. 1996, Houghton et al. 1996). The barnacle *Chthamalus dalli* and oligochaetes showed the opposite pattern of enhancement on oiled shores (statistically significant only for *C. dalli*). By the final sampling in summer 1991, recovery was incomplete in this sheltered rocky habitat of Prince William Sound, with many taxa still significantly depressed on oiled shores (Highsmith et al. 1996). NOAA Hazmat also showed little progress toward recovery of epibiota in sheltered rocky shores of Prince William Sound by 1991 (Houghton et al. 1996). CHIA did not report results of its infaunal sampling on mixed soft-sediment shores to compare to the results of the NOAA study. For the habitat and geographic region held in common, the similarities in the results of these 2 studies are striking. Given the fundamental contrast in approach of these 2 studies, one using a stratified random sampling and the other a subjective choice of fixed sites, the similarity in results provides confidence in the robustness of the conclusions.

Results and conclusions of the SRS portion of SEP contrast rather sharply with those of the other 2 studies of intertidal injury and recovery in Prince William Sound. Of 141 tests conducted on densities of individual species in the sheltered rocky habitat, 13.5% showed significant effects of oiling, but more of the significant differences indicated enhanced abundance rather than depressed abundance (Gilfillan et al. 1995a). The important provider of structural habitat, *Fucus,* tended to exhibit substantially lower biomass at oiled sites in the sheltered rocky habitat, although statistical tests only detected significance in some contrasts (Gilfillan et al. 1995a). So, except for the lack of statistical significance in most contrasts, this response mirrors those of the NOAA Hazmat and CHIA studies. The total abundance of all species of limpets combined was reported to be significantly depressed by light and moderate oiling but not by heavy oiling in the high intertidal zone of sheltered bedrock shores (Gilfillan et al. 1995a), a pattern not immediately explicable. Data on the species *Tectura persona* alone by tidal elevation were not presented, so a direct contrast with CHIA results for that species is not possible. The NOAA Hazmat study detected significant and large declines for all limpets pooled only at mid tidal levels (Houghton et al. 1996) and CHIA only occasionally showed significant responses (declines) in the total limpet category despite the significant and large reductions with oiling in the one component species, *T. persona* (Highsmith et al. 1996). SEP failed to detect any significant response of mussel abundance to the oil spill (Gilfillan et al. 1995a), although both CHIA (Highsmith et al. 1996) and NOAA Hazmat (Houghton et al. 1996) showed generally lower mussel densities at oiled sites in this sheltered rocky habitat. Thus, for limpets and mussels in this habitat and region where all 3 studies overlapped, the results of SEP are discordant.

While no additional results for tests of other taxa on sheltered rocky shores are provided in the SEP publication to permit contrasts with the CHIA and NOAA Hazmat results, SEP does provide information on tests of total invertebrate abundance and algal biomass: no significant differences were detected in the sheltered rocky habitat for either of these parameters (Gilfillan et al. 1995a). This lack of response contrasts sharply with the large, generally significant, and consistent depressions on oiled shores detected in both the NOAA Hazmat and CHIA studies. Because 97% of all community-level tests of impact of the oil spill (on total invertebrate density, algal biomass, species richness, and Shannon-Wiener diversity) failed to detect a significant effect of the oil spill in this habitat, SEP reported that one estimate of degree of recovery was 97% recovery for the sheltered bedrock shore of Prince William Sound in 1990 (Gilfillan et al. 1995a). This contrasts dramatically with the large and still statistically significant differences in many species- and community-level parameters of the intertidal community by the end of summer 1991 that were demonstrated in both CHIA (Highsmith et al. 1996, Stekoll et al. 1996) and NOAA Hazmat (Driskell et al. 1996, Houghton et al. 1996).

