United States
Department
of Agriculture

Forest Service

**Rocky Mountain
Research Station**

General Technical
Report RMRS-GTR-122

March 2004

Testing Common Stream Sampling Methods for Broad-Scale, Long-Term Monitoring

Eric K. Archer
Brett B. Roper
Richard C. Henderson
Nick Bouwes
S. Chad Mellison
Jeffrey L. Kershner

Abstract

Archer, Eric K.; Roper, Brett B.; Henderson, Richard C.; Bouwes, Nick; Mellison, S. Chad; Kershner, Jeffrey L. 2004. **Testing common stream sampling methods for broad-scale, long-term monitoring**. Gen. Tech. Rep. RMRS-GTR-122. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain research Station. 15 p.

We evaluated sampling variability of stream habitat sampling methods used by the USDA Forest Service and the USDI Bureau of Land Management monitoring program for the upper Columbia River Basin. Three separate studies were conducted to describe the variability of individual measurement techniques, variability between crews, and temporal variation throughout the summer sampling season. We quantified the variability between crews and through time, and described the percent of the total variability attributed between crew and seasonal variability. We then estimated the number of samples needed to detect change between managed and reference sites.

Differences among streams accounted for a larger share of the total variability than did differences among observers. Stream variability was greater than 80 percent of the total variability for 12 of the 16 variables measured. This is somewhat surprising given the similarities between the study streams. Observer variability was minimal for stream habitat methods describing reach, streambank, and cross-section variables. Conversely, variability was higher for pool, large woody debris, and substrate variables. Seasonal variation was minimal for stream channel variables with the exception of substrate particle sizes. Sample sizes derived from both observer and stream variability (type I error 0.1, type II error 0.9, minimum detectable change 10 percent) ranged from 10 to 3,502 sites to detect changes between two populations. We believe that these estimates represent an unambiguous and powerful way to display the consequences of variability to scientists and managers.

Keywords: habitat, measurement, seasonal variability, Columbia River Basin

The Authors

Eric K. Archer is a Fishery Biologist with the Fish and Aquatic Ecology Unit in Logan, UT. Eric received a B.S. degree in fisheries management and an M.S. degree in aquatic ecology in 1999 from Utah State University, Logan. Eric joined the Forest Service in 1999 and has worked as a Supervisor of Field Operations and Data Analyst for the PACFISH/INFISH Effectiveness Monitoring Project.

Brett B. Roper is currently the National Aquatic Monitoring Coordinator for the Forest Service. Brett received a B.S. degree in environmental studies and an M.S. degree in forest resource management from Utah State University, Logan, and a Ph.D. degree in fisheries management from the University of Moscow, ID. Brett joined the Forest Service in 1994 and has worked as a District Biologist in Region 6 and a Forest Biologist in Region 1 before moving to his current position with the Fish and Aquatic Ecology Unit in 2001.

Richard C. Henderson is a Fishery Biologist with the Fish and Aquatic Ecology Unit in Logan, UT. Rick received a B.S. degree and an M.S. degree in fisheries biology from Colorado State University and Utah State University, Logan. Rick joined the Forest Service in 1998 and has been acting as the Project Leader for the PACFISH/INFISH Effectiveness Monitoring Project.

Nick Bouwes received a B.S. degree in zoology from the University of Wisconsin, Madison, and an M.S. degree and Ph.D. degree in aquatic ecology from Utah State University, Logan. He currently serves as an Adjunct Professor at Utah State University in the Aquatic, Watershed, and Earth Resource Department, and owns and operates Eco Logical Research in Logan, UT.

S. Chad Mellison is currently a Fish and Wildlife Biologist with the U.S. Fish and Wildlife Service in Reno, NV. Chad received a B.S. degree and an M.S. degree in ecology at Utah State University, Logan. Chad began working for the Forest Service in 2000 and recently began Section 7 consultation as related to the National Fire Plan and on Paiute cutthrout trout recovery.

Jeffrey L. Kershner is the National Fish Ecologist with the Forest Service. Jeff received his B.S. degree in fisheries and an M.S. degree in natural resources from Humboldt State University and a Ph.D. degree in ecology from the University of California, Davis. Jeff currently works as Project Leader for the Fish and Aquatic Ecology Unit and as an Associate Research Professor at Utah State University, Logan.

Acknowledgments

The authors wish to thank all summer technicians from the 1998, 1999, and 2000 field crews for their quality work, comments, and critical review of the sampling methods. We greatly appreciate the statistical consultation and guidance from David Turner, Susan Durham, and Richard Cutler. Reviews by David Hohler, John Potyondy, and Ken Roby greatly increased the clarity and content of this manuscript. Finally, we would like to thank Regions 1, 4, and 6 of the Forest Service; the Idaho, Oregon, and Washington offices of the Bureau of Land Management; and the Fish and Aquatic Ecology Unit for supporting this effort.

Contents

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and number.

Telephone (970) 498-1392

FAX (970) 498-1396

E-mail rschneider@fs.fed.us

Web site http://www.fs.fed.us/rm

Mailing Address Publications Distribution
Rocky Mountain Research Station
240 West Prospect Road
Fort Collins, CO 80526

Testing Common Stream Sampling Methods for Broad-Scale, Long-Term Monitoring

Eric K. Archer
Brett B. Roper
Richard C. Henderson
Nick Bouwes
S. Chad Mellison
Jeffrey L. Kershner

Introduction

The concern over the status of anadromous and resident salmonids has prompted renewed interest in the relationship among habitat quality, quantity, and fish populations in the Columbia River Basin. Chinook salmon (*Oncorynchus tshawytscha*), steelhead (*O. mykiss*), and bull trout (*Salvelinus confluentus*) have all received protection under the Endangered Species Act (1973) because of threats to their habitat, competition from nonnative species, and declining population sizes. While there is concern over the perceived decline in habitat, surprisingly little quantitative data has been consistently collected that can adequately describe the status of stream habitat at the basin scale.

While monitoring that describes changes in habitat conditions at large scales is desirable, it is often difficult to implement a monitoring plan of this magnitude for several reasons. The large land area and associated stream networks provide a logistical challenge for data gathering. The diversity of elevations, basin areas, geologies, and climatic regimes within the region provides a great deal of natural variation among streams. In addition, data gathered at large scales are generally aggregated from a variety of sources, crews, and agencies, resulting in potential incompatibilities and inconsistencies that could inhibit broad-scale data analysis and interpretation.

These challenges are considerable, but the potential benefit of understanding how human-induced habitat changes affect aquatic ecosystems across large landscapes has prompted several agencies to undertake large-scale monitoring programs. The Environmental Protection Agency (EPA), through the Environmental Monitoring and Assessment Program, has developed a nationwide environmental monitoring program to "develop the scientific understanding for translating monitoring data from multiple spatial and temporal scales and assessments of ecological conditions" (Stevens 1994). The U.S. Geological Survey, through its National Water Quality Assessment Program (NAWQA), is developing "long-term consistent and comparable information on streams, groundwater, and aquatic ecosystems to support sound management decisions and policy" (Fitzpatrick and others 1998). The U.S. Department of Agriculture Forest Service, U.S. Department of the Interior Bureau of Land Management, National Marine Fisheries Service, and Fish and Wildlife Service have developed a large-scale monitoring program to determine whether habitat within the areas covered by the Pacific Anadromous Fish Strategy (PACFISH) (USDA/USDI 1995) and the bull trout biological opinion (USFWS 1999) is improving, declining, or being restored as a result of land management activities. A similar effort is occurring in Oregon, Washington, and northern California related to the Northwest Forest Plan (Hohler and others, in press; Mulder and others 1999).

Central to these large-scale monitoring efforts is the assumption that anthropogenic influences on ecosystems can be detected across space and through time. For change to be detected, several things must occur:

1. Monitoring programs must be able to focus on attributes that are altered by human disturbance.
2. The variables and the methods used to measure them must be sensitive enough to detect changes in ecosystems that are a result of human activities.
3. Enough measurements must be taken to adequately characterize the variability of the system of interest.
4. Methods should be quantifiable and repeatable by different personnel at different locations and over different time periods to reduce sampling variability.

While there are a number of studies that have attempted to address the sensitivity of specific attributes

to evaluate stream condition (Wang and others 1996), few have sought to address the variability associated with both observer variability and the environmental heterogeneity (Kaufman and others 1999).

