On November 16, 2005, the ISRP and ISAB produced a revised and updated portion of the Research, Monitoring and Evaluation section of the Retrospective Report (pp 18-30) for publication, titled Study Designs for Research, Monitoring and Evaluation. The ISRP and ISAB's intent was to make the section accessible to a broader audience beyond the Columbia River Basin.

The pre-publication draft uses some different terminology than the previous section in the Retrospective Report. Namely, the term "observational study" is used instead of "mensurative experiment," and "manipulative experiment" is used instead of "randomized treatment experiment" as the two basic classifications of research, monitoring, and evaluation. The basic definitions for these two classifications was not changed.

Corresponding author
Lyman L. McDonald
WEST, Inc.
Western EcoSystems Technology, Inc.
217 South First Street
Laramie, Wyoming 82070
307-634-1756
lmcdonald@west-inc.com
USA

STUDY DESIGNS FOR RESEARCH, MONITORING, AND EVALUATION

Independent Scientific Advisory Board (ISAB) to the Northwest Power and Conservation Council (Council), Portland, Oregon, NOAA Fisheries, Seattle, Washington, and the Columbia River Basin Indian Tribes, Portland, Oregon

and

Independent Scientific Review Panel (ISRP) to the Northwest Power and Conservation Council, Portland, Oregon.

Members:
Lyman L. McDonald

Robert Bilby

Peter A. Bisson

Charles C. Coutant

John M. Epifanio

Daniel Goodman

Susan Hanna

Nancy Huntly

Brian Riddell

William Liss

Eric J. Loudenslager

David P. Philipp

William Smoker

Richard R. Whitney

Richard N. Williams

McDonald is Senior Biometrician, Western EcoSystems Technology, Inc., Cheyenne, Wyoming. He can be reached at lmcdonald@west-inc.com.

Bilby is an Ecologist at Weyerhaeuser Company, Tacoma, Washington.

Bisson is Senior Scientist at the Forestry Sciences Laboratory of the U.S. Forest Service, Olympia, Washington.

Coutant is Distinguished Research Ecologist, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Epifanio is Director and Associate Professional Scientist for the Center for Aquatic Ecology at the Illinois Natural History Survey, Champaign, Illinois.

Goodman is Professor of Ecology, Montana State University, Bozeman, Montana.

Hanna is Professor of Marine Economics, Department of Agricultural and Resource Economics, Oregon State University, Corvallis, Oregon.

Huntly is Professor of Ecology, Idaho State University, Pocatello, Idaho.

Liss is Professor of Fisheries, Oregon State University, Corvallis, Oregon.

Loudenslager is Adjunct Professor of Fisheries Biology and Hatchery Manager, Humboldt State University, Arcata, California.

Philipp is Director of the Illinois Fisheries Genetics Laboratory, Center for Aquatic Ecology at the Illinois Natural History Survey, Champaign, Illinois.

Riddell is a Fisheries Ecologist, Pacific Biological Station, Department of Fisheries and Oceans Canada, Nanaimo, British Columbia.

Smoker is Professor of Fisheries, University of Alaska Fairbanks, Juneau Center for Fisheries and Ocean Sciences, Juneau, Alaska.

Whitney is a Consulting Fisheries Scientist, Leavenworth, Washington.

Williams is a Fisheries Biologist, Center for Salmonids and Freshwater Species at Risk, University of Idaho, Hageman, Idaho.

Key Words: mensurative experiment, observational study, manipulative experiment, quasi-experiment, quasi-treatment, model-based inference, design-based inference

ABSTRACT

Terminology defining *Research, Monitoring, and Evaluation* (RM&E) varies widely among fields of science. This variation is particularly evident when one considers the scientific basis for "effectiveness monitoring" of active management actions that are, for example, intended to improve habitat or recover threatened or endangered populations. We classify RM&E studies into manipulative experiments and observational studies following their long-standing usage by statisticians and empirical scientists. We explore specifically the scientific basis behind inferences in manipulative experiments and observational studies, recognizing that observational studies, including quasi-experiments, can be just as complex and just as effective as manipulative experiments. A review of these concepts that leads to the development of common terminology is an important step for promoting effective communication among researchers, administrators, and policy makers. Although the concepts, terminology, and recommendations are illustrated with fishery and wildlife applications, we believe that the audience for this perspective includes all areas of science. We conclude by making recommendations by which RM&E can meet the challenges of large-scale monitoring of complex entities such as ecosystems.

Introduction

The Independent Scientific Advisory Board (ISAB) for the Northwest Power and Conservation Council (Council), Portland, Oregon, NOAA-Fisheries, Seattle, Washington, and the Columbia River Basin Indian Tribes, Portland, Oregon, and the Independent Scientific Review Panel (ISRP) for the Council have existed from 1996 and 1997, respectively. The responsibilities of the ISRP, specified by the U.S. Congress, (Williams, 2006), include review of project proposals and quality control/quality assurance for implementation of the Council's Fish and Wildlife Program. One of the ISRP's statutory responsibilities is to evaluate whether or not project proposals have adequate provisions for monitoring and evaluation. The primary role of the ISAB is to examine the broader scientific basis underpinning various management decisions concerning the recovery or maintenance of fish and wildlife populations in the Columbia River Basin.