This same habitat, the sheltered rocky shore, was also studied in the Gulf of Alaska region independently in 2 of the major assessments, CHIA and GOA, allowing a further contrast of how studies differing in methodology, means of removing bias, and power of their assessment designs may have produced accordingly different conclusions. In CHIA, where the geography was segregated into the Kenai-lower Cook Inlet shores versus the Kodiak-Alaska Peninsula shores, some large differences in response to the oil spill emerged between the 2 geographic regions. In both regions, the *Fucus* cover was significantly lower and the unoccupied space significantly higher on oiled shores than on reference sites in the upper and mid intertidal zones of the sheltered rocky habitat for most sample dates in 1990 and 1991 (Highsmith et al. 1996). In the low intertidal, *Fucus* cover and biomass were significantly greater on oiled sheltered rocky shores in the Kenai region but not in Kodiak-Alaska Peninsula: *Fucus* cover in the lower shore of the Kenai region was higher on oiled shores, where this species was replacing other algae, especially some annual reds and browns (Highsmith et al. 1996). The mussel *Mytilus trossulus* exhibited significantly reduced densities on oiled shores within the sheltered rocky habitat on the Kenai but not in the Kodiak-Alaska Peninsula region (Highsmith et al. 1996). The limpet *Tectura persona* occasionally exhibited lower abundance on oiled shores at the mid tidal level on the Kodiak-Alaska Peninsula region's sheltered rocky coast but not on the Kenai Peninsula region shores in the CHIA study results (Highsmith et al. 1996). The barnacle *Chthamalus dalli* tended to be more abundant on oiled shores in the sheltered rocky habitat, especially on the Kodiak-Alaska Peninsula region, although the pattern was reversed in the high intertidal on the Kenai coast (Highsmith et al. 1996). The other barnacles, *Balanus glandula* and *Semibalanus balanoides*, exhibited inconsistent patterns of differences in the sheltered rocky habitat across tidal levels and between the 2 geographic regions. GOA reported only some community-level parameters that allow comparisons with CHIA results for this region (combining the Kenai and Kodiak-Alaska Peninsula into a single region). Both total biotic cover on bedrock (pooling sheltered and exposed habitats) and total epifaunal abundance declined substantially (at all 3 elevations) and significantly (at the mid intertidal only) with increased levels of oiling (Gilfillan et al. 1995b). This result is fully consistent with the observation of widespread and significant increases in unoccupied intertidal space on the sheltered rocky shores of the Kenai and Kodiak regions demonstrated in CHIA (Highsmith et al. 1996). Thus, these 2 studies did not produce inconsistent results where comparisons can be drawn, but the heterogeneous response between the 2 geographic regions at the species level exhibited in CHIA could not possibly be reflected in GOA results, even if species-level results were reported, because of the failure to stratify geographically. GOA conducted its assessments of biological response in 1989, whereas the redesigned CHIA study, like SEP, assessed impacts beginning in 1990. This difference in timing of assessments may help crudely compensate for the difference in character of oil between regions: the Kenai-Kodiak regions received the oil after 1 to 8 wk of transport and weathering, whereas the Prince William Sound region was oiled rapidly and more heavily by less weathered oil (Wolfe et al. 1994). The GOA study to assess the Kenai-Kodiak region detected larger and more consistent negative impacts of the oiling than the SEP study detected for Prince William Sound. Differences in year of assessment and in sampling design probably both contribute to the contrasting levels of impact reported in these 2 studies.

One other habitat and geographic area was held in common by pairs of the major intertidal assessments, allowing some limited further contrasts. The infauna of the boulder/cobble habitat in Prince William Sound was studied by a common technique of coring in both SEP and NOAA Hazmat. SEP failed to detect an oiling effect on total infaunal abundance at any elevation in this habitat despite substantially lower mean abundances at mid intertidal elevations and somewhat higher mean abundances at low intertidal elevations on oiled shores (Gilfillan et al. 1995a). No analyses of responses of individual infaunal taxa were presented for SEP. Because 85% of the community-level analyses failed to identify a significant effect of the oil spill on the infauna in this habitat, SEP reported that one estimate of the degree of recovery was 85% recovery of the infauna of the boulder/cobble habitat (Gilfillan et al. 1995a). The NOAA Hazmat coring study for this habitat in Prince William Sound demonstrated large and significant depressions in infaunal abundance on oiled + treated shores as compared to unoiled and oiled-untreated shores for each of 4 years 1989 to 1992 (Driskell et al. 1996). The dominant infaunal taxa also exhibited substantial depression in abundance at oiled + treated boulder/cobble sites such that the community composition differed between oiled + treated shores, on the one hand, and both unoiled and oiled-untreated shores. Between 1991 and 1992, the compositions of the 3 types of shores began to converge, but densities of major taxa were still depressed on oiled + treated shores (Driskell et al. 1996). Thus, the SEP and the NOAA Hazmat studies conflicted in their ability to detect impact and in the magnitude of the estimated responses. They also differed in the claims about recovery (85% by 1990 reported in SEP vs grossly divergent still in 1992 in NOAA Hazmat).