Observer variability can come from a number of sources (Simonson 1993). Monitoring protocols that use subjective rating systems or rely on ocular ratings may be imprecise without some estimate of observer variability (Poole and others 1997; Roper and Scarnecchia 1995). Variability can also be introduced if the measurement relies on subjective evaluation of stream attributes. For example, it is often difficult for multiple observers to consistently identify the starting and ending points of a pool. Temporal variability may also be a problem when the characteristics that are being measured change during a sampling season. Channel characteristics measured during the early summer at the end of snowmelt runoff may look very different during the late summer or early fall (Herger and others 1996; Hilderbrand and others 1999). Sources of variability due to observers can generally be minimized with proper protocol development and crew training, but investigators must still conduct quality assurance testing to understand how much variability exists in their sampling.

Many authors have stressed the need to employ quality assurance testing to understand and minimize the variability associated with field measurements of stream attributes used in monitoring studies (Lazorchak and others 1998; MacDonald and others 1991; Poole and others 1997). In general, quality assurance testing is necessary to ensure that the conclusions drawn from the data are not overwhelmed by the error associated with inconsistent methods and data collection. Specifically, results from quality assurance tests can be used to define the level of precision and accuracy associated with each estimate or a stated probability that the estimate is correct (Taylor and Stanley 1983). These estimates can then be used to evaluate whether the methods are sensitive enough to detect ecologically significant changes (Poole and others 1997).

Past quality assurance tests associated with stream inventory and monitoring have generally tried to define the reliability of a small subset of sampling methods for channel and bank descriptors (Platts and others 1983, 1987), habitat units (Poole and others 1997; Roper and Scarnecchia 1995), and substrate (Marcus and others 1995). More recently, large-scale monitoring efforts have begun to evaluate the full range of methods included in their sampling. Wang and others (1996) evaluated the accuracy and precision of stream habitat, riparian vegetation, and land use variables for Wisconsin's Priority Watershed Program. They were the first to compare precision estimates among observers with *a priori* acceptable levels of variability based on monitoring objectives. The Environmental Monitoring and Assessment Program (EMAP) incorporated quality assurance testing into their annual sampling efforts (Kaufmann and others 1999; Robison 1997). They report the precision of in-channel and riparian vegetation assessment methods used during three sampling seasons.

In 1998, the U.S. Forest Service and Bureau of Land Management designed a large-scale monitoring program to determine whether current land management practices are resulting in improved stream habitat and riparian conditions on Federal lands in parts of the Columbia River Basin (Kershner and others 2001). Measurement variables include a wide range of stream channel, riparian vegetation, biotic, and watershed descriptors. Informal quality testing was conducted during the 1998 and 1999 sampling seasons, and results were used to modify field protocols.

During the spring of 2000, we developed a more formal quality assurance program with the goal of defining the variability associated with the evaluation of stream habitat. We evaluated sources of variability associated with differences among streams, differences among observers at a specific site, differences between crews evaluating the same reach, and changes in stream conditions throughout a summer sampling season.

Our objectives were to quantify the variability in the measurement of attributes used to monitor stream systems and to partition that variability among its sources. We used this information to determine the sample sizes needed to detect change and to identify field methods that should be modified to reduce the error associated with observer measurement.

Study Area Description

The study area is located within the Interior Columbia River Basin on lands managed by the USDA Forest Service and USDI Bureau of Land Management (fig.1). Sample reaches are typically third-order streams that have a gradient less than 2 percent and have bankfull widths ranging from 2 to 10 m (tables 1 and 2). We stress that the results and conclusions in this report should only be viewed in the context of these stream descriptions.

Methods

The field methods for the monitoring project were developed from a variety of published stream survey methods and were designed to describe general site characteristics and stream channel habitat (table 3; Kershner 2002). Monitoring crews received training for a 2-week period prior to the start of the sampling season. Stream surveyors were divided into three groups of six people, and they learned stream sampling

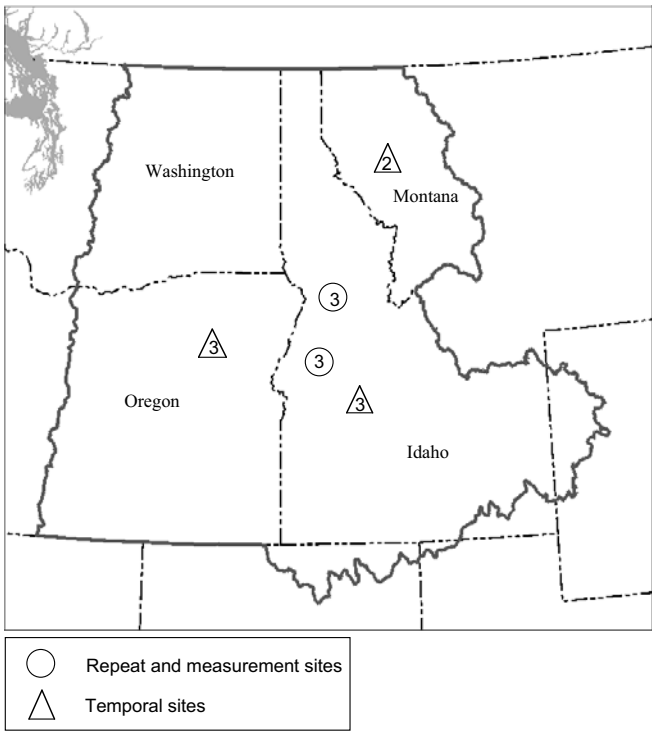


Figure 1—Map of the PACFISH/INFISH study area and the number and type of site at each study location.

methods from one of three crew supervisors. Crews were reassembled for the last 5 days and practiced the sampling protocol on separate streams each day.

Quality Assurance Tests

We assessed the four separate components of variability associated with the evaluation of stream attributes: (1) variability associated with the individual measurement protocol, (2) variability associated with different crews assessing the same sample reach, (3) variability associated with measurements within the same sample reach but collected at different times throughout the summer, and (4) variability associated with streams (environmental heterogeneity). We assessed the variability associated with measurement protocols by having different observers repeat each protocol at the same location (hereafter called "measurement"). Variability associated with different crews measuring the same reach was assessed by giving each crew a fixed starting point for a reach, having the crew establish the boundaries of the reach, and then performing all reach measurements (hereafter called "repeat"). Seasonal variability was measured by identifying fixed reaches and sampling those reaches over three time periods during the summer (hereafter called

Table 1—General information for sites used in the measurement and repeat studies.

Study site	Stream	Date	Geology	Elevation	Gradient	Bankfull width
				ft	*percent*	*m*
Measurement	Boulder	August	Volcanic	4,900	1.60	7.0
	E .B. Price	August	Volcanic	4,200	.66	3.3
	Jack	June	Granitic	5,260	.50	2.5
	Meadow	June	Granitic	3,250	.59	7.0
Repeat	Big	June	Granitic	6,360	.33	4.4
	Boulder	August	Volcanic	4,750	.34	7.7
	Jack	June	Granitic	5,280	.73	2.7
	L. Goose	August	Volcanic	5,000	1.57	3.5
	Lost	August	Volcanic	4,855	.59	5.4
	Meadow	June	Granitic	3,200	.37	8.3

Table 2—General information for sites used in the temporal study.

Stream	State	Geology	Elevation	Gradient	Bankfull width
			ft	*percent*	*m*
Bearskin	Idaho	Granitic	6,460	0.21	7.2
Bench	Idaho	Granitic	6,950	.87	4.3
Fivemile	Oregon	Volcanic	4,720	.39	3.3
Kelsay	Oregon	Volcanic	4,780	1.37	3.3
L. Thompson	Montana	Sedimentary	4,120	1.35	6.5
McGinnis	Montana	Sedimentary	4,080	1.58	8.8
S. F. Desolation	Oregon	Granitic	5,300	.63	9.4
W. F. Sixmile	Idaho	Granitic	4,430	.93	5.4

Table 3—List of aquatic and riparian variables used in the quality assurance testing, included the measurement descriptions or derivations for calculated variables.