Since their formation, the ISRP and ISAB have observed inconsistent use of *Research, Monitoring, and Evaluation* (RM&E) terminology by various fields of science (e.g., fisheries, wildlife, hydrology, genetics, statistics) and an inconsistent scientific basis for "effectiveness monitoring" of active management actions of responsible organizations (e.g., NOAA Fisheries, Bonneville Power Administration, U.S. Army Corps of Engineers, and Bureau of Reclamation). For example, the words mensurative, implementation, observational, retrospective, non-experimental, pseudo-experiments, quasi-experiments, Tier 1 monitoring, Tier 2 monitoring, or simply "monitoring" have been used to identify one general category of scientific studies. The words manipulative, true experiment, effectiveness monitoring, Tier 3 monitoring, randomized treatment, and response monitoring have been used to identify a second general category of RM&E. Furthermore, the roles of each of the core types of inference-supporting monitoring in large-scale environmental and ecological programs are not well described in the literature.

We elect to classify RM&E studies into manipulative experiments and observational studies following long-standing usage by statisticians and empirical scientists. In some disciplines, the term “comparative experiment” has been used historically, but in the ecological sciences “manipulative experiment” has dominated for decades and is now fully entrenched (e.g., Scheiner and Gurevitch 1993, Underwood 1997, Krebs 1999, Quinn and Keough 2002, Gotelli and Ellison 2004). Hurlbert (1984) coined the term “mensurative experiment” in an attempt to accommodate the precedent that scientists have long used “experimental” as a synonym for “empirical” and applied it to observational studies of various sorts. Hurlbert also made clear, however, that guidance on the planning of "mensurative experiments" was to be sought in books on sampling design, not those on experimental design, and he has recently expressed concern about how the term "mensurative experiment" seems to have fostered confusion in the literature (Hurlbert 2003). In strongly statistical contexts, he now feels the term is best not used at all (S. Hurlbert, San Diego State University, personal communication).

More useful than "mensurative experiment" is the older term "quasi-experiment," which is loosely defined as an observational study "in which the research person can introduce something like experimental design into his scheduling of data collection procedures … even though he lacks the full control over the scheduling of experimental stimuli [i.e., imposition of independent variables]...which makes a true experiment possible" (Campbell & Stanley 1966:34). Also, see the text authored by Cook and

Campbell (1979). The label, quasi-experiment, generates other useful terms such as “quasi-experimental design" and "quasi-treatments." Clarity of the literature will be well served if statisticians and scientists stop using the term treatments outside the context of manipulative experiments, and the term quasi-treatments may help do that. For example, in the observational study of survival of naturally spawned parr before and after a habitat improvement project is implemented in a stream, it would be useful to refer to the project as a quasi-treatment.

At the broadest level we propose adopting the traditional observational-experimental dichotomy, recognizing that observational studies, including quasi-experiments, can be just as complex and just as effective as manipulative experiments. They have different functions and ideally complement each other. We propose that a review of these concepts and development of common terminology are important for effective communication among researchers, administrators, and policy makers. Although we illustrate this paper and our various recommendations with fisheries and wildlife applications because these are the disciplines most affected by our supporting agencies, we believe that the audience includes workers from all areas of science. We conclude by making recommendations for how RM&E should meet the challenges of large-scale monitoring of complex entities such as ecosystems. Although our recommendations on terminology differ somewhat from those in previous reports (e.g., ISRP 2005) as a result of input from reviewers in preparation of this manuscript for publication, the basic definitions and substantive content remain unchanged.

Observational Studies

Observational studies involve the collection of data at one or more points in space or time *WITHOUT* some type of random assignment of treatments on entities that would constitute experimental units (Table 1). Quasi-treatments may be involved that include management actions intended to improve fisheries or wildlife habitat at selected sites. Some or all of the areas under study may have been deliberately influenced by the researcher or inadvertently impacted by others apart from the treatments themselves, e.g., uncoordinated implementation of different management actions involving different treatments to improve freshwater tributary habitat or unplanned spills of hazardous substances. The researcher usually does not have control over these actions or other external influences on the system studied. Routine monitoring studies typically yield data that are compared over time and space and as a result fall into this category, e.g., counts of adult anadromous fish passing Bonneville Dam on the Columbia River, and periodic meteorological measurements.

Implementation Monitoring In RM&E.

Implementation monitoring, the monitoring of task completion in a specific project, is one of the simplest types of observational study (Table 2). For example, implementation monitoring data may report miles of stream fenced to exclude domestic livestock, number of culverts removed, acres of invasive plants removed, or numbers of fish tagged. In the initial proposal reviews conducted by the ISRP (ISRP 1997-1, ISRP 1998-1), implementation monitoring was often the only monitoring objective addressed, with no mention of tracking the ecological effects of the proposed restoration actions. Implementation monitoring is needed to evaluate progress of management projects, but rigorous science requires that project results also be measured in terms of benefits to fish and wildlife. In addition to implementation monitoring, the ISRP has recommended that ALL projects should be monitored or, preferably, included in an overarching monitoring program to establish the basic benefit of the project or the cumulative benefit of multiple projects for fish and wildlife (see below for additional discussion).

Census Monitoring in RM&E.

Census monitoring involves the collection and analysis of data at one or more study areas in space or time, with data collected on all units (sites, individuals) within areas (Table 2), i.e., a complete census of units in the project areas or of individuals in the populations is available and measurements are made on each. There is no randomization at any level. Inferences beyond the areas or populations on which measures are taken are based on subjective judgment. Estimates of "sampling" error (e.g., standard errors of estimates and confidence intervals) are not appropriate, because data are available on all units in the area or individuals in the population. Calculation of summary statistics such as means and standard deviations or plotting of frequency distributions will be useful for documenting properties and variability of units in the entire area/population. Often the objective is to quantify trends or changes over time specific to a single study area.