This set of contrasts of the comparable conclusions of the 4 major studies of intertidal injury and recovery reveals a pattern in which CHIA and NOAA Hazmat demonstrate substantially more impacts of the 'Exxon Valdez' oil spill on intertidal populations and communities in Prince William Sound than SEP. In addition, the CHIA program outside Prince William Sound likewise showed more responses than GOA. These differences in outcomes can be related to differences among studies in the suite of design decisions that affect the power of the environmental assessment. To provide contrasts among the 4 studies in how different design decisions affected power to detect impacts of the oil spill, we tabulated results of our analysis of design decisions (Table 5). We did not attempt to quantify magnitudes of differences among studies for each design decision but instead provide rankings of the 4 studies for each design parameter. Although such rankings necessarily involve professional judgment, the basis for the each ranking has been presented and defended in the corresponding sections above. We also did not attempt to weight these decisions by their relative importance in determining the overall power of the study, but we discuss the relative importance of various decisions below. We merely list each separately and then provide a sum of ranks for each study. We rank studies in order of lowest (1) to highest (4) power in this table.

Of the 18 design decisions that we isolated for analysis, 3 largely concern philosophy of assessment and do not represent separate criteria for assessing power contrasts. Of the 15 remaining design decisions, SEP included choices that reduced its power to detect impacts of the oil spill below that of CHIA in 12 decisions, with 1 tie (Table 5). Only in 2 (no. 11–controls for shoreline treatment and oiling intensity, and no. 13–pseudoreplication) of these remaining 14 decisions did SEP employ a more powerful approach. In assigning a greater power to SEP for using 4 levels of oiling in its design instead of just the 2 used by CHIA, we provide a more conservative estimate of the cumulative power difference between these studies because we assume that biological differences exist between light oiling and no oiling and between moderate and heavy oiling. In addition, although SEP receives a higher

Table 5. Summary of how design decisions influenced power to detect impacts of the 'Exxon Valdez' oil spill in the 4 major impact assessments. Entries are ranks for each decision variable (with 1 indicating lowest power). Ties are averaged to preserve the sum of ranks across rows. na: not applicable

Design decision	SEP	GOA	CHIA	NOAA
(1) Area covered per sample	1	2.5	4	2.5
(2) Sample replication	1.5	1.5	3	4
(3) Numbers of study sites per category	2.5	2.5	2.5	2.5
(4) Numbers of sampling dates	1.5	1.5	3	4
(5) Total area sampled	1	2	4	3
(6) Philosophical support for targeting putative affected areas	na	na	na	na
(7) Random site selection vs matched pairs	na	na	na	na
(8) Sampling frame	1	2.5	4	2.5
(9) Treatment of habititat heterogeneity within sites	1	2.5	4	2.5
(10) Interspersion of sites	1	2	4	3
(11) Controls for shoreline treatment and oiling intensity	3	2	1	4
(12) ANCOVA with covariate affected by treatment	3	1.5	4	1.5
(13) Pseudoreplication	4	3	1.5	1.5
(14) Inferring degree of recovery	1.5	1.5	3	4
(15) Power analysis	na	na	na	na
(16) Taxonomic level used for analysis	1.5	1.5	3.5	3.5
(17) Pooling of disparate communities	1.5	1.5	3	4
(18) Scope of communities and habits assessed	2	1	4	3
Sum of rank scores	27	29	48.5	45.5