Variable	Method
Reach descriptors	**(Variables are measured along a reach defined as a longitudinal section of stream approximately 20 times the bankfull width.)**
Gradient	Measured as the water surface gradient using a level and stadia rod, elevations recorded at the top and bottom of the reach and calculated as the change in elevation by reach length.
Sinuosity	Measured as the reach length along the thalweg divided by the straight-line distance between the top and bottom of the reach.
Bank transects	
Bank angle	Measured by laying clinometer on depth rod at point on bank perpendicular to flow, angle measured to nearest degree, measured at 20 locations on both sides of stream.
Undercut depth	Measured at same location as bank angle, measured as the maximum distance from under bank to bank edge, average of measurements at 20 locations on both sides of stream.
Percent undercut banks	Measured as the number of undercut banks divided by 40; nonundercut banks are rated as "0."
Bank stability	Measured at 30-cm rectangular plots at each bank sampling location and calculated as the number of "stable" plots divided by the total number of plots in the reach.
Covered/uncovered	Perennial vegetation, root cover, or large woody debris cover 50 percent of the bank.
Pools	
Pool tail depth	Measured as the depth at the downstream crest of each pool, measured at the most shallow point in the pool tail.
Maximum pool depth	Measured as the deepest point in a pool.
Residual pool depth	Calculated as the difference between the maximum depth and tail depth.
Pool length	Measured as the length along the thalweg from the head crest to tail crest.
Percent pools	Calculated as the sum of all pool lengths divided by reach length.
Channel cross-sections	
Bankfull width	Measured at the widest point in four riffles and averaged for the reach.
Width-to-depth ratio	Calculated as the average of bankfull width divided by average depth for four cross-sections.
Cross-section maximum depth	Measured as the deepest point in the cross-section.
Cross-section average depth	Depth measured at 10 equally spaced points across the transect and averaged for the cross-section.
Bankfull height	Estimated by the observer as the maximum height at bankfull flow on both banks.
Substrate	
D_{50}	Measured as the median diameter of a minimum of 100 particles sampled from three to four consecutive riffles.
Surface fine sediment in riffles	Uses particle count data from a minimum of 100 particles sampled from three to four consecutive riffles. Fines calculated as the number of particles less than 6 mm divided by the total.
Pool tail fine sediment	Fifty-intersection grid was randomly tossed three times within each pool tail. Percent fines calculated by dividing the number of intersections with fine sediment less than 6 mm by total number of intersections (150) per pool. Total percentage averaged for four pools.
Large wood (LWD)	
LWD pieces per 100 m	Measured as large woody debris that are at least partially within the bankfull channel. Estimate the length and diameter of each piece by placing the depth rod across (width) and along (length) the piece. All singular pieces greater than 3 m in length and at least 10 cm in diameter one-third of the way up from the base are counted. The total number of pieces is divided by the reach length * is multiplied by, times, x 100.
LWD volume per 100 m	Estimated using the length and diameter estimates for each piece and then summed for the total volume of large wood. The total volume was then divided by the reach length * is multiplied by, times, x 100.
LWD submerged volume per 100 m	Observers estimated the percent of each piece that extended below the bankfull elevation. The volume of each piece was multiplied by the percent submerged and then summed for the total volume of large wood submerged at bankfull flows. The total volume was then divided by the reach length * is multiplied by, times, x 100.

"temporal"). Stream variability was estimated by calculating the variance among stream reaches used for evaluating crews at the same reach (hereafter called "environmental heterogeneity").

The measurement, repeat, and environmental heterogeneity studies were conducted during two 1-week periods: the first in mid-June on the Nez Perce National Forest and the second in early August on the Payette National Forest. Both of these Forests are located in central Idaho (fig. 1). Sample sites were specifically chosen to represent the range of stream sizes, gradients, elevations, geology, and conditions observed in this geographic area (table 1). No precipitation fell during either session and streamflows remained stable. During each of the two 5-day sessions, each crew sampled two measurement sites and evaluated three reaches. Each day crews were randomly assigned to either a "repeat" or "measurement" site.

We sampled eight reaches on three separate occasions to assess the variability during the summer sampling period (table 2). The eight sites were distributed throughout the project area, with two in western Montana, three in eastern Oregon, and three in central Idaho. All reaches were sampled during June 27 to 30, August 7 to 9, and September 1 to 7. Crew supervisors identified the starting point for each reach, and crews established the boundaries of the sampling reach using their normal sampling procedure. Each sample location or transect was marked using a combination of survey flags and tags to allow repeated sampling at the same locations.

Analysis

We used a variety of graphical and statistical techniques to look at the data. Data were initially examined for the presence of data entry problems and corrected. All variables used in the statistical models were examined to determine whether they were normally distributed. Non-normal data were transformed using logarithmic or square root transformations prior to statistical testing.

Measurement Data—Descriptive statistics were computed from individual measurements of each of the attributes evaluated at a site and then pooled among sites. The means, standard deviations (SD), and coefficient of variation (CV) were computed for each of the continuous variables. Percent agreement matrices were used to examine agreement between crews for categorical variables. All possible between-crew comparisons were used and summarized as the percent of comparisons where the crews agreed.

Repeat Data—We used a random effects model (PROC MIXED; SAS 2000) to estimate the means and variance associated with crews (observer variability) and streams (environmental heterogeneity). Variance estimates were evaluated by treating both stream and observer as a random effect within the model (Littell and others 1996). We used a random effects model instead of a fixed model because we wished to generalize inferences to a larger population of streams and observers.

Crew error—The crew variance component provided an estimate of variability among crews (all error not associated with streams was assumed to be associated with crews) and an estimate of variability among streams.

We evaluated crew precision by calculating the coefficient of variation among crews and the 95 percent confidence interval. In addition, we calculated the percent variation attributed to crews relative to the overall variability among the streams sampled. Kaufmann and others (1999) considered values less than 33 percent to be acceptable and less than 9 percent to be excellent.

Calculation of minimum sample sizes for observer, stream, and total variation—Sample size estimates are a good tool to evaluate monitoring attributes because they indicate the amount of effort needed to be confident that changes in an attribute will be detected (Eckblad 1991). Minimum sample sizes were calculated using specified differences between two means. This was done independently for observer, stream, and total variation. We evaluated differences between means that ranged from 5 to 50 percent. This range was chosen because differences of these magnitudes likely included changes in attributes that would result in a biological response. We limited our evaluation to a single Type I error rate, $\alpha = 0.1$, and two Type II error rates, $\beta = 0.1$ and $\beta = 0.25$. This single Type I error was chosen because α in published journals are seldom reported above 0.10. Two different Type II error rates were evaluated because these error types are rarely presented and may be more important to management decisions than Type I errors (Peterman 1990).

Estimates of sample size were calculated following the iterative procedure outlined by Zar (1996: 133, equation 8.22):

$$n \geq \frac{2 S_p^2}{d^2}(t_{\alpha(2),V} + t_{\beta(1),V})^2$$

$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$ is the pooled estimate of variance

$v = (n_1 - 1) + (n_2 - 1)$ is the degrees of freedom for S_p^2

$t_{\alpha(2),V}$ = 2-tailed t-value on v df for a Type-I error rate of α (also used for 1-α, two-sided confidence intervals)

$t_{\beta(1),V}$ = 1-tailed upper t-value on v df where β is the acceptable Type II error rate

d = minimum difference to be detected

For these calculations, we used variance estimates from the repeat study as estimates of S_p^2. Total variance was calculated as the sum of streams and observer variation (Clark and others 1996; Montgomery 1984; Ramsey and others 1992). This equation calculates the number of samples needed from each population, and assumes equal sample sizes. If the number of samples from one population is constrained (for example, few unmanaged stream reaches), it will be necessary to adjust the sample size of the unconstrained population. When n exceeded 30, values for infinite sample size were substituted because differences in results were minimal.

One final consideration is when taking a sizeable (more than 5 percent) sample without replacement from a finite population, each observation "carries" more information than when sampling with replacement or from an infinite population. This "extra information" results in a slight decrease in the variance, accomplished by multiplying the usual variance by the finite population correction factor, $(1 - n/N)$ where N is the number of elements of the population and n is the sample size. n/N is known as the sampling fraction. Corrections for finite populations were not included in our sample size estimates.

Temporal Variation—To estimate temporal variation we used a random effects model (Littell and others 1996) to partition the variance associated with differences among the three sampling dates and differences among stream reaches (because of multiple sites). The coefficient of variation, 95 percent confidence interval, and the proportion of total variability attributed to date (percent variability) were used to evaluate temporal precision.

Results

Measurement Study

The ability of observers to use the prescribed field methods and obtain consistent results was somewhat mixed. Variables used to characterize reaches (gradient and sinuosity) were within 20 percent of the mean for all sites (CV 8.4 percent and 6.9 percent, respectively; table 4).

Bank descriptors varied in their usefulness in characterizing streambank condition. Observers were unable to consistently measure bank angle (CV 19.0 percent; table 4), particularly angles from 45 to 90° and 90 to 135°. Most of this variability was associated with the inability of the crews to identify which angle to measure on a complex bank. Observers could not consistently measure undercut depth (CV 20.6 percent; table 4). Estimates of undercut depth were within 20 percent of the mean at 89 of 156 locations. The ability of crews to determine if a bank was undercut varied among observers. Observer agreement at a specific location was generally high with a mean agreement among observers of 88.0 percent (range 84.5 to 95.7 percent). Most of the disagreements occurred

Table 4—Observer summary statistics for each variable at measurement sites.