The Action Agencies (2003) chose to refer to census monitoring as Tier 1 (Table 3), because they envisioned that census monitoring would be most used in tracking status, trend, and changes at the landscape scale (very large study areas, such as the entire Columbia or John Day River Basins)[1]. In these applications, census monitoring

[1] The assumption that census monitoring would be used at this spatial level of monitoring prompted the ISRP and ISAB to refer to census monitoring as "Tier 1" in earlier reports. It is possible, however to use statistical monitoring on large scale projects, an activity that is usually uneconomical but possible. To eliminate the inconsistency, we drop the term "Tier 1" in reference to census monitoring.

(Action Agencies' Tier 1) can be a low cost, low level of monitoring on large areas. For example, aerial photography or other remote sensing would typically be used to create census data layers in a Geographic Information System (GIS) for long term monitoring of trends or changes in riparian and other terrestrial habitat in subbasins or watersheds. Often, no (or few) changes are expected on decadal time scales (e.g., geology, soils, land surface form), or changes are expected to be relatively slow (e.g., land use; riparian vegetation patterns).

The ISRP and ISAB have used the words "Tier 1, Tier 2, and Tier 3" in a slightly different manner in past reports, referring more to the way data are collected (i.e., census versus sample) than to the scale of the study. To eliminate potential confusion, we have dropped the use of the word "Tier" when referring to the way data are collected (Table 3).

Census monitoring is appropriate to document direct effects of a project. For example, census monitoring in a project to improve aquatic habitat or to supplement a weak stock of naturally spawning fish with hatchery fish might include complete counts of hatchery and naturally produced adults passing a weir to the spawning grounds. Census monitoring is not necessarily prohibitively expensive or time consuming.

The proper role for census monitoring is often to provide low cost, repeatable, long term, daily (or yearly) data with enough accuracy and precision to detect trend, change, differences, or correlations in the face of background noise. For example, complete counts of adults passing a weir on a study stream to gain access to natural spawning grounds might indicate an increasing trend in the percentage of hatchery fish. The question would arise – Why does that trend exist? When trends or changes are detected, then relatively short-term and hypothesis-driven research projects (i.e., manipulative experiments, see below) can be developed to help explain why the trend or change occurred.

Statistical Monitoring (Sampling) in RM&E.

Statistical monitoring projects are also observational studies involving collection of data on a probabilistic (e.g., a simple random) sample of units from one or more study areas (populations) at one or more points in time (Table 2). Statistical monitoring differs from census monitoring in that classical statistical sampling and analysis methods must be employed to detect status and trends of parameters for the study areas or populations as a whole. Statistical conclusions apply to the total areas or populations sampled, not just the units on which data were collected. Inference based on probabilistic sampling is the topic of statistics books with the word “sampling” in the title (e.g., Cochran 1977, Green 1979, Thompson 1992).

When the objectives include study of habitat, vegetation, water quality, fish populations, etc., using on-the-ground field data collection methods in relatively small study areas (e.g., watersheds compared to entire river basins; “index reaches” instead of whole streams), the Action Agencies (2003) chose to refer to the studies at this scale as Tier 2. They envisioned that statistical monitoring would be most economical and hence most used in these studies.[2] These data collection methods are often labor intensive, and it is not economically feasible to collect data on a census of all units in a study area (e.g., data from all reaches in a branch of the John Day River).

[2] The assumption that statistical monitoring would be used at this spatial level of monitoring prompted the ISRP and ISAB to refer to statistical monitoring as “Tier 2” in earlier ISRP reports. It is possible, however to use census monitoring at this spatial level, an activity that is usually uneconomical but possible. To eliminate the inconsistency, we drop the term “Tier 2” in reference to statistical monitoring.

A good model for the use of statistical sampling in monitoring of salmon abundance status and trend is the Oregon Plan for Salmon and Watersheds Monitoring Program (www.nwr.noaa.gov/pcsrf/Moore/) as implemented in Oregon for coho salmon in coastal streams. The Oregon Plan applied a rigorous sampling design for probabilistic site selection to answer key monitoring questions for estimation of coho distribution and abundance**.** The Council's Fish and Wildlife Program Project #200301700 "Develop and Implement a Pilot Status and Trend Monitoring Program for Salmonids and their Habitat in the Wenatchee and Grande Ronde River Basins" is an example of the current development of statistical monitoring (Action Agencies' Tier 2) for status and trend of salmonids and aquatic habitat over three large subbasins in the Columbia River Basin (C. Jordan, NOAA Fisheries, personal communication).

Statistical monitoring reverts to census monitoring if data are collected on all units in the study areas or populations. For example, if upstream and downstream movement of adult anadromous fish can be perfectly counted at a weir 24 hours per day for a migration season, then the total count is census monitoring of escapement above the weir. If counts are made on a random or systematic sample of 24 hour periods distributed during the migration season, then the total count must be estimated by statistical methods and the result is statistical monitoring.

Individual projects within a set of projects should support broader scale statistical monitoring projects by using common methods to select study sites and common methods for data collection. For example, different projects to monitor habitat in a watershed can most easily provide data for monitoring of habitat at a larger scale if the same probabilistic site selection and field data collection methods are used. The more site selection and data collection methods differ, the more difficult it is to aggregate data to make inferences about larger regions.