power score for the decision to use pseudoreplicated transects as replicates in constructing *F*-ratios in tests of oiling (Gilfillan et al. 1995a report that this decision enhanced power), the spatial confounding may have led to bias in estimates of effects. Similarly, the SEP study included decisions that reduced detection power relative to the NOAA Hazmat study in 12 of the 15 parameters, with 1 tie (Table 5). The only decisions that favored power of the SEP study were again the problematic decision to use pseudoreplication in constructing its *F*-ratios (no. 13) and the measurement and use of covariates to remove effects of environmental heterogeneity (no. 12). As a means of providing contrasts in cumulative detection power among studies, for each study we summed the ranks over all 15 decision criteria (Table 5). This calculation reveals the great disparity between SEP at the bottom with a sum of ranks of 27 as compared to CHIA (48.5) and NOAA Hazmat (45.5) at the top. This disparity in power of overall assessment design is likely the explanation for SEP reporting fewer impacts of the oil spill than CHIA or NOAA Hazmat. The most important decisions that enhanced power to identify real impacts of the oil spill in the CHIA and NOAA Hazmat studies are probably the total area sampled per category, the use of multiple sampling dates, and the restriction of the sampling frame so that reference sites better mimicked the nonrandom selection of sites by the oiling process. Each of these factors differed greatly between the SEP and the CHIA studies (Tables 1 to 4). Other decisions, such as the problem of using a covariate that was affected by the treatment, probably had relatively small influence on the overall conclusions of the tests and studies.

The CHIA and NOAA Hazmat studies differ in one major way that affects power, by the control of shoreline treatment history. The short- and long-term studies of NOAA Hazmat both suggest that the high-pressure hot-water wash that was widely applied as a shoreline treatment had more serious consequences than the oiling itself (Driskell et al. 1996, Houghton et al. 1996, Lees et al. 1996). Thus, the inability to control for type and amount of shoreline treatment in the CHIA study probably greatly increased the unexplained error variance and accordingly lowered its power to detect true oil spill effects. This may explain why some of the impacts reported in the NOAA Hazmat study are often somewhat larger than those identified in the CHIA study.

Although the GOA study like the SEP study was supported by Exxon Corporation, shared some principle investigators, and had many similarities in methodology, the 2 studies differed in important ways. The sum of the ranks of design decisions that affected power for the GOA study was 29 as compared to 27 for SEP (Table 5). Yet, these 2 studies contrasted greatly in some key design decisions that probably affected power most, namely the choice of quadrat size, total area sampled, and use of a sampling frame that was similar for oiled and control sites (Tables 1 to 4). GOA used an areal cover quadrat almost an order of magnitude larger in surface area than the scrape sample used by SEP (Table 1). This led to conclusions based upon more total area sampled for a given habitat (Table 1) and allowed a better characterization of the local community composition in this system of such patchiness (Paine & Levin 1981). GOA sampled 7 times the total area per habitat stratum. And GOA did not appear to share the problem of SEP that involved confounding of oil effects with stress from glacier ice-melt at reference sites. These differences between SEP and GOA in design decisions that affect power may explain the paradox that GOA demonstrated larger and more consistently negative impacts of the oil spill in the geographic region that received weathered oil some 1 to 8 wk after much of the oil had initially grounded on the nearby shores of Prince William Sound. Alternatively or additionally, the year of assessment may play an important role in dictating differences between SEP and GOA conclusions: GOA was conducted in 1989 whereas SEP assessed biological patterns in 1990.

Despite the apparent complexity and sophistication of designs for environmental assessments, multiple simple design decisions are often ultimately the determinants of much of the power to detect true impacts. We take the opportunity provided by redundancy, or at least overlap, of intensive studies of how the same ecological system responded to a large environmental perturbation to show how even complex studies can be analyzed by extracting the set of basic design decisions that influence their power to detect impacts, accurately estimate them, and describe the course of recovery. We then recombine those separate decisions into a single metric to contrast the multiple studies. When any study combines several decisions that fail to reduce error variance and fail to enhance power, the outcome of such multiple decisions is almost guaranteed to be a set of inconclusive results unable to detect even large impacts of environmental perturbation. Wiens (1996) argued that government agency science suffers in general and suffered in specific seabird analyses following the ‘Exxon Valdez’ oil spill from becoming overwhelmed by environmental advocacy. What Wiens failed to acknowledge are the strong financial and institutional incentives for corporate science also to overwhelm dispassionate and objective scientific judgment. Should we place more trust, for example, in the human health science promulgated and publicized by the tobacco industry than that conducted by government health laboratories? To achieve a balanced understanding of environmental impacts, assessment designs should include estimation approaches and insure high power to detect and estimate impacts of importance. Government and corporate entities both often have substantial resources to invest in scientific studies: any inequality in such investment can greatly affect sampling effort and thereby create an imbalance in detection power. Evaluating reliability of conclusions of scientific studies includes a need not only for comparing effort but also for synthesizing consequences of all the multiple design decisions that may affect that power.