Category	Variable	Data type	N	SD	CV	Mean	Min	Max	Average agreement
Reach	Gradient	Numeric	8	0.06	8.42	0.81	0.34	1.52	
	Sinuosity	Numeric	8	.11	6.88	1.67	1.06	3.23	
Banks	Bank angle	Numeric	298	17.13	19.00	90.18	.00	179	
	Percent undercut	Categorical	98						88.01
	Undercut depth	Numeric	156	.06	20.60	.29	.00	1.20	
	Bank stability	Categorical	298						82.65
	Stable/unstable	Categorical	298						92.02
X sections	Bankfull width	Numeric	26	.47	10.00	4.66	1.95	13.94	
	Width-to-depth ratio	Numeric	26	2.06	13.75	14.95	4.57	67.30	
Pools	Pool length	Numeric	36	2.88	25.09	11.48	2.40	58.6	
	Pool tail depth	Numeric	36	.03	18.19	.19	.06	.44	
	Maximum pool depth	Numeric	36	.03	4.48	.61	.26	1.05	
LWD	LWD length	Numeric	48	1.61	24.58	6.58	.05	35.00	
	LWD diameter	Numeric	48	.05	20.14	0.27	.10	1.10	
	LWD submerged volume	Numeric	48	.37	55.22	.67	.0	3.17	
Substrate	Pebble B-axis	Numeric	400	2.30	6.61	37.78	4.00	173.0	

where undercuts were small (undercut depths were near 5 cm, and observers disagreed on whether to consider the bank uncut. Observers were consistently able to measure bank stability. When stability was grouped into five different categories, observers agreed 82.7 percent of the time (range = 76.1 to 91.5 percent). Observer agreement increased to 92.0 percent when categories were combined to "stable" or "unstable" (range = 88.4 to 96.3 percent).

Observers precisely measured bankfull width and width-to-depth ratio. Observer variability when measuring bankfull width was generally low (CV 10.0 percent; table 4). We observed similar results when observers calculated the width-to-depth ratio, which is not surprising given our use of the bankfull width in the calculation. The observer CV was less than 20 percent of the mean at 83 percent of the measurement sites (CV 13.8 percent; table 4). The largest deviation in the width-to-depth ratio from the grand mean by an observer was three.

Our ability to characterize pool habitats was influenced by the ability of observers to consistently characterize some pool variables and the variability of the streams sampled. Observers had trouble consistently determining the starting and ending point of pools at measurement sites, and were within 20 percent of the mean length at roughly half (53 percent) of the pools measured (CV 25.1 percent; table 4). Variability in the measurement of residual pool depth occurred from two sources: identification of the pool tail crest and the subsequent measurement of the maximum crest depth, and the identification and measurement of the maximum pool depth. Measurements of the maximum crest depth was less than 20 percent of the mean at 69 percent of the pools (CV 18.2 percent; table 4). Observer variability averaged 20 percent of the mean for maximum depth (CV 4.5 percent; table 4). The subsequent variability between observers for residual pool depths were intermediate with a CV of 11.3 percent (table 4).

Volume estimates of large wood required an observer to correctly measure the length and the diameter of each piece. When measuring length, observer variability was within 20 percent of the mean for 25 of 48 pieces of large wood (CV 24.6 percent; table 4). Precision declined as length increased. Sixty-eight percent of the diameter measurements were within 20 percent of the mean (CV 20.1 percent; table 4).

The submerged volume of each piece of large wood was difficult for observers to estimate, and there was considerable variability at all four sites. Individuals had to correctly identify the height of the channel at bankfull flow and then estimate the amount of wood that was submerged at that flow. Only 30 percent of the pieces measured had a CV less than 20 percent of the mean (CV 55.2 percent; table 4).

Observer error associated with measuring the diameter of substrate particles was low (CV 6.6 percent; table 4). Observers were within 20 percent of the mean value for 94 percent of the particles measured.

Repeat Study

Reach Measurements—The precision of measurements characterizing sample reaches was moderate. Only 2.1 percent of the total variability in measured gradient was attributed to crews at repeat sites (table 5). The largest variation among crews at a site

Table 5—Summary statistics for each variable at repeat sites.

Variable	Mean	Upper confidence interval	Lower confidence interval	Width (±) of confidence interval	Coefficient variation	Percent variability
Gradient	0.56	0.66	0.47	0.09	14.58	2.08
Sinuosity	1.95	2.17	1.72	.23	5.99	13.86
Bankangle	96.44	111.90	80.98	15.46	8.18	15.06
Undercut depth	.10	.14	.07	.04	17.33	17.34
Percent undercut banks	40.92	54.34	27.50	13.42	16.73	14.45
Bank stability	91.80	100.09	83.52	8.29	4.60	47.21
Bankfull width	5.30	6.64	3.95	1.35	12.97	17.38
Percent pools	69.98	87.86	52.10	17.88	13.04	56.11
Residual pool depth	.37	.45	.30	.08	10.50	20.07
LWD pieces per 100 m	8.64	12.52	4.76	3.88	22.89	6.27
LWD volume per 100 m	1.35	2.10	.78	.66	16.64	2.92
LWD submerged volume per 100 m	.84	1.08	.30	.39	29.12	7.94
D_{50}	22.40	33.17	13.73	9.72	11.07	9.56
Surface fine sediment in riffles	31.61	43.46	19.77	11.85	35.71	30.54
Pool tail fine sediment	26.16	43.50	8.82	17.34	33.82	19.55

was 21 percent of the reach mean (CV 14.6 percent, 95 percent confidence interval [CI] 0.09) with more variability occurring at the three longest reaches. We observed similar results for sinuosity. The variability attributed to crews at repeat sites was 13.9 percent (table 5). Precision estimates for sinuosity were low with a CV of 6.0 percent and a CI of 0.23.

Sample size estimates needed to detect changes of 10 percent with a Type I error of 0.1 and a power of 0.9 varied for gradient and sinuosity. Sample sizes to detect changes in gradient are high (n = 2,106; table 8) and low (n = 50) for sinuosity if both crew and sample site variability are included. In both cases the majority of this variability is associated with stream heterogeneity (table 7).

Bank Transects—Fifteen percent of the total variability in measuring bank angle at repeat sites was attributed to crews (table 5). The largest deviation from the grand mean by a crew was 23 percent (CV 8.1 percent, CI 15.5 degrees). The increased precision between measurement and repeat sites suggests that differences at individual points become less important when multiple locations are averaged to compute a reach mean. Sample size estimates to detect a change of 10 percent were generally high (n = 82; table 8), and most of the variability was associated with stream heterogeneity rather than crew variability (tables 6 and 7).

Crew measurements of the undercut depth accounted for 17.3 percent of the total variability (table 5). The largest difference between a crew and the grand mean for a site was 0.05 m (CV 17.3 percent, CI 0.04 cm). Measurement precision was generally higher when averaged by reach than at individual locations. The sample sizes needed to detect a 10 percent change (power – 0.9) were high (n = 321) with stream heterogeneity having the greatest influence (tables 7 and 8).

Fourteen percent of the variability in percent undercut banks was attributed to crews (table 5). The largest difference between a crew mean and the grand mean at a site was 19 percent (CV 16.7 percent, CI 13.4 percent). Sample size estimates were again influenced more by the variability among streams than among crews (tables 6 and 7). Sample sizes needed to detect a 10 percent (power – 0.9) change were large (n = 391) relative to many of the other variables that we measured (table 8).

Observer variability accounted for 47 percent of the variability at repeat sites for bank stability (table 5). The lack of heterogeneity among the six sites (range of 85.7 to 97.3 percent) was reflected in the high value for percent variability attributed to observers. The maximum deviation between two crews at a site was 18 percent (CV 4.6 percent, CI 8.3 percent). Sample sizes to detect a 10 percent change were lower for bank stability than any other variable (n = 10; table 8).

Table 6—Minimum sample sizes needed to detect differences among physical variables for observer variability only. Sample size estimates assume equal size, so values listed below indicate half the total sample needed. The first value listed in each column is the sample size needed to detect the stated change with a Type I error of 0.1 and a Type II error of 0.1. The value in parentheses has the same Type I error rate but a Type II error rate of 0.25.