Census and statistical monitoring both qualify as serious research in the sense that full census data are being provided or probabilistic conclusions are being drawn about entire study areas or populations. For example, statistical estimates of the number of chinook salmon redds in the Wenatchee River Basin in 2035 might be based on counts in a probabilistic sample of sites from the basin. These approaches, however, limit learning about why trends, changes, or correlations occurred. The causes of the effects detected by census or statistical monitoring usually remain elusive.

Many important census or statistical monitoring projects may not yield results of interest to managers until a significant period of time has passed to establish "baselines" for the study areas, trends or changes are detected, or correlations are replicated. The experience of the ISRP and ISAB is that often 10 to 15 years must pass before status and trend monitoring projects may be effectively assessed and appreciated by managers. It is important that the level of long-term commitment to funding be adequate to conduct the monitoring and fully analyze the data. Uncertainties in funding continuation can threaten the investment made to that point in time.

Impact-Control (IC), Before-After (BA), and Before-After-Control-Impact (BACI) studies.

Census or statistical monitoring can be implemented on one or more points in space and time to give rise to Impact-Control (IC) comparisons between areas, Before-After (BA) comparisons on an area(s), and Before-After-Control-Impact (BACI) designs. We prefer to use the word "reference" rather than "control" and note that these quasi-experiments are observational studies to avoid implying that they can document cause-effect relations with the same certitude that manipulative experiments can. There is usually no possibility for random assignment of the quasi-treatments of reference and impact. Measurements often are taken with the objective of asking if there is: 1) a real pre-impact difference between a potential impact area and a reference area, 2) a real difference between years on a site, or 3) a real change in difference between impact and reference sites pre- and post-impact (e.g., Green 1979, Manly 1992).

If random or systematic samples of study units are selected from the study areas and time periods, then these quasi-experimental studies lead to “design-based[3]” statistical inferences concerning the specific study areas and time periods under study. Estimates of parameters with confidence intervals and statistical measures of precision and accuracy apply to the areas and times studied. Results of statistical tests of “no difference” between sites, statistical power, and regression modeling apply only to the whole areas and time period~~s~~ studied. The researcher may conclude that real trends or differences existed between the areas or times, but cause and effect relationships between the quasi-treatments and documented differences cannot be conclusively demonstrated. General application of results outside of the specific areas and times remains open to question.

Most statistical inferences about cause-effect relations in observational or quasi-experimental studies, including tests of hypotheses, power, and regression modeling,

[3] Design based inferences are inferences that are justified by the design of the study and do not require additional assumptions.

are “model-based[4]”, i.e., the inferences are partially based on assumptions additional to those concerning the numerical properties of the data. This is the case for statistical tests of hypotheses (e.g., Smith et al. 1993), simple and multiple-regression modeling (e.g., Zar 1999), and generalized linear modeling (McCullagh and Nelder 1989). Although assumptions about the numerical properties of the data are often easy to justify, those about the absence of confounding factors or extraneous variables are not. Non-subjective conclusions are restricted to changes, differences, and trends among the specific study areas and time periods involved. Causal relationships remain elusive.

Statistical inferences concerning cause and effect relationships beyond the observed “real” differences among specific study areas and time periods are made under a set of assumptions that is often difficult to justify and are subject to criticism. For example, hard conclusions concerning cause and effect relationships assessed with a BACI design require the assumption that, absent any effects of the imposed quasi-treatment, the magnitude of differences between reference and impact areas would have remained constant over time. Conjectured causal relations might be stated as tentative working hypotheses warranting further study.

Other Model-Based Evaluation Methods

The preceding discussion of evaluation and analysis issues reflects the classical “frequentist” approach to the study of probability and statistical inference. There is a large and growing literature on the use of Bayesian and other model-based tools (e.g., geostatistical methods) that can be applied in the evaluation/analysis stage of research and monitoring (e.g., Isaaks and Srivastava (1989), Lee (1997))**.** Because of the extra

[4] Model based inferences are those that require assumptions (models) on the part of the researcher. For example, confidence intervals around an estimate of a parameter typically require the assumption that the estimator has a normal distribution.

assumptions or models required, such methods have their detractors and cautions, but can be useful when the conditions are judged to be reasonably well satisfied.

Manipulative Experiments

Manipulative experiments incorporate treatments that are randomly assigned to experimental units (Table 1) (Fisher 1935). One or more of the treatments may be designated as a control (or reference). The key difference between observational or quasi-experimental studies and manipulative experiments is that, in the latter, treatments (including control treatments) ARE randomly assigned to study units. These “true” experiments generate the strongest conclusions of research designs concerning causal relations between treatments and effects. They require the minimum amount of assumptions or professional judgment to reach these conclusions. Even in a manipulative experiment, however, the *mechanisms* producing a conclusively demonstrated effect remain conjectural.

Statistical conclusions concerning causal relationships are “design based” in the sense that they are justified by the randomized assignment of treatments and design of the study. Thus, cause and effect conclusions do not require strong assumptions on the part of the researcher, given that other factors potentially influencing the response variables on all experimental units remain relatively constant throughout the study. Classical parametric statistical procedures as well as bootstrapping, permutation methods, and other non-parametric statistical methods are directly applicable and require no strong subjective assumptions about the absence of confounding variables on the part of the researcher.