Acknowledgements. Financial support for the preparation of this paper came from the PEW Charitable Trusts and the Exxon Valdez Oil Spill Trustee Council. The manuscript was improved greatly by comments from A. Stewart-Oaten and anonymous reviewers. The views expressed in this paper are solely the authors' and do not represent opinions of the Exxon Valdez Trustee Council.

LITERATURE CITED

ADEC (Alaska Department of Environmental Conservation) (1989) Impact maps and summary reports of shoreline surveys of the Exxon Valdez oil spill site–Prince William Sound. Report to the Exxon Valdez Oil Spill Trustee Council, Anchorage, AK

Andrew NL, Mapstone BD (1987) Sampling and the description of spatial pattern in marine ecology. Oceanogr Mar Biol Annu Rev 25:39–90

Babcock MM, Irvine GV, Harris PM, Cusick JA, Rice SD (1996) Persistence of oiling in mussel beds three and four years after the Exxon Valdez oil spill. Am Fish Soc Symp 18:286–297

Barber WE, McDonald LL, Erickson WP, Vallarino M (1995) Effect of the Exxon Valdez oil spill on intertidal fish: a field study. Trans Am Fish Soc 124:461–476

Bernstein BB, Zalinski J (1983) On optimum sampling design and power tests for environmental biologists. J Environ Manage 16:35–43

Box GEP, Hunter WG, Stuart J (1978) Statistics for experimenters: an introduction to design, data analysis, and model building. John Wiley and Sons, New York

Clarke KR, Ainsworth M (1993) A method of linking multivariate community structure to environmental variables. Mar Ecol Prog Ser 92:205–219

Coats DA, Imamura E, Fukuyama AK, Skalski JR, Kimura S, Steinbeck J (1999) Monitoring of biological recovery of Prince William Sound intertidal sites impacted by the Exxon Valdez oil spill: 1997 biological monitoring survey. NOAA Technical Memorandum NOS OR&R I. NOAA Hazardous Materials Response Division, Seattle, WA

Cochran WG (1977) Sampling techniques, 3rd edn. John Wiley and Sons, New York

Cochran WG, Rubin DB (1973) Controlling bias in observational studies: a review. Sankhya A35:417–446

Connell JH (1972) Community interactions on marine rocky intertidal shores. Annu Rev Ecol Syst 3:169–192

Cox DR (1952) Some recent work on systematic experimental designs. J R Stat Soc B 14:211–219

Cox DR (1957) The use of a concomitant variable in selecting an experimental design. Biometrika 44:150–158

Cox DR (1958) Planning of experiments. John Wiley and Sons, New York

Dayton PK (1998) Reversal of the burden of proof in fisheries management. Science 279:821–822

Dean TA, Jewett SC, Laur DR, Smith RO (1996) Injury to epibenthic invertebrates resulting from the Exxon Valdez oil spill. Am Fish Soc Symp 18:424–439

Digby PGN, Kempton RA (1987) Multivariate analysis of ecological communities. Chapman and Hall, New York

Driskell WB, Fukuyama AK, Houghton JP, Lees DC, Mearns AJ, Shigenaka G (1996) Recovery of Prince William Sound infauna from Exxon Valdez oiling and shoreline treatments, 1989 through 1992. Am Fish Soc Symp 18:362–378

Eberhardt LL, Thomas JM (1991) Designing environmental field studies. Ecol Monogr 61:53–73

Fairfield Smith H (1957) Interpretation of adjusted treatment means and regressions in analysis of covariance. Biometrics 13:282–308

Fairweather PG (1991) Statistical power and design requirements for environmental monitoring. Aust J Mar Freshw Res 42:555–567