Stream attribute	Sample sizes needed to detect changes of:				
	5	10	20	30	50
	Percent				
Gradient[a]	175 (110)	44 (29)	12 (8)	6 (4)	3 (2)
Sinuosity	29 (18)	8 (6)	3 (2)	2 (2)	2 (2)
Bank angle	50 (31)	14 (9)	5 (3)	3 (3)	2 (2)
Undercut depth	223 (141)	56 (36)	15 (10)	8 (5)	4 (3)
Percent undercut banks	226 (142)	57 (36)	15 (10)	8 (5)	4 (3)
Bank stability	18 (12)	6 (4)	3 (3)	2 (2)	2 (2)
Bankfull width	124 (78)	31 (21)	9 (6)	5 (3)	3 (3)
Width-to-depth ratio[a]	27 (17)	8 (5)	3 (3)	2 (2)	2 (2)
Percent pools	137 (86)	35 (22)	10 (6)	5 (4)	3 (3)
Residual pool depth	82 (51)	22 (14)	6 (5)	4 (3)	3 (2)
LWD pieces per 100 m	467 (293)	117 (74)	30 (20)	14 (9)	6 (4)
LWD volume per 100m[a]	259 (163)	65 (41)	18 (11)	9 (6)	4 (3)
LWD submerged volume per 100m[a]	795 (500)	199 (125)	50 (32)	23 (15)	9 (6)
D_{50}[b]	91 (57)	24 (16)	7 (5)	4 (3)	3 (2)
Surface fine sediment in riffles	967 (608)	242 (152)	61 (38)	28 (18)	11 (7)
Pool tail fine sediment	784 (493)	196 (124)	49 (31)	23 (15)	9 (6)

[a] Estimates based on natural logarithm transformations.
[b] Estimates based on square root transformations.

Table 7—Minimum sample sizes needed to detect differences among physical variables for stream heterogeneity only. Sample size estimates assume equal size, so values listed below indicate half the total sample needed. The first value listed in each column is the sample size needed to detect the stated change with a Type I error of 0.1 and a Type II error of 0.1. The value in parentheses has the same Type I error rate but a Type II error rate of 0.25.

Stream attribute	Sample sizes needed to detect changes of:				
	5	10	20	30	50
	Percent				
Gradient[a]	8,249 (5,182)	2,063 (1,296)	516 (324)	230 (144)	83 (52)
Sinuosity	171 (107)	43 (28)	12 (8)	6 (4)	3 (2)
Bank angle	279 (175)	70 (44)	19 (12)	9 (6)	4 (3)
Undercut depth	1,063 (668)	266 (167)	67 (42)	30 (20)	12 (8)
Percent undercut banks	1,338 (841)	335 (211)	84 (53)	38 (25)	15 (10)
Bank stability	20 (13)	6 (4)	3 (2)	2 (2)	2 (2)
Bankfull width	930 (584)	233 (146)	59 (37)	27 (18)	11 (7)
Width-to-depth ratio[a]	121 (76)	31 (20)	9 (6)	5 (3)	3 (2)
Percent pools	107 (67)	28 (18)	8 (6)	4 (3)	3 (2)
Residual pool depth	334 (204)	81 (51)	22 (14)	10 (7)	5 (4)
LWD pieces per 100 m	6,964 (4,374)	1,741 (1,094)	436 (274)	194 (122)	70 (44)
LWD volume per 10 0m[a]	8,608 (5,407)	2,152 (1,352)	538 (338)	240 (151)	87 (55)
LWD submerged volume per 100 m[a]	9,213 (5,787)	2,304 (1,447)	576 (362)	256 (161)	93 (58)
D_{50}[b]	857 (538)	215 (135)	54 (34)	25 (16)	10 (7)
Surface fine sediment in riffles	2,198 (1,381)	550 (346)	138 (87)	62 (39)	23 (15)
Pool tail fine sediment	3,227 (2,027)	807 (507)	202 (127)	90 (57)	33 (22)

[a]Estimates based on natural logarithm transformations.
[b]Estimates based on square root transformations.

Table 8—Minimum sample sizes needed to detect differences among physical variables when both observer variability and stream heterogeneity are accounted for. Sample size estimates assume equal size, so values listed below indicate half the total sample needed. The first value listed in each column is the sample size needed to detect the stated change with a Type I error of 0.1 and a Type II error of 0.1. The value in parentheses has the same Type I error rate but a Type II error rate of 0.25.

Stream attribute	Sample sizes needed to detect changes of:				
	5	10	20	30	50
	Percent				
Gradient[a]	8,424 (5,291)	2,106 (1,323)	527 (331)	234 (147)	85 (53)
Sinuosity	198 (125)	50 (32)	14 (9)	7 (5)	3 (2)
Bank angle	328 (206)	82 (52)	22 (14)	10 (7)	5 (3)
Undercut depth	1,282 (806)	321 (202)	81 (51)	36 (24)	15 (9)
Percent undercut banks	1,564 (983)	391 (246)	98 (62)	44 (29)	17 (11)
Bank stability	35 (23)	10 (7)	4 (3)	3 (2)	2 (2)
Bankfull width	1,054 (662)	264 (166)	66 (42)	30 (20)	12 (8)
Width-to-depth ratio[a]	151 (95)	38 (25)	11 (7)	6 (4)	3 (3)
Percent pools	243 (153)	61 (39)	17 (11)	8 (6)	4 (3)
Residual pool depth	405 (255)	102 (64)	27 (17)	13 (8)	6 (4)
LWD pieces per 100 m	7,430 (4,667)	1,858 (1,167)	465 (292)	207 (130)	75 (47)
LWD volume per 100 m[a]	8,867 (5,570)	2,217 (1,393)	555 (349)	247 (155)	89 (56)
LWD submerged volume per 100 m[a]	1,008 (6,286)	2,502 (1,572)	626 (393)	278 (175)	101 (63)
D_{50} (cm)[b]	947 (595)	237 (149)	60 (38)	28 (18)	16 (11)
Surface fine sediment in riffles	3,165 (1,988)	792 (497)	198 (125)	88 (56)	32 (20)
Pool tail fine sediment	4,011 (2,519)	1,003 (630)	251 (158)	112 (70)	41 (26)

[a]Estimates based on natural logarithm transformations.
[b]Estimates based on square root transformations.

Channel Cross-Section Variables—Approximately 12 percent of the total variability in determining the average bankfull width was attributed to crews (table 5). The largest deviation from the grand mean by a crew was 1.6 m (CV 13.0 percent, CI 1.35 m). Crews were able to consistently define and measure the bankfull elevation at three sites, but were inconsistent at the other three repeat sites.

Seventeen percent of the total variability in measuring bankfull elevation was attributed to crews (table 5). Much of the variability that was observed again occurred at the three sites where crews had difficulty agreeing on the bankfull elevation and the subsequent bankfull width. Sample sizes needed to detect a 10 percent difference (power 0.9) in bankfull width and width-to-depth ratio were (n = 264) and (n = 38), respectively (table 8). Stream heterogeneity accounted for most of the observed variability (table 7).

Pool Variables—Fifty-six percent of the total variability in the percent pools was attributed to crew measurement (table 5). The maximum deviation from the mean value was 30 percent (CV 13.0 percent, CI 17.9 percent). Observer and stream variability almost equally contributed to total sample size necessary to detect a 10 percent change (n = 61; table 8).

Approximately 20 percent of the total variability associated with measuring the residual pool depth was attributed to crews (table 5). Among crews, the largest deviation from the mean was 0.17 m (CV 10.5 percent, CI 0.08 m). Minimum sample sizes needed to detect a 10 percent change (power – 0.9) were again primarily influenced by stream heterogeneity (n = 102; tables 7 and 8).

Large Wood—Large wood was present at three repeat sites. Six percent of the total variability associated with counting large wood at repeat sites was attributed to crews (table 5). Crews rarely agreed on the exact number of pieces within a site with the largest difference among crews being 6.8 pieces per 100 m (CV 22.9 percent, CI 3.9 pieces). Sample sizes needed to detect a change of 10 percent were very high (n = 1,858; table 8) and were primarily influenced by stream heterogeneity (table 8). Three percent of the variability in measuring the volume of large wood was attributed to crew error (table 5). Crew agreement was imprecise with a CV of 16.7 percent and CI of 0.66 m^3. Sample sizes were again considerable for this variable if one wishes to detect a change of 10 percent (n = 2,217; table 8). Crew variability as a percentage of the total variability in measuring the submerged volume of large wood was small (7.9 percent), but their ability to consistently describe the submerged volume of a reach was limited (CV 29.1 percent, CI 0.39 m^3; table 5). Sample sizes for change detection in this variable were the largest of any variable sampled (n = 2,502; table 8).

Substrate—The majority of the variability in characterizing the D_{50} was attributed to stream differences with only 9.6 percent of the total error associated with crews (table 5). The largest difference between a crew D_{50} and a stream average was 17.7 mm (CV 11.1 percent, CI 9.7 mm). Sample sizes needed to detect a 10 percent change in D_{50} are generally large (n = 237; table 8) and most of the variability is associated with stream heterogeneity (n = 215; table 7). The percent surface fines in riffles were highly variable with 30.5 percent of the total variability associated with crew error (table 5). Variability between crews was large with a CV of 35.7 percent and CI of 11.9 percent (table 5). Sample sizes for this variable were among the largest of the variables that we measured (n = 792; table 8).