Generally, manipulative field experiments are conducted for a relatively short time period, i.e., perhaps for only one to five years. Manipulative experiments are relatively more common in laboratory studies than in field studies. Unfortunately, manipulative field experiments on anadromous species of fish with complicated life histories may require several generations and it may be difficult to insure that other factors are either relatively constant throughout the study or that annual variation in factors, such as streamflow, affect all study sites approximately the same.

Effectiveness Monitoring in RM&E.

Effectiveness monitoring is intended for use with those manipulative experiments whose objectives include establishing mechanistic or causal links between management actions and fish or wildlife population response (Table 2). Bisbal (2001) terms this level of effort as effects or response monitoring, the repeated measurement of environmental variables to detect changes caused by external influences. The key phrases here are "establishing mechanistic links" and "detect changes caused by external influences." The Action Agencies (2003) chose to refer to this scale of monitoring as Tier 3 in their classification of study designs (Table 3).

Examples of manipulative experiments leading to effectiveness monitoring would include: 1) projects to evaluate the effects of different levels of fertilization on growth and survival of juvenile salmonids with streams selected randomly for allocation to reference and treatment groups; 2) projects to evaluate the effectiveness of spillways in moving out-migrating smolts past a dam on the Columbia River, in which spill levels were chosen to cover the likely operating range (say, 20% of river flow, 40%, 60% and 80%), and these levels were applied in randomized order for relatively short time periods during the out-migration; 3) laboratory experiments to evaluate the ability of lamprey to navigate two types of fish ladders with lamprey from an available population randomly assigned to the two types of ladders; and 4) projects to evaluate the effectiveness of various watershed habitat treatments on survival of parr with treatments randomly or systematically assigned to watersheds. "Action Effectiveness Monitoring" as defined by the Action Agencies (2003) in their RM&E Plan falls into this category.

A good example of planning for large scale effectiveness monitoring in the Columbia River Basin with a manipulative experiment was the original design of the Idaho Supplementation Study on chinook salmon (Lutch et al. 2003). Randomized assignments of treatments to streams in this large-scale study, however, were not made, and the study has reverted to an observational quasi-experiment. The result is that objective unambiguous conclusions concerning the effects of supplementing naturally spawning fish with hatchery fish (the treatment) are not justified by the study design. Inferences will be based on subjective judgment concerning the validity of assumptions, correlations, and models.

Effectiveness Monitoring with Observational Studies

Observational studies may be conducted when the objective is to establish the effectiveness of various management actions. In these studies the management actions are NOT probabilistically assigned to study units. If census or statistical monitoring is carried out in multiple similar observational or quasi-experimental studies over time and space, corroborative results of the studies can provide compelling evidence for the actual effect of an action. In this inductive sense, census and statistical monitoring in numerous observational studies do allow assessment of causal relations (e.g., Shipley 2000). The ISRP believes that this is the most useful type of study design for determining effectiveness of management actions in large ecosystems such as the Columbia River Basin. Good sampling designs are still required, however, and conclusions still require subjective judgment. Hard conclusions as to causal relations are not justified by the design of the study in even these replicated observational studies, because many uncontrolled factors still can influence the results. For example, it may be possible to infer that the magnitude of flow in the Snake River is correlated with survival of out-migrating anadromous smolts, however it is difficult to reach the conclusion that increased flow causes increased survival because other uncontrolled factors such as water temperature and turbidity are also correlated with survival and flow.

The Evaluation Component of RM&E.

It is important to distinguish evaluation based on data collected as part of long-term monitoring programs (usually standard, everyday, every-year data collection from large areas and over long time periods) and those collected in more focused experimental or observational research projects. Research projects are usually relatively short term, often three to five years, are designed for the testing of specific hypotheses and must have well defined plans for analysis and evaluation. Funding agencies should require that plans for evaluation be described in a proposal for a research project and that results of the evaluation be reported. Evaluation is an equally important part of all long term ecological monitoring, and there MUST BE a perceived need and clear procedure for analysis and full and timely interpretation of data being collected. Real-time evaluation as the data are being collected is important, because it allows detection of unusual events or changes in time for them to be subject to additional scrutiny. It also allows for real-time detection of possible recording or measurement errors. Evaluation in long term ecological monitoring should be possible using simple methods with few assumptions, and periodic re-evaluations are appropriate, because evaluation methodologies will gradually improve. The data should have a long shelf life (in the range of 50 to 100 years minimum). The methods that will be available in, say 2055, for evaluation of long term monitoring data probably have not yet been conceived.

Large-Scale Ecosystem Monitoring and Evaluation

Monitoring is difficult in large ecosystems such as the Columbia River Basin where there are numerous state, county, and city governments, autonomous Indian tribes, and a host of federal government agencies that all have vested interests in the ecosystem. A basic problem common to monitoring large ecosystems is that most fish and wildlife agencies and private organizations have ongoing research and monitoring efforts using different site selection criteria, indicator variables, and data collection methods. Many of these groups now have “good old data,” collected by an array of methods for varying amounts of time, and the agencies and other organizations are reluctant to change methods. Although these groups, in combination, may be spending an exceptional amount of effort and money for monitoring, the disparate methods and metrics mean that it is often difficult or impossible to combine these data into meaningful evaluations on larger scales.

