ISAB

Independent Scientific Advisory Board
for the Northwest Power and Conservation Council,
Columbia River Basin Indian Tribes,
and NOAA Fisheries
851 SW 6th Avenue, Suite 1100
Portland, Oregon 97204
ISAB@nwcouncil.org

Review of the Pacific Northwest Aquatic Monitoring Partnership's "Study Design for Comparing Monitoring Protocols"

Robert E. Bilby
Peter A. Bisson
Charles C. Coutant
Daniel Goodman
Susan Hanna
Nancy Huntly
Eric Loudenslager
Lyman McDonald
David Philipp
Brian Riddell
Dick Whitney, ISRP

March 18, 2005
ISAB 2005-1

Review of PNAMP's "Study Design for Comparing Monitoring Protocols"

Contents

Review of PNAMP's "Study Design for Comparing Monitoring Protocols"

Introduction

At the request of the Northwest Power and Conservation Council and the Pacific Northwest Aquatic Monitoring Partnership (PNAMP), the ISAB reviewed the design and proposed statistical analyses in a study comparing stream habitat monitoring protocols, which will be carried out by PNAMP. The study objectives are as follows:

- *Goal:* To compare wadeable stream protocols used by monitoring groups within the Pacific Northwest.
    1. *Objective 1:* Do protocols differ in their assessment [of] the same physical stream attributes?
    2. *Objective 2:* Does the amount of variation due to crew differ among monitoring groups?
    3. *Objective 3:* Which monitoring group's protocols permit the best discrimination among streams?

According to the review request, "Large-scale monitoring efforts in the Pacific Northwest are currently using different protocols for a basic set of attributes, making it difficult (if not impossible) to share and compare data. There is now widespread understanding that by standardizing protocols or providing statistical crosswalks (regression analysis) among protocols, monitoring efforts will have the ability to share data and increase sample sizes, thereby increasing the statistical power to describe spatial and temporal trends."

PNAMP proposes to compare mid-summer stream measurements taken by crews representing different monitoring organizations in (yet to be) selected stream reaches from the John Day River Basin, Oregon. The streams will be "wadeable", with bankfull widths less than 15 m and gradients of <1% to 6%. Although the study plan did not state why these criteria were chosen, streams bearing such characteristics probably represent a large percentage of small streams surveyed by habitat inventory crews because they include the majority of 2nd- and 3rd-order (HUC7-8) tributaries used by anadromous salmonids. We note that:

- Study reaches comprise only part of the fish-bearing channel network in watersheds and therefore constrain the comparison of monitoring protocols to a subset of conditions.

- The PNAMP comparison will not necessarily apply to monitoring protocols for the habitats of headwater trout or of salmonids spawning and rearing in larger rivers.

Sampling Methods

Descriptions of sampling methods do not provide sufficient detail for full scientific review. For instance, there are many "if possible" contingencies in the description related to the number of crews, number of sites, number of intensive surveys, and methods for determining some of the "true values". The study plan would benefit from explicit description of rationale for experimental design, including, for example the expected need for and value of a particular number of crews or protocols to be compared. The study design now states "at least 3 (if possible)", then "at least two", and does not indicate whether either of these would be a useful sample size. The ISAB believes that at least 3 groups should be used to provide a measure of variance within a protocol, but that the implied emphasis on comparing protocols (i.e., likely comparison of more protocols than of sampling crews within each protocol) is the appropriate emphasis and allocation of effort. The minimum number of crews should be derived by statistical evaluation of the sampling design that would be needed to provide data that can be useful in addressing the project's goals.

Selection of Test Crews

The study plan proposes to compare survey results among several independent survey crews from each of 4-8 different monitoring groups. These groups potentially include federal (AREMP, EMAP), state (WDFW, ODF, CDFG), and tribal organizations. Thus, if 6 monitoring groups volunteered to participate in the study, and each group fielded 3 independent survey crews (2-3 people each?), a total of 18 habitat survey crews would participate in the comparison. Each survey crew will visit 12-18 stream reaches pre-selected for the study to measure in-channel and riparian habitat attributes according to the standard monitoring protocols of their respective organizations. All surveys will be conducted between July 20 and August 15.

The PNAMP study plan does not describe how survey crews will be selected by participating monitoring organizations, other than that they be "independent". If one of the primary objectives of the study is to determine the variation in survey results among crews (Objective 2), then selection of truly representative crews is crucial to obtaining an unbiased result. The ISAB believes it is critical for the crews to represent the full range of survey skills typical of crews working for their sponsoring organizations, so that results reflect the quality and consistency of that organization's monitoring data. There might be a tendency for the organizations and crews to feel they are in competition with one another to come closest to the *true* values for different habitat attributes. Such a feeling could lead to the most experienced field workers being selected to participate in the comparison. This could bias the study, as it could yield an incorrectly low level of variation among survey results if experienced field workers tend to produce more consistent measurements than inexperienced workers.