The percent fine sediment in pool tails was also highly variable. Twenty percent of the total variability was associated with crew measurement error (table 5). Crews varied from 15 to 20 percent in their estimates at most sites with the largest difference between a crew and the grand mean being 19 percent (CV 33.8 percent, CI 17.3 percent). Sample sizes needed to detect a 10 percent (power 0.9) change in percent fines were one of the largest of any variable surveyed (n = 1,003; table 8).

Temporal Study

Bank Transects—Bank characteristics changed little throughout the summer. The percent of the total error attributed to seasonal differences were small for bank angle, undercut depth, and percent undercut (10.4, 3.8, and 8.0 percent, respectively; table 9), whereas a higher value was observed for bank stability (25.2 percent). Precision estimates of CV and CI were similar to estimates from repeat sites, suggesting that most of the variability we observed between sampling dates was due to observer error and not seasonal variability (table 9). The only exception was for undercut depth measurements, which had a larger CV at temporal sites (17.3 percent; table 9) than at repeat sites (14.9 percent; table 5).

Pool Variables—Only a small proportion of the variability in percent pools and residual depth was due to seasonal differences (1.2 and 2.9 percent, respectively; table 9). Precision estimates showed little seasonal variability with estimates lower at temporal sites than repeat sites for both variables (tables 5 and 9).

Substrate—Temporal differences in substrate were observed at some sites with no apparent trend. Substrate measurements in riffles remained relatively constant at four sites, whereas the D_{50} increased and the percent surface fines decreased at four sites (figs. 2 and 3). Pool tail fines varied with measurements

Table 9—Summary statistics for each variable at temporal sites.

Variable	Mean	Upper confidence interval	Lower confidence interval	Width (±) of confidence interval	Coefficient variation	Percent variability
Bank angle	106.75	122.01	91.49	15.26	7.29	10.38
Undercut depth	.09	.11	.06	9.00	15.77	8.03
Bank stability	98.33	101.51	95.16	3.17	1.65	25.20
Percent pools	58.22	62.43	54.02	4.21	3.69	1.24
Residual pool depth	.38	.45	.31	.07	9.41	2.93
D_{50}	32.83	53.33	12.33	20.50	31.85	5.95
Surface fine sediment in riffles	35.24	58.12	12.37	22.88	36.94	32.51
Pool tail fine sediment	31.42	58.54	4.29	27.13	44.05	23.93

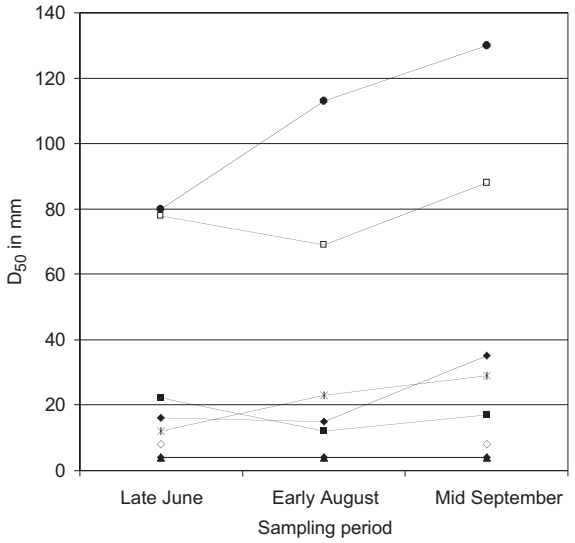


Figure 2—Seasonal change in D_{50} in temporal study streams.

declining at five sites between June and August and then increasing at three sites by September (fig. 4). The percent of the total variability was small for D_{50} (6.0 percent) and higher for the percent surface fines in riffles and pool tail fines (32.5 and 23.9 percent, respectively).

Discussion

The primary purpose of most stream habitat monitoring efforts is to detect differences in habitat characteristics that are caused by anthropogenic changes. Our ability to detect these differences is often clouded by the inherent heterogeneity of streams and the ability of human observers to consistently characterize stream habitat. Understanding the magnitude of these differences and where sources of error might occur allows scientists and managers to

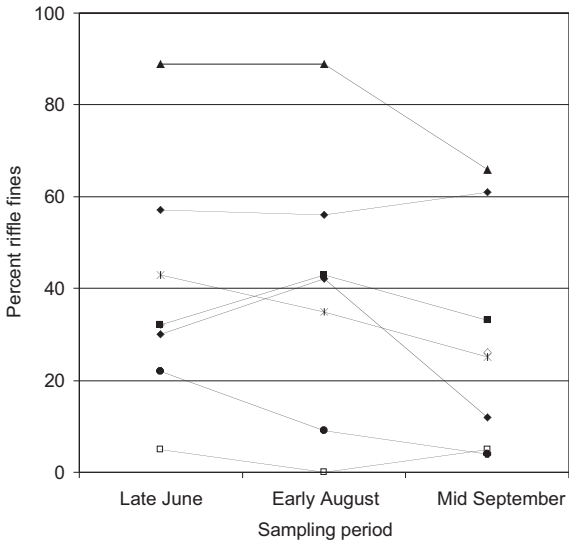


Figure 3—Seasonal change in percent surface fines in riffles in temporal study streams.

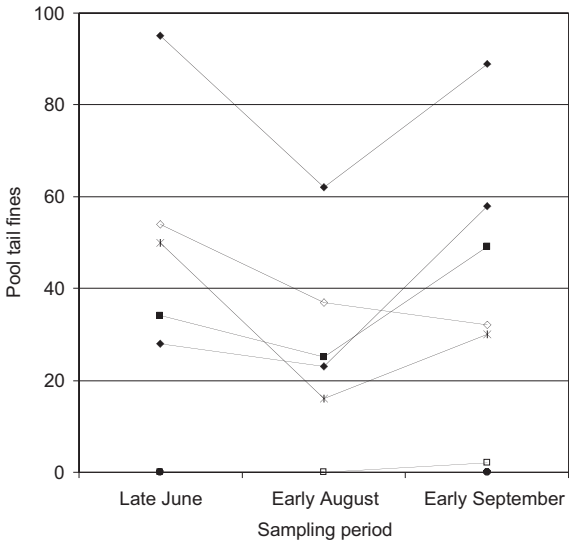


Figure 4—Seasonal change in pool tail fines in temporal study streams.

design meaningful monitoring studies that can account for this variability.

The differences among streams accounted for the largest share of the variability we observed in the repeat study. This is somewhat surprising given the limited differences among the streams where we conducted the sampling. Stratifying sites into homogeneous subareas is generally recognized as one way to reduce site heterogeneity in monitoring studies (Bauer and Ralph 2001; Conquest and others 1994; Green 1984). The use of classification systems may be one way to reduce the variability associated with stream selection (Conquest and others 1994). In particular, spatially nested systems that have defined criteria for both physical and biological attributes may be useful ways to partition streams (Cupp 1989; Frissell and others 1986; Naiman and others 1992; Wang and others 1996). In this study, we confined our repeat sites to two geologic types (granitic and volcanic), and stream gradient was less than 2 percent for all streams. There was significant variability among estimates of gradient in the six repeat streams. Almost all of the variation in gradient was due to differences in streams and not to differences in observer measurement. Given this variability, and the associated sample sizes necessary to detect a change due to management, further stratification may be necessary to improve its use as a monitoring tool.

The variability associated with observer measurement has been the primary focus of most stream habitat quality assurance testing (Kaufman 1999; Marcus and others 1995; Platts and others 1983; Wang and others 1996,). Sources of error generally include inadequate sampling methods to characterize the variable of interest, inconsistent or poorly defined measurement protocols, and a lack of training for field observers (Platts and others 1983; Poole and others 1997; Roper and Scarnecchia 1995). We found only two variables where the variability associated with observer bias was roughly equal to stream heterogeneity (percent pools and bank stability).

Observer identification and measurement of pools has been identified as a problem by other investigators. Variability is associated with the observer's ability to correctly identify a pool and correctly describe the dimensions of the pool (Peterson and Wollrab 1999; Poole and others 1997). Observers in our study generally agreed on pool identification, but had difficulty locating the starting and ending points of the pool. Seasonal differences can provide an additional source of variability in pool identification and measurement (Herger and others 1996; Hilderbrand and others 1999). However, our study suggests this is a minor concern when compared to other sources of variation that must be accounted for in the evaluation of pool habitat.