Field JC, Clarke KR, Warwick RM (1982) A practical strategy for analyzing multispecies distribution patterns. Mar Ecol Prog Ser 8:37–52

Fisher RA (1971) The design of experiments, 9th edn. Hafner, New York

Folks JL (1984) Combination of independent tests. In: Krishnaiah PR, Sen PK (eds) Handbook of statistics 4: nonparameteric methods. North Holland, New York

Ford RG, Page GW, Carter HR (1987) Estimating mortality of seabirds from oil spills. In: Proceedings, 1987 Oil Spill Conference, American Petroleum Institute Publication 4452, Washington, DC, p 848–851

Gilbert RO (1987) Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, New York

Gilfillan ES, Page DS, Harner EJ, Boehm PD (1995a) Shoreline ecology program for Prince William Sound, Alaska, following the Exxon Valdez oil spill: Part 3–Biology. In: Wells PG, Butler JN, Hughes JS (eds) Exxon Valdez oil spill: fate and effects in Alaskan waters. ASTM STP 1219, American Society for Testing and Materials, Philadelphia, PA, p 398–443

Gilfillan ES, Suchanek TH, Boehm PD, Harner EJ, Page DS, Sloan NA (1995b) Shoreline impacts in the Gulf of Alaska region following the Exxon Valdez oil spill. In: Wells PG, Butler JN, Hughes JS (eds) Exxon Valdez oil spill: fate and effects in Alaskan waters. ASTM STP 1219, American Society for Testing and Materials, Philadelphia, PA, p 444–481

Gilfillan ES, Harner EJ, O'Reilly JE, Burns WA (1996) A comparison of shoreline assessment study designs used for the Exxon Valdez oil spill. Arctic Mar Oil-Spill Prog Tech Sem 19:615–630

Green RH (1979) Sampling design and statistical methods for environmental biologists. John Wiley and Sons, New York

Green RH (1989) Power analysis and practical strategies for environmental monitoring. Environ Res 50:195–205

Green RH (1994) Aspects of power analysis in environmental monitoring. In: Fletcher DJ, Manly BFJ (eds) Statistics in ecology and environmental monitoring, Otago Conference Series, Univ Otago Press, Dunedin, p 173–182

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, New York

Highsmith RC, Stekoll MS, Barber WE, McDonald L, Strickland D, Erickson WP (1994) Comprehensive assessment of coastal habitat. Exxon Valdez oil spill state/federal natural resource damage assessment. Final report of coastal habitat study 1A. Exxon Valdez Oil Spill Trustee Council, Anchorage, AK

Highsmith RC, Rucker TL, Stekoll MS, Saupe SM, Lindeberg

MR, Jenne RN, Erickson WP (1996) Impact of the Exxon Valdez oil spill on intertidal biota. Am Fish Soc Symp 18: 212–237

Houghton JP, Lees DC, Driskell WB, Lindstrom SC, Mearns AJ (1996) Recovery of Prince William Sound epibiota from Exxon Valdez oiling and shoreline treatments, 1989 through 1992. Am Fish Soc Symp 18:379–411

Houghton JP, Gilmour RH, Lees DC, Driskell WB, Lindstrom SC, Mearns A (1997) Prince William Sound intertidal biota seven years later–has it recovered? International Oil Spill Conference Paper No. 260. American Petroleum Institute, Washington, DC

Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211

Jewett SC, Dean TA, Smith RO, Blanchard A (1999) The Exxon Valdez oil spill: impacts and recovery in the soft-bottom benthic community in eelgrass habitats. Mar Ecol Prog Ser 185:59–83

Johnson B, Rogers J, Chu A, Flyer P, Dorrier R (1989) Methods for evaluating the attainment of clean-up standards. Vol 1. Soils and solid media. WESTAT Research, Inc, Rockville, MD. Prepared for the US Environmental Protection Agency, Washington, DC, EPA 230/02–89–042

Johnson DH (1999) The insignificance of statistical significance testing. J Wildl Manage 63:763–772

Krebs CJ (1989) Ecological methodology. Harper and Row, New York

Kuehl RO (2000) Design of experiments: statistical principles of research design and analysis, 2nd edn. Duxbury, Pacific Grove, CA