We urge PNAMP to provide strict guidelines to participating monitoring organizations when selecting crews. The guidelines could include:

- Crew members should accurately reflect the training and field experience of survey crews typical of the organization. If crews are trained by more than one person, then crews trained by different people should be selected by the organization for this study. On-the-ground experience should range from novice to seasoned veteran. One way to ensure that all skill levels are represented on the crews would be to require that each organization select crew members at random from among candidates rated novice, moderately experienced, and very experienced.
- Crews should have experience with a variety of stream settings. Crew members should not be selected because they have worked in the dry forest conditions typical of the John

Day River, but rather should represent a cross section of crews working in all biogeoclimatic regimes included in the organization's management area.

Channel Types

The PNAMP study plan does not describe how the 12-18 stream reaches will be selected, nor was there an explanation of why the study will be confined to the John Day River (the latter was presumably for logistical reasons, but this was not explicitly stated). The ISAB believes the value of the study will be enhanced if stream reaches selected as test sites span the range of geomorphological channel types that meet the site selection criteria, i.e., "wadeable", <15 m wide, and 0-6% gradient. We believe that site selection be stratified according to channel types, with each channel type represented by at least one of the test sites.

One of the most widely used channel classification systems for streams is that of Rosgen (1994), which identifies 7 major and 42 minor channel types according to channel pattern, entrenchment, width-to-depth ratio, sinuosity, slope, and predominant substrate. Although the Rosgen system is used widely in the Columbia River Basin, it is not a process-based system and lacks explanation of how different channel types are formed and maintained.

The ISAB recommends that test sites be stratified according to the channel classification system of Montgomery and Buffington (1997), which is a process-based (and simpler) system of channel typing.

Five Montgomery and Buffington (1997) reach types potentially satisfy the criteria specified by the PNAMP study plan. They are described in the following table.

	Bedrock	**Cascade**	**Step-pool**	**Plane-bed**	**Pool-riffle**
Predominant bed material	bedrock	boulder	cobble/boulder	gravel/cobble	gravel
Bedform pattern	variable	chaotic	vertically oscillatory	none	laterally oscillatory
Dominant roughness elements	streambed, banks	boulders, banks	bedforms (steps, pools) boulders, large wood, banks	boulders and cobbles, banks	bedforms (bars, pools) boulders and cobbles, large wood, sinuosity, banks
Dominant sediment sources	fluvial, hillslope, debris flows	fluvial, hillslope, debris flows	fluvial, hillslope, debris flows	fluvial, bank erosion, debris flows	fluvial, bank erosion, inactive channels, debris flows
Slope (%)	variable	4-25	2-8	1-4	0.1-2
Typical confinement	strongly confined	strongly confined	moderately confined	variable	unconfined
Pool spacing (channel widths)	variable	< 1	1-4	none	5-7
Bankfull recurrence interval (years)	variable	variable	variable	1-2	1-2

Ideally, each reach type could be represented by 3 test reaches in the PNAMP study; however, we recognize that some types may be rare in the John Day River system and some areas may be difficult to access. In addition, it may be difficult to locate cascade and step pool channel types that meet the study requirements for channel gradient. Nevertheless, stratification of test reaches according to channel types would add value to the study by (1) helping to identify certain types of channels where there is considerable discrepancy in habitat survey results, and (2) providing a basis for extrapolating study results to other regions.

Benchmark Habitat Survey

The PNAMP study plan states that during the same time test crews are carrying out their surveys, the sites will be intensively surveyed to map the entire reach. Results will be considered the "true" habitat values to which test survey data from different crews will be compared. Complete mapping of 12-18 reaches will be very time consuming, and this task, along with the surveys using other protocols, may not be possible within the allotted time. Also, it is not clear that this benchmark measurement will provide any more of the "true" estimate for streambank stability than would some of the other protocols. (Streambank stability is often a subjective judgment based on the presence of eroding banks.) Similar difficulties could be encountered with measurements of temperature, shade, and riparian characteristics. It would help if the study plan specified whether reference values for specific attributes will be determined, what methods will be used, and how the "true" data will be verified (or, how variance of the "truth" is estimated). The ISAB believes that certain parameters can be measured on an intensively surveyed map with variance so small relative to the variance of the other competing field procedures that the variance of the former can be ignored. However, this is not the case for many of the parameters under study, and a variance measure should be attached to each "true value" prior to assessing the significance of deviations from "true values" by test crews. A "true" value may not be practically obtainable for many of the parameters that crews and protocols are to estimate, but the comparisons among protocols (both relative variances and correlations between them) and the crosswalk models that relate protocols remain of value.