Surface substrate characterization has been used by a number of investigators to examine anthropogenic influences on stream habitat (Potyondy and Hardy 1994; Schnackenberg and MacDonald 1998). Substrate characterization in this study was confounded by the differences in the streams that we sampled, the observers who took the measurements, and to a lesser degree the time period of the observations. Sample sizes to detect changes at the 10 percent level in pool tail fines (n = 1,003) and percent surface fines in riffles (n = 792) may make them impractical for many monitoring studies. The variability associated with these two attributes is primarily environmental, with less than 30 percent of the variability associated with the measurement technique. Characterization of the percent fines at repeat sites was limited by the ability of observers to measure fine particles. This is consistent with recommendations by other authors to use other means to measure finer particle sizes (less than 4 mm) (Kondolf and Li 1992).

It has been suggested by other authors that more complex categorizations may make it more difficult for crews to consistently characterize environmental features (MacDonald and others 1991; Poole and others 1997). This was particularly apparent for the variable "bank stability." We found that observers were more often in agreement when bank stability was determined to be stable versus unstable as compared to using five categories of bank stability. Roughly half of the variability associated with assessing bank stability at repeat sites was due to observer variability. While this may seem large, the total variation is low compared to other evaluated attributes, and the sample sizes need to detect a change in bank stability were generally small (n = 10; 10 percent change), indicating that there is a reasonable chance to detect changes in stability with a small sample size. Other authors have had mixed success in describing bank stability and attribute this difficulty to the subjective nature of many bank stability protocols (Bauer and Burton 1993; Bauer and Ralph 2001; Platts and others 1987).

Inherent to most monitoring programs is the assumption that the measured stream attributes remain constant throughout the sampling season. Many aquatic monitoring programs sample during the summer low flow season from June to September when discharge is assumed to be relatively constant (Kaufmann and others 1999; MacDonald and others 1991). This assumption has rarely been tested, and most studies either exclude this component by sampling short time periods (usually weeks) or by ignoring seasonal differences. Our results indicate that there were few seasonal differences with the exception of substrate measurements. We observed a general trend of decreasing fines within riffles throughout the summer, indicating that these particles may be transported out of this

habitat type at low flows. Pool tail fines were highly variable between sampling dates but displayed no apparent trend.

Characterization of stream habitat variables and their usefulness in monitoring studies has been attempted by a number of authors. The width of the confidence interval around the mean has been suggested as a useful way to examine the precision of stream habitat measurements when measuring observer variability (Platts and others 1983; Wang and others 1996). The percent variability of an attribute has been suggested as an alternative way to look at measurement precision when one seeks to understand the relationship between observer variability and environmental heterogeneity (Kaufmann and others 1999).

In understanding the way variation among observers affects the results of stream surveys, two concepts are important: percent of the total variation due to observers (signal-to-noise) and the magnitude of variation. The first of these two concepts, percent of the total variation due to observers, has often been used in quality control studies to minimize the intrasite variations. Through extensive training and protocol modification we found that it was possible to reduce observer variability to less than 20 percent of the total variability for most attributes. Studies evaluating water chemistry (Clark and others 1996), geochemistry (Garrett 1969; Ramsey and others 1992), and stream attributes (Kaufmann and others 1999) have suggested that when sampling variance due to observers accounts for less than 20 percent of the total variability, it is likely to be a reliable monitoring attribute. Using this criterion, all but four of our variables appear to provide reliable values.

The problem with relying on the percent total variability attributable to observers ("signal to noise") as the sole guide for selecting monitoring attributes is that this measure fails to account for the total variance associated with the measurement of an attribute. For example, observer variability accounted for 20 percent of the total variability associated with evaluating pool tail fines, but the CV among observers was 34 percent of the mean value.

High CVs result in two problems when evaluating streams: little reliability in interpreting results from a single survey and reduced power in detecting trends in stream attributes. For example, the lower the CV the more likely each individual value will be similar to the mean value if the values are normally distributed. For example, if the CV is 10 then 66 percent of the values will be within 10 percent of the mean value. Increasing variation as expressed as a percentage of the mean (CV) also results in larger sample sizes to achieve a given level of precision (n = 1,003 for pool tail fines) and, therefore, lower power in detecting trends (Eckblad 1991; Peterman 1990) when attributes change at similar rates.

Stream attributes are not all equally affected by land management (MacDonald and others 1991). Some stream attributes such as substrate composition and large wood may change rapidly (less than 5 years) (Benda and Dunne 1997). In contrast, other variables such as gradient, sinuosity, and percent of a reach in pool habitat are likely to respond more gradually (less than 5 years). Therefore, guidance for using CVs to choose attributes to evaluate must account for differences in attribute response timing. Coefficients of variation of 10 or less are likely necessary to detect changes in gradient, sinuosity, or pool percent because changes in these attributes exceeding 10 percent of the reach value are unlikely in short time scales. However, coefficients of variation less than 25 may be suitable when evaluating attributes that change rapidly (substrate, wood). Because of these rapid changes, gross characterization may be sufficient to characterize the large changes these attributes are prone to exhibit.

Combining both of these approaches can provide a unique insight into the problems associated with both observer measurements and differences in streams. Observer measurements of certain variables were consistently rated "good" or "excellent" using the ratings developed by Platts and others (1983). For example, our results for bank stability indicated that observers were generally "good" at describing bank conditions at repeat sites. However, the variability associated with both observers and sites might be a concern when looking at the percent variability estimates. While these measures provide information on where problems might be occurring, a summary is needed to help investigators understand the consequences to proposed or ongoing field studies.

Combining estimates of variability for both streams and observers gives a more realistic view of the significance of source variability. When combined to derive a sample size calculation necessary to detect a change due to management, we believe that this represents an unambiguous and powerful way to display the consequences of variability to scientists and managers. For example, the number of samples necessary to calculate differences in pool tail fines is relatively large (n = 1,003, 1 treatment, 10 percent difference; table 8). Approximately 20 percent of the variability is due to observers and the rest to stream heterogeneity. While pool tail fines is given a rating of "poor" for observer variability, there is no context to understand the consequences to an investigator who may wish to design a monitoring study. Does "poor" mean that we can improve crew training or change methods and potentially improve the precision? In this case, the variability associated with both streams and crews is so high that it is unrealistic to expect a field scientist

or manager to design a stream comparison study that evaluates changes in fine sediment and get meaningful results on a small management area. The cost of gathering data on multiple streams for a small field study might be prohibitively high, given the number of samples that might be needed.

Consequently, field scientists and managers should use caution when developing monitoring studies. Our particular study question in the PACFISH/INFISH monitoring study asks if changes are occurring over a broad landscape. Quality assurance testing is an important component of the study because we sample large numbers of streams and there are multiple crews that collect the data. By estimating variability associated with crew measurements and sample sites it is possible to estimate and obtain the sample sizes necessary to evaluate change. Monitoring studies that attempt to detect changes on a specific stream or small subset of streams may be better suited to establishing permanent cross-sections at specific locations and then establishing a sampling frequency that is adequate to detect change (Elzinga and others 1998). The advantage of this approach is that the differences due to stream heterogeneity are reduced, and the scientist or manager can focus on improving crew performance. While this approach is useful for small, control-treatment studies, it may provide limited insights into processes or management consequences operating at larger spatial scales.

Regardless of the size of the study, it is important to conduct quality control testing for studies where there are hopes of detecting meaningful change and/or statistically significant differences. Understanding the variability associated with the sites that one hopes to measure and the observers that take the measurements allows the investigator to understand the consequences of both when designing a monitoring study.

References

Bauer, S. B.; Burton, T. A. 1993. Monitoring protocols to evaluate water quality effects of grazing management of Western rangeland streams. EPA 910/R-93-017. Seattle, WA: U.S. Environmental Protection Agency. 179 p.

Bauer, S. B.; Ralph, S. C. 2001. Strengthening the use of aquatic habitat indicators in Clean Water Act programs. Fisheries. 26(6): 14–25.

Benda, L.; Dunne, T. 1997. Stochastic forcing of sediment routing and storage in channel networks. Water Resources Research. 33: 2865–2880.

Clark, M. J.; Laidlaw, M. C. A.; Ryneveld, S. C.; Ward, M. I. 1996. Estimating sampling variance and local environmental heterogeneity for both known and estimated analytical variance. Chemosphere. 32: 1133–1151.

Conquest, L. L.; Ralph, S. C.; Naiman, R. J. 1994. Implementation of large-scale stream monitoring efforts: sampling design and data analysis issues. In: Loeb, S. L.; Spacie, A., eds. Biological monitoring of aquatic systems. Boca Raton, FL: Lewis Publishers. 381 p.

Cupp, C. E. 1989. Identifying spatial variability of stream characteristics through classification. Seattle, WA: University of Washington, School of Fisheries. Thesis. 92 p.

Eckblad, J. W. 1991. How many samples should be taken? BioScience. 41: 346–348.