Lees DC, Houghton JP, Driskell WB (1996) Short-term effects of several types of shoreline treatment on rocky intertidal biota in Prince William Sound. Am Fish Soc Symp 18: 329–348

Legendre P (1993) Spatial autocorrelation: trouble or a new paradigm? Ecology 74:1659–1673

Leigh EG Jr, Paine RT, Quinn JF, Suchanek JH (1987) Wave energy and intertidal productivity. Proc Natl Acad Sci 84: 1314–1318

Manly BFJ (1997) Randomization, bootstrap, and Monte Carlo methods in biology. Chapman and Hall, London

Mapstone BD (1995) Scalable decision rules for environmental impact studies: effect size, type I, and type II errors. Ecol Appl 5:401–410

McArdle BH (1996) Levels of evidence in studies of competition, predation, and disease. NZ J Ecol 20:7–15

McDonald LL, Erickson WP, Strickland MD (1995) Survey design, statistical analysis, and basis for statistical inferences in Coastal Habitat Injury Assessment. In: Wells PG, Butler JN, Hughes JS (eds) Exxon Valdez oil spill; fate and effects in Alaskan waters. ASTM STP 1219, American Society for Testing and Materials, Philadelphia, PA, p 296–311

Mearns AJ (1996) Exxon Valdez shoreline treatment and operations: implications for response, assessment, monitoring, and research. Am Fish Soc Symp 18:309–328

Menge BA (1995) Indirect effects in marine rocky intertidal interaction webs: patterns and importance. Ecol Monogr 65:21–74

Naeem S (1997) Biodiversity enhances ecosystem reliability. Nature 390:507–509

National Research Council (NRC) (1986) Oil in the sea—inputs, fates, and effects. National Academy of Sciences Press, Washington, DC

Neff JM, Owens EW, Stoker SW, McCormick DM (1995) Shoreline oiling conditions in Prince William Sound following the Exxon Valdez oil spill. In: Wells PG, Butler JN, Hughes JS (eds) Exxon Valdez oil spill: fate and effects in Alaskan waters. ASTM STP 1219, American Society for Testing and Materials, Philadelphia, PA, p 312–345

Page DS, Gilfillan ES, Boehm PD, Harner EJ (1995) Shoreline ecology program for Prince William Sound, Alaska, following the Exxon Valdez oil spill: Part 1-study design and methods. In: Wells PG, Butler JN, Hughes JS (eds) Exxon Valdez oil spill: fate and effects in Alaskan waters. ASTM STP 1219, American Society for Testing and Materials, Philadelphia, PA, p 263–295

Paine RT, Levin SA (1981) Intertidal landscapes: disturbance and the dynamics of pattern. Ecol Monogr 51:145–178

Paine RT, Ruesink JL, Sun A, Soulanille EL, Wonham MJ, Harley CDG, Brumbaugh DR, Secord DL (1996) Trouble on oiled waters: lessons from the Exxon Valdez oil spill. Annu Rev Ecol Syst 27:197–235

Peterman RM, M'Gonigle M (1992) Statistical power analysis and the precautionary principle. Mar Pollut Bull 24: 231–234

Peterson CH (1991) Intertidal zonation of marine invertebrates in sand and mud. Am Sci 79:236–249

Peterson CH (1993) Improvement of environmental impact analysis by application of principles derived from manipulative ecology: lessons from coastal marine case histories. Aust J Ecol 18:21–52

Peterson CH (2001) The 'Exxon Valdez' oil spill in Alaska: acute, indirect and chronic effects on the ecosystem. Adv Mar Biol 39:1–103

Peterson CH, Kennicutt MC II, Green RH, Montagna P, Harper DE Jr, Powell EN, Roscigno PF (1996) Ecological consequences of environmental perturbations associated with offshore hydrocarbon production: a perspective on long-term exposures in the Gulf of Mexico. Can J Fish Aquat Sci 53:2637–2654

Piatt JF, Ford RG (1996) How many seabirds were killed by the Exxon Valdez oil spill? Am Fish Soc Symp 18:712–719

Piatt JF, Lensink CJ (1989) Exxon Valdez bird toll. Nature 342:865–866

Pielou EC (1966) The measurement of diversity in different types of biological collections. J Theor Biol 13:131–144