Flow and Temperature Changes During the Test Period

The study plan apparently assumes the surveys will be carried out during summer low flow conditions when habitat attributes are not likely to vary over the test period, so that results will reveal differences among monitoring protocols, not survey differences caused by environmental change. The ISAB believes there is good reason to suspect that both flow and temperature will change over the sampling window, and these changes could bias study results. Changes in streamflow will affect pools and residual pool depth, and natural seasonal and diurnal thermal changes will make it very hard to obtain comparable temperature data.

We found no long-term stream discharge records for small streams in the John Day basin; however, the USGS has monitored discharge in the North Fork John Day River at Monument, Oregon, since 1925 (http://waterdata.usgs.gov/or/nwis/uv?14046000). The long-term average flow in the North Fork on July 20, the beginning date of the PNAMP survey comparison, is 264 cfs. On August 15, the final day of the study, the long-term average flow is 125 cfs. This represents a reduction in discharge of about 50% over the study interval. We do not know if small streams in the vicinity experience a similar reduction, but the magnitude of difference between July 20 and August 15 suggests that discharge should be measured frequently at the study reaches and appropriate correction factors applied to pool and residual pool depth data to standardize for discharge. Furthermore, care should be taken when determining reference conditions during the intensive surveys. The window for intensive benchmark surveys is July 1 – August 30. Over this interval, long-term average flows in the North Fork John Day at Monument drop from 736 cfs to 112 cfs – a reduction of 85%! Thus, the larger the time differential between test surveys and reference surveys, the greater the potential for measurement discrepancy in data that may not result from inaccurate monitoring methods, but simply from changes in streamflow.

Likewise, stream temperature is likely to change over the same period, as well as from day to day due to local weather and from hour to hour due to natural diurnal variation. In small streams, our experience is that seasonal differences in water temperature over this period of 5^{o}C are not uncommon, daily differences due to weather can result in shifts of several degrees, and a diurnal temperature spread of 5-10^{o}C is not atypical for small streams in this area. The ISAB believes that comparisons of temperature monitoring among test crews are best carried out if the crews measure water temperature at the same reach, on the same day, as temporally close to each other as possible – say, within one-half hour, and during mid-afternoon, when water temperatures are typically at their daily peak and relatively stable.

In-channel Attributes

It was somewhat surprising that the monitoring comparison did not include stream habitat type (Hawkins et al. 1993) surveys. Categories of riffles and pools are often included in fish habitat inventories, and habitat classification data are commonly used to assess suitability of stream reaches for different life history needs, such as spawning areas or overwintering sites.

It was not clear from the PNAMP study plan how comparisons would deal with monitoring protocols that measured different aspects of some attributes. Examples include:

- Sediment. Commonly used methods to assess sediment include pebble counts, estimates of substrate "embeddedness", the percentage of fine sediment (as determined by visual estimates or quantitative substrate sampling), and turbidity. Additionally, there are alternative ways of expressing sediment abundance and particle size composition. The study plan implies monitoring comparisons will enable correlations among habitat survey results that permit data crosswalks, but this category will be problematic when the aspects of substrate composition being measured or estimated are quite different.
- Riparian. This is another category for which it will be difficult to compare survey results. Some protocols simply include the percentage of the riparian zone covered by vegetation of different types (herbaceous, shrub, tree); others focus on the potential for delivery of large wood to the channel by determining the number of potentially recruitable trees within a specified distance from the channel's edge; still others characterize plant communities using traditional ordination or plot-transect methods. These protocols measure very different things, and correlations will likely be meaningless unless monitoring comparisons focus on a particular aspect of riparian vegetation.

Objective 1 is to answer the question "Do protocols differ in their assessment of the same physical stream attributes?" As illustrated by the above two examples, the protocols may measure different aspects of a stream attribute. Table 1 in the study plan (reproduced below) lists the parameters to be measured. We encourage PNAMP to give precise definitions for each parameter (or adequate reference to published definitions) to be measured by a given protocol. Our experience is that it often is difficult to decide what protocol is best because the protocols are designed to measure different parameters. The definitions may be precise and well defined for some of the parameters, such as "bankfull width", but are likely not well defined for other parameters such as "wood", "sediment", "bank stability", and "riparian characteristics".

The ISAB also encourages the participants to carefully evaluate the time required to complete the various protocols and the associated costs of material and labor. Our judgment is that the standardization of protocols is more about balancing budgets while allowing data to be collected on a large sample size (number of sites), well dispersed throughout the study areas with "reasonable" coefficients of variation and signal to noise ratios, than it is about obtaining the "gold standard" protocol, which may be too time-consuming and expensive to implement over a large number of sites.

Table 1. Physical in-channel attributes to be measured by each monitoring group (or specific protocol).