In the Columbia River Basin there are some coordinated efforts for monitoring of anadromous fish. Samples of juvenile anadromous fish are tagged for studies of survival through the hydropower system to the ocean and for estimation of harvest of commercially valuable species. Adult anadromous fish must pass the large dams via fish ladders, and procedures are in place to sample or census the returning adults. Unfortunately, there are no coordinated efforts with common techniques for basin wide monitoring of fish and wildlife habitat, for monitoring basic life history parameters such as reproductive success and survival of populations, or for monitoring meta-populations of fish. Without such coordination, it is usually impossible to combine individual monitoring efforts at local levels to draw meaningful conclusions for large subbasins that cross local political boundaries.

Probably the most promising attempts to provide guidance and achieve some coordination of monitoring methods in the Columbia River Basin are being made by a group of professional statisticians and biologists called the Pacific Northwest Aquatic Monitoring Partnership (PNAMP, see www.reo.gov/pnamp/) and the Collaborative Systemwide Monitoring and Evaluation Project (CSMEP, see www.cbfwa.org) in the Council's Fish and Wildlife Program funded by Bonneville Power Administration. The purpose of PNAMP "…is to coordinate important scientific information at the appropriate scales needed to inform public policy and resource management decisions." Although, the members of PNAMP originally came together to coordinate monitoring of watershed condition, objectives have been expanded to include status and trend monitoring of habitat and populations and effectiveness monitoring of management actions for anadromous fish. The CSMEP seeks to inventory and standardize fish monitoring data for the Columbia River Basin. It is coordinated by the Columbia Basin Fish & Wildlife Authority and collaborators include four state, two Federal agencies, several Indian Tribes, and other entities.

Standardization of ongoing monitoring methods among the various government agencies is a large challenge facing development of successful monitoring of the impact of environmental and conservation management actions on large areas of the Columbia River Basin. It remains to be seen if the group of professionals in PNAMP or the CSMAP can implement real change in methods used by the various government agencies.

Case Study: Large-Scale RM&E Programs for Tributary Habitat Recovery

We draw heavily from the Independent Scientific Advisory Board (ISAB 2003-2) report entitled “A Review of Strategies for Tributary Habitat Recovery.” That report contains recommendations on design of research and monitoring studies to evaluate the effects of actions intended to recover or improve tributary habitat for fish and wildlife. Although the material is written primarily in terms of RM&E of fish populations and aquatic habitat, the basic principles apply equally well to RM&E of terrestrial wildlife populations and habitat.

Two general approaches exist for the collection of empirical data for evaluation of the effectiveness of tributary habitat restoration activities. The first includes an extensive collection of study units, such as a large number of pairs of sites where the primary difference is that one site has a certain habitat improvement (e.g., removal of roads) and the other does not. Any future changes in management should be applied uniformly to both members of a pair. As discussed above, this is considered a type of observational study because the quasi-treatments (e.g., road removal and reference) are typically not randomly assigned within each pair of sites. Hard conclusions (e.g., that road removal reduces sedimentation in streams) are not justified by the design of the study. If enough pairs of sites are obtained and if the various results corroborate one another, then evidence is obtained in an inductive sense that a cause and effect relationship is probable. If strong assumptions are made (e.g., that pairs of sites are well matched in all characteristics except those influenced by the quasi-treatments) then model-based statistical methods can be used to quantify the strength of the relationship.

The second approach is to focus intensive evaluations in a smaller number of units (e.g., watersheds), a monitoring approach the state of Washington has termed Intensive Watershed Monitoring (IWM). The basic premise of IWM is that cause-effect relationships in complex ecosystems can best be understood by concentrating monitoring and research efforts at a few locations. Closely spaced measurements of many variables in space and time are often required to develop a thorough understanding of the processes responsible for habitat or fish and wildlife population response to a management action. Concentration of effort may be able to focus sufficient resources and research expertise to begin to tease apart some of the complex interactions governing ecosystem response to restoration activities.

In the first approach there are many replications of study units or pairs of units, which yield an extensive sampling design. In the second, there are repeated measurements of a larger number of variables within a few (perhaps more controlled) units, giving rise to an intensive study design. Although both approaches would constitute observational quasi-experiments, smaller scale manipulative experiments are often embedded within both intensive and extensive sampling designs.

There are obvious advantages and disadvantages to each approach (Roni et al. 2003). The first approach allows reasonable inferences based on the design of the study, but the number of restoration activities or combinations of activities that can be compared is severely limited. The second approach (e.g., IWM) limits inferences to a smaller number of sites with limited geographical coverage and combinations of restoration activities, but with intense study of more variables, processes and their relationships.

Inferences concerning applicability of conclusions to large regions are based on professional judgment in both of these extensive and intensive designs. The primary disadvantages of both approaches are costs, limited inductive inferences to large regions, and logistical difficulties of dealing with relatively large and long-term monitoring/research projects. Logistical difficulties with unavoidable changes in the study designs, however, should be less with the IWM approach, because fewer sites are required.

Based on our collective professional judgment, we recommend the IWM approach for the evaluation of effectiveness of large scale actions affecting tributary habitat. At the time of this writing, it appears that the Columbia River Basin is moving toward probability based statistical monitoring for status and trend of fish and wildlife populations and habitat, combined with intensive study of a few watersheds using the IWM approach.