Rice WR (1990) A consensus combined P-value test and the family-wide significance of component tests. Biometrica 46:303–308

Schmitt RJ, Osenberg CW (1996) Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, CA

Schoener TH (1993) On the relative importance of direct versus indirect effects in ecological communities. In: Kawanabe H, Cohen JE, Iwasaki K (eds) Mutualism and community organization. Oxford Univ Press, Oxford, p 365–411

Searle SR, Casella G, McCulloch CE (1992) Variance components. John Wiley and Sons, New York

Skalski JR, Robson DS (1992) Techniques for wildlife investigations: design and analysis of capture data. Academic Press, New York

Spies RB, Hardin DD, Teal JP (1988) Organic enrichment or toxicology? A comparison of the effects of kelp and crude oil in sediments on the colonization and growth of benthic infauna. J Exp Mar Biol Ecol 124:261–282

Spies RB, Rice SD, Wolfe DA, Wright BA (1996) The effects of the Exxon Valdez oil spill on the Alaskan coastal environment. Am Fish Soc Symp 18:1–16

Steidl RJ, Hayes JP, Schauber E (1997) Statistical power analysis in wildlife research. J Wildl Manage 61:270–279

Stekoll MS, Deysher L (1996) Recolonization and restoration of upper intertidal *Fucus gardneri* (Fucales, Phaeophyta) following the Exxon Valdez oil spill. Hydrobiologia 326/327:311–316

Stekoll MS, Deysher L, Highsmith RC, Saupe SM, Guo Z, Erickson WP, McDonald L, Strickland D (1996) Coastal habitat injury assessment: intertidal communities and the Exxon Valdez oil spill. Am Fish Soc Symp 18:177–192

Stewart-Oaten A, Murdoch WW, Parker KR (1986) Environmental impact assessment: pseudoreplication in time? Ecology 60:1225–1240

Stewart-Oaten A, Bence JR, Osenberg CW (1992) Assessing effects of unreplicated perturbations: no simple solutions. Ecology 73:1396–1404

Sundberg K, Deysher L, McDonald L (1996) Intertidal and supratidal site selection using a geographical information system. Am Fish Soc Symp 18:167–176

Teal JM, Howarth RW (1984) Oil spill studies: a review of ecological effects. Environ Manage 8:27–44

Thompson SK (1992) Sampling. John Wiley and Sons, New York

Tilman D, Downing JA (1994) Biodiversity and stability in grasslands. Nature 367:363–365

Toft CA, Shea PJ (1983) Detecting community-wide patterns: estimating power strengthens statistical inference. Am Nat 122:618–625

Underwood AJ (1981) Techniques of analysis of variance in experimental marine biology and ecology. Oceanogr Mar Biol Annu Rev 19:513–605

Underwood AJ (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecol Appl 4:3–15

Underwood AJ (1997) Experiments in ecology, their logical design and interpretation using analysis of variance. Cambridge Univ Press, Cambridge

van Tamelen PG, Stekoll MS, Deysher L (1997) Recovery processes of the brown alga, *Fucus gardneri* (Silva), following the Exxon Valdez oil spill: settlement and recruitment. Mar Ecol Prog Ser 160:265–277

Warwick RM (1988) The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. Mar Pollut Bull 19:259–268

Warwick RM, Clarke KR (1993) Comparing the severity of disturbance: a meta-analysis of marine macrobenthic community data. Mar Ecol Prog Ser 92:221–231

Wiens JA (1996) Oil, seabirds, and science: the effects of the Exxon Valdez oil spill. Bioscience 46:587–597

Wiens JA, Parker KR (1995) Analyzing the effects of accidental environmental impacts: approaches and assumptions. Ecol Appl 5:1069–1083

Winer BJ (1971) Statistical principles in experimental design, 2nd edn. McGraw-Hill, New York

Wolfe DA and 11 authors (1994) The fate of the oil spilled from the Exxon Valdez. Environ Sci Tech 28:561A-568A

Editorial responsibility: Anthony Underwood (Contributing Editor), Sydney, Australia

Submitted: November 1, 1998; Accepted: August 9, 2000
Proofs received from author(s): December 27, 2000