Bankfull width
Reach length
Bankfull depth
Sediment
Wood
Gradient
Pools
Residual pool depth
Bank stability
Temperature
Shade
Riparian characteristics

Finally, it seems highly unlikely that the rather meager attempt to compare macroinvertebrate samples in the PNAMP study plan will yield useful information. Only a single crew from each organization may be asked to collect macroinvertebrate samples, eliminating any chance of estimating variation among crews within a single macroinvertebrate sampling protocol. There is no information given about how the "true" values for macroinvertebrates will be determined at each site (or for that matter what "true" is, because these organisms vary so much). Macroinvertebrates are notoriously difficult to sample because of their patchy distributions and seasonal changes in abundance. Karr and Chu (1998) recommend taking 3 replicate samples from a single riffle in September. Even if this could be done – the recommended sample time is outside the test survey window – it is doubtful that the macroinvertebrate assemblage of a single riffle in a small stream would remain undisturbed by 6-8 crews, each taking 3 replicate samples. The ISAB therefore suggests that the macroinvertebrate component of this study be delayed until such time that a more complete and thorough comparison of monitoring protocols can be undertaken. Such a study should include not only a comparison of collecting methods, but also a comparison and analysis of indices commonly derived from macroinvertebrate bioassessment data, e.g., the Index of Biotic Integrity (IBI) and the River Invertebrate Prediction and Classification System (RIVPACS).

Analytical Methods

The statistical analysis is inadequately described, and it is not clear that survey data will be appropriately analyzed by forcing them into a standard software package such as PROC MIXED of SAS. In particular, standard ANOVA with the assumption of fixed effects for protocols and streams will require transforming data to obtain "equal variances" before testing formal null hypotheses. By back-transforming the results and keeping track of the original unequal variances, the analysis should be useful, but PNAMP should take care to focus on the project's original objectives, which may be better served by a different analysis. The plan appears to overemphasize hypothesis testing, rather than focusing on quantification of contributions of different sources (e.g., crews, protocols) to patterns of variance in the survey data.

We suggest that the analysis be reduced to its basic structure, e.g., estimate the variance between crews for each habitat feature by measurement protocol and stream type (assuming streams are stratified as suggested above). Compute the average variance for each habitat feature by measurement protocol, averaging across streams. Estimate the magnitude of differences between survey data for a given habitat feature using confidence intervals accounting for unequal variances. Estimate differences between measurements obtained by different protocols and the "truth" for those parameters that have a reliable "true" value. The emphasis should be on "estimation" of results of different survey protocols with measures of precision rather than on "testing" a null hypothesis of no difference among protocols with a particular size test (alpha), coupled with conservative multiple comparison procedures such as Tukey's adjustment. A skilled biometrician can get statistical parameters such as coefficients of variation, signal to noise ratios, and the estimates of variances, effects, and contrasts between effects with unequal variances needed to establish significant differences among protocols using appropriate software, but we would suggest checking the results by obtaining basic descriptive statistics and simply doing direct comparisons.

Careful attention should be given to institutional barriers to "…providing statistical crosswalks (regression analysis) among protocols." We encourage the agencies to consider adoption of standardized protocols, but agencies (and individuals) are often reluctant to change methods. In addition, if a long data record exists and a protocol is to be changed, it is often necessary to use both protocols for a few years and correlation/regression models to relate results of the two methods. Data collected in this study will allow regression models to be developed for prediction of the value of a parameter as measured by Protocol A using data obtained with Protocol B (and vice versa) for wadeable steams in the John Day Basin in the summer of 2005. The precision of predictions from these models should be among the criteria used to evaluate the protocols. Lack of correlation between two protocols for measuring the same attribute indicates that one or the other or both have unpredictable or inconsistent biases or excessively large variances. If biases are consistent and variances are small, these crosswalk models, based on data collected from the same streams using different protocols over time, may be the most important results of the study.

Literature Cited

Hawkins, C. P., J. L. Kershner, P. A. Bisson, M. D. Bryant, L. M. Decker, S. V. Gregory, D. A. McCullough, C. K. Overton, G. H. Reeves, R. J. Steedman, and M. K. Young. 1993. A hierarchical approach to classifying stream habitat features. Fisheries **18**:3-12.

Karr. J. R., and E. W. Chu. 1998. Restoring life in running waters: better biological monitoring. Island Press, New York, N.Y., USA. 200 p.

Montgomery, D. R., and J. M. Buffington. 1997. Channel reach morphology in mountain drainage basins. Geological Society of America Bulletin **109**:596-611.

Rosgen, D. L. 1994. A classification of natural rivers. Catena **22**: 169-199.

w:\em\ww\isab projects and reports\1 isab final reports\isab 2005-1 pnamp study design.doc