Conclusion: Recommendations for Monitoring Ecosystems

We recognize the difficulties inherent in monitoring the many dimensions of ecosystem complexity given limitations on mandate, finances, expertise, and personnel. Moreover, it is not easy to condense our advice into a simple set of recommendations that apply to all conceivable situations on research and monitoring in large ecosystems, such as habitat restoration actions in a major tributary system of the Columbia River Basin. Furthermore, the situations in different parts of an ecosystem are likely to require different approaches. For example, evaluation of effectiveness of habitat actions on forest lands might be integrated with the U.S. Forest Service monitoring procedures, whereas evaluation on private lands may require development of new survey procedures. We believe the following four steps contain the essential elements for developing an appropriate RM&E plan in a large ecosystem:

1. Develop a sound census monitoring procedure for trends based on remote sensing, photography, and data layers in a GIS. Land use and landscape changes in terrestrial and aquatic habitat should be monitored for the smallest units possible (i.e., pixels or sites). Future technology may allow low cost remote sensing of important parameters such as water temperature. Accuracy and precision of data layers in the GIS should be evaluated using “blind” classification of randomly selected units by on-the-ground verification during field visits.

Large-scale census monitoring for trends in populations or habitats might include complete fish counts and condition in juvenile bypass systems at dams, adult fish counts at weirs, or measurement of the volume of large wood in all reaches of a river.

In practice, however, statistical monitoring (Action Agency Tier 2) is often more cost-effective because measurements can be made during a random or systematic sample of units or time periods.

2. Cooperate with system-wide attempts to develop common probabilistic site selection procedures for population and habitat status and trend monitoring. Use common protocols for on-the-ground or remotely sensed data collection. As far as possible, measurements of the different indicator or response variables should be made in close proximity to each other on the same sites. Use of probabilistically selected sites should be implemented as soon as possible to avoid inherent biases in subjectively selected sites, e.g., sites chosen for ease of access.

3. As data are obtained on status and trends of wildlife or fish populations and habitat, develop empirical (e.g., regression) models for prediction of current abundance, presence-absence of focal species, and models for population selection of “preferred” habitat. Potential predictor variables include not only physical habitat variables (vegetation, flow, temperature, etc.), but also measures of habitat recovery actions that are currently in place or are implemented in the future. Use the empirical models to evaluate the relative importance of physical factors and habitat improvements and to predict abundance or presence-absence throughout major sections of the ecosystem.

4. Employ best professional judgment, based on available data, as to whether or not any new research in the spirit of the Intensive Watershed Monitoring approach should be instigated. Most new intensive research should arise as a result of the interaction of existing inventory data with new data arising from population and habitat status and trend monitoring.

We judge that the approach in these four steps is the most likely one to accomplish successful large-scale, long-term RM&E programs in large ecosystems. An extensive long-term status and trend monitoring program identifies important and unexplained trends and changes that leads to identification of the intensive research that, if conducted, would help explain the "why." Although census monitoring by remote sensing procedures and statistical monitoring provide indications of trend and change in indicator variables, the "why" of certain trends and changes is usually not well understood. For example, future status monitoring may indicate that a major and unexpected increase in juvenile fish production occurred in a watershed with high summer water temperature and low flow during the period 2010 to 2020. Why? A population of bull trout might be detected in an area in 2035 where current knowledge and logic indicate they should not exist. Why? Appropriate and relatively short-term research projects should be designed when the causes of trends and changes observed in long-term M&E programs are not obvious and causal mechanisms require elucidation.

Acknowledgements

We acknowledge significant contribution to this document from Erik Merrill, Coordinator for the ISAB and ISRP, Northwest Power and Conservation Council. We appreciate helpful comments on manuscript drafts from Richard Alldredge, present member of the ISRP, Chris Jordan, Tracy Hillman, Stuart Hurlbert and Roland Lamberson, present members of the ISAB, Bruce Marcot, David Marmorek, Peter

Paquet, and Charlie Paulsen. In particular, the considerable experience of Stuart Hurlbert improved the consistency of terminology for use in all ecological sciences.

Literature Cited

Action Agencies. 2003. Research, Monitoring & Evaluation Plan for the NOAA-Fisheries 2000 Federal Columbia River Power System Biological Opinion. Draft September 11, 2003. Action Agencies: NOAA Fisheries, Bonneville Power Administration, U.S. Army Corps of Engineers, and Bureau of Reclamation. Available at www.salmonrecovery.gov/RME_Plan_09-2003.pdf.

Bisbal, G. A. 2001 Conceptual design of monitoring and evaluation plans for fish and wildlife in the Columbia River Ecosystem. Environmental Management 28: 433-453.

Campbell, D.T., and J.C. Stanley. 1996. Experimental and quasi-experimental designs for research. Houghton Mifflin, Boston.

Cochran, W. G. 1977. Sampling Techniques, Third Edition. John Wiley & Sons, New York.

Conover, W. J. 1999. Practical Nonparametric Statistics, 3rd Edition. John Wiley & Sons, New York.

Cook, T.D. and D.T. Campbell. 1979. Quasi-experimentation: Design and analysis issues for field experimentation, Rand McNally, Chicago. 405 pp.

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh.

Gotelli, N.J. and A.M. Ellison. 2004. A primer of ecological statistics. Sinauer, Sunderland, Massachusetts.

Green, R. H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley & Sons, New York.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187-211.

Hurlbert, S.H. 2003. On misinterpretations of pseudoreplication and related matters: a reply to Oksanen. Oikos 104: 591-597.

ISAB 2003-2. A Review of Strategies for Recovering Tributary Habitat. Northwest Power and Conservation Council, the National Marine Fisheries Service, and the Columbia River Basin Indian Tribes, 851 SW 6th Avenue, Suite 1100, Portland, Oregon 97204. Available at: www.nwcouncil.org/library/isab/isab2003-2.htm.

ISRP 1997-1. Review of the Columbia River Basin Fish and Wildlife Program as directed by the 1996 amendment to the Power Act. Northwest Power and Conservation Council, 851 SW 6th Avenue, Suite 1100, Portland, Oregon 97204. Available at: www.nwcouncil.org/library/isrp/isrp_97-1.pdf.

ISRP 1998-1. Review of the Columbia River Basin Fish and Wildlife Program for Fiscal Year 1999. Northwest Power and Conservation Council, 851 SW 6th Avenue, Suite 1100, Portland, Oregon 97204. Available at: www.nwcouncil.org/library/isrp/isrp_98-1.pdf.

ISRP 2005-14. Retrospective Report 1997-2005. Northwest Power and Conservation Council, 851 SW 6th Avenue, Suite 1100, Portland, Oregon 97204. Available at: www.nwcouncil.org/library/isrp/isrp2005-14.htm.

Isaaks, E. H., and R. M. Srivastava. 1989. An Introduction to Applied Geostatistics. Oxford University Press, Oxford.

Krebs, C.J. 1999. Ecological methodology, 2d edn. Addison Wesley Longman, Menlo Park, California.

Lee, P. M. 1997. Bayesian Statistics: An Introduction, Second Edition. Oxford University Press, Oxford.

Lutch, J., C. Beasley, and K. Steinhorst. 2003. Evaluation and Statistical Review of Idaho Supplementation Studies. Idaho Department of Fish and Game Report Number 03-16. Idaho Department of Fish and Game, 600 South Walnut Street, P.O. Box 25, Boise, ID 83707.

Manly, B. F. J. 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London.

Manly, B. F. J. 1992.The Design and Analysis of Research Studies. Cambridge University Press, Cambridge.

McCullagh, P., and J. A. Nelder. 1999. Generalized Linear Models, Second Edition. Chapman & Hall/CRC, London.

Quinn, G.P., and M.J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press, New York.

Roni P., M. Liermann, and A. Steel. 2003. Monitoring and evaluating fish response to instream restoration. Pages 318-339 *In* D. R. Montgomery, S. Bolton, D. B. Booth, and L. Wall, editors. Restoration of Puget Sound Rivers. University of Washington Press, Seattle, Washington.

Scheiner, S.M., and J. Gurevitch, eds. 1993. Design and analysis of ecological experiments. Chapman and Hall, New York.

Shipley, B. 2000. Cause and correlation in biology: A user's guide to path analysis, structural equations, and causal inference. Cambridge University Press, Cambridge.

Smith, E.P., D.R. Orvos, and J. Cairns. 1993. Impact assessment using the before-after-control-impact (BACI) model: concerns and comments. Can. J. Fish. Aquat. Sci. 50: 627-637.

Thompson, S. K. 1992. Sampling. John Wiley & Sons, New York, New York.

Underwood, A. J. 1997. Experiments in ecology. Blackwell, London.

Williams, R. N. [Editor] 2006. Return to the river: Restoring salmon to the Columbia River. Elsevier Academic Press, Amsterdam.

Zar, J. H. 1999. Biostatistical Analysis, Fourth Edition. Prentice Hall, Upper Saddle River, New Jersey.

Table 1. Basic definitions of observational studies and manipulative experiments.

	Basic Definition
Observational study	Collection of data at one or more study units in space or time *WITHOUT* random assignment of treatments to units.
Manipulative experiment	Collection of data at two or more study units in space or time *WITH* random assignment of treatments to units.

Table 2. Objectives of four common types of monitoring classified as observational studies or manipulative experiment.

Monitoring Type	**Class**	**Objective**
Implementation	Observational study	Monitoring of task completion
Census	Observational study	Monitoring of one or more areas in space or time, with data collected on all study units to detect changes and trends, compare areas, etc.
Statistical	Observational study	Monitoring of one or more areas in space or time, with data collected on a probabilistic sample of study units to detect changes and trends, compare areas, etc.
Effectiveness	Observational study	Establishment of mechanistic or causal links between management actions and population responses with conclusions justified by correlation methods, replicated results, and subjective judgment.
Effectiveness	Manipulative experiment	Establishment of mechanistic or causal links between management actions and population responses with conclusions justified by the design of the experiment.

Table 3. Relationship of census and statistical monitoring to Action Agency (2002) Tier 1, 2 and 3 monitoring.

	Census Monitoring	**Statistical Monitoring**
Large-scale Tier 1 monitoring	Usually census monitoring is most appropriate (e.g., remote sensing to create GIS data layers).	Statistical monitoring could be useful in special cases (e.g., in monitoring range condition on BLM land in Oregon)
Small-scale Tier 2 monitoring	Usually census monitoring is not appropriate because of high costs of large number of experimental units and/or on-the-ground labor intensive methods.	Statistical monitoring with known precision and confidence based on a sample of units is usually most appropriate (e.g., juvenile chinook salmon abundance in a sample of reaches of the John Day River).
Effectiveness Tier 3 monitoring	Usually census monitoring is not appropriate because of high costs of large number of experimental units and/or on-the-ground labor intensive methods.	Statistical monitoring with known precision and confidence based on a sample of units is usually most appropriate. Rigorous experimental design is

	required (e.g., evaluation of survival of juvenile salmonids past John Day Dam with different levels of spill).