Elzinga, C. L.; Salzer, D. W.; Willoughby, J. W. 1998. Measuring and monitoring plant populations. Tech. Reference 1730-1. Denver, CO: Bureau of Land Management/RS/ST-98/005+1730.

Fitzpatrick, F. A.; [and others]. 1998. Revised methods for characterizing stream habitat in the National Water Quality Assessment Program. Water Resources Investigations Rep. 98-4052. Raleigh, NC: U.S. Geological Survey.

Frissell, C. A.; Liss, W. J.; Wissmar, R. C.; Hurley, M. D. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. Environmental Management. 10(2): AAI99–124.

Garrett, R. G. 1969. The determination of sampling and analytical errors in exploration geochemistry. Economic Geology. 64: 568–574.

Green, R. H. 1984. Statistical and nonstatistical considerations for environmental monitoring studies. Environmental Monitoring and Assessment. 4: 293–301.

Herger, L. G.; Hubert, W. A.; Young, M. K. 1996. Comparison of habitat composition and cutthroat trout abundance at two flows in small mountain streams. North American Journal of Fisheries Management. 16: 294–301.

Hilderbrand, R. H.; Lemly, A. D.; Dolloff, C. A. 1999. Habitat sequencing and importance of discharge in inferences. North American Journal of Fisheries Management. 19: 198–202.

Kaufmann, P. R.; Levine, L.; Robison, E. G.; Seeliger, C.; Peck, D. V. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. Washington, DC: U.S. Environmental Protection Agency.

Kershner, J. L. 2002. Effectiveness monitoring of aquatic and riparian resources in the area of PACKFISH/INFISH and the biological opinions for bull trout, salmon, and steelhead. Final Plan. Logan, UT: U.S. Department of Agriculture, Forest Service. 56 p.

Kershner, J. L.; Cowley, E.; Henderson, R.; Kratz, K.; Martin, D.; Quimby, D.; Stein, K.; Turner, D.; Ulmer, L.; Vinson, M.; Young, D. 2001. Effectiveness monitoring of aquatic and riparian resources in the area of PACFISH/INFISH and the biological opinions for bull trout, salmon, and steelhead. Draft plan. Logan, UT: U.S. Department of Agriculture, Forest Service, and U.S. Department of the Interior, Bureau of Land Management. 50 p.

Kondolf, G. M.; Li, S. 1992. The pebble count technique for quantifying surface bed material size in instream flow studies. Rivers. 3: 80–87.

Lazorchak, J. M.; Klemm, D. J.; Peck, D. V. 1998. Environmental monitoring and assessment program—surface waters: field operations and methods for measuring ecological condition of wadeable streams. EPA/620/R-94/004F. Washington, DC: U.S. Environmental Protection Agency.

Littell, R. C.; Miliken, G. A.; Stroup, W. W.; Wolfinger, R. D. 1996. SAS® System for mixed models. Cary, NC: SAS Institute Inc. 633 p.

MacDonald, L. H.; Smart, A.; Wissmar, R. C. 1991. Monitoring guidelines to evaluate effects of forestry activities on streams in the Pacific Northwest and Alaska. EPA/910/9-91-001. Seattle, WA: U.S. Environmental Protection Agency, Region 10, Water Division. 166 p.

Marcus, W. A.; Ladd S.; Stock, J.; Stoughton, J. 1995. Pebble counts and the role of user-dependent bias in documenting sediment size distributions. Water Resources Research. 31: 2625–2631.

Montgomery, D. C. 1984. Design and analysis of experiments. New York: John Wiley and Sons.

Mulder, B. S.; Noon, B. R.; Spies, T. A.; Raphael, M. G.; Palmer, C. J.; Olsen, A. R.; Reeves, G. H.; Welsh, H. H. 1999. The strategy and designing of the effectiveness monitoring program for the Northwest Forest Plan. Gen. Tech. Rep. PNW-GTR. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Station.

Naiman, R. J.; Lonzarich, D. G.; Beechie, T. J.; Ralph, S. C. 1992. Stream classification and the assessment of conservation potential. In: Boon, P. J.; Calow, P.; Petts, G. E., eds River conservation and management. London, England: 93–124.

Overton, C. K.; Wollrab, S. P.; Roberts, B. C.; Radko, M. A. 1997. R1/R4 (Northern and Intermountain Regions) fish and fish habitat standard inventory procedures handbook. Gen. Tech. Rep. INT-GTR-346. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station.

Peterman, R. M. 1990. Statistical power can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences. 47: 2–15.

Peterson, James T.; Wollrab, S. P. 1999. An analysis of potential stream fish and fish habitat monitoring procedures for the Inland Northwest. Annual Report 1999 to Bonneville Power Administration, Portland, OR. Contract No. 92AI25866, Project No. 92-032-00, 61 electronic pages (BPA Report DOE/BP-25866-1).

Platts, W. S.; Armour, C.; Booth, G. D.; Bryant, M.; Bufford, J. L.; Cuplin, P.; Jensen, S.; Lienkaemper, G. W.; Minshall, G. W.; Monsen, S. P.; Nelson, R. L.; Sedell, J. R.; Tuhy, J. S. 1987. Methods for evaluating riparian habitats with applications to management. Gen. Tech. Rep. INT-221. Ogden, UT. U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. 177 p.

Platts, W. S.; Megahan, W. F.; Minshall, G. W. 1983. Methods for evaluating stream, riparian, and biotic conditions. Gen. Tech. Rep. INT-138. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. 70 p.

Poole, G. C.; Frissell, C. A.; Ralph, S. C. 1997. In-stream habitat unit classification: inadequacies for monitoring and some consequences for management. Journal of the American Water Resources Association. 33(4): 879–896.

Potyondy, J. P.; Hardy, T. 1994. Use of pebble counts to evaluate fine sediment increase in stream channels. Water Resource Bulletin. 30: 509–520.

Ramsey, M. H.; Thompson, M.; Hale, M. 1992. Objective evaluation of precision requirements for geochemistey analysis using robust analysis of variance. Journal of Geochemical Exploration. 44: 23–36.

Reeves, G. H.; Hohler, D. B.; Larsen, D. P.; Kratz, K.; Reynolds, K.; Stein, K. F.; Busch, D. E.; Hays, P.; Tehan, M.; Atzet, T. [In Press]. Aquatic and riparian effectiveness monitoring program for the Northwest Forest Plan. Corvallis, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Robison, E. G. 1997. Reach scale sampling metrics and longitudinal pattern adjustment of small streams. Corvallis: Oregon State University. Dissertation. 254 p.

Roper, B. B.; Scarnecchia, D. L. 1995. Observer variability in classifying habitat types in stream surveys. North American Journal of Fisheries Management. 15: 49–53.

SAS Institute, Inc. 2000. SAS procedures guide. Release 8.0 ed. Cary, NC: SAS Institute, Inc. 441 p.

Schnackenburg, E. S.; MacDonald, L. H. 1998. Detecting cumulative effects on the Routt National Forest, Colorado. Journal of American Water Resources Association. 34(5): 1163–1177.

Simonson, T. D. 1993. Correspondence and relative precision of stream habitat features estimated at two spatial scales. Journal of Freshwater Ecology. 8: 363–372.

Stevens, D. L., Jr. 1994. Implementation of a national environmental monitoring program. Environmental Management. 42: 1–29.

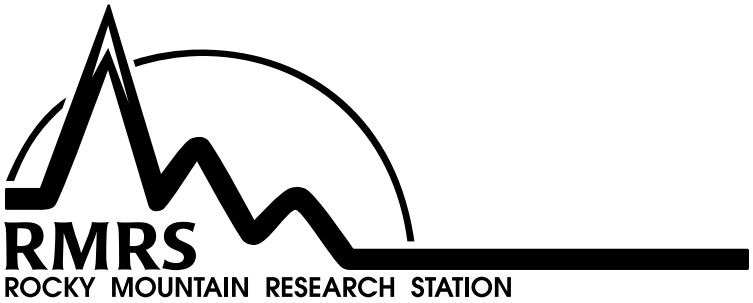
Taylor, J. K.; Stanley, T. W. 1983. Quality assurance for environmental measurements. Proceedings of the conference on quality assurance for environmental measurements; 1983 August 8–12; Boulder, CO. ASTM Special Tech. Publ. 867: 5–11.

USDA Forest Service and USDI Bureau of Land Management. 1995. Environmental assessment for the implementation of interim strategies for managing anadromous fish-producing watersheds in eastern Oregon and Washington, Idaho, and portions of California. Washington, DC. (SS/X.A. 1.3).

USDI Fish and Wildlife Service. 1999. Determination of threatened status for bull trout in the coterminous United States. Federal Register 64(210): 58909–58933.

Wang, L.; Simonson, T. D.; Lyons, J. 1996. Accuracy and precision of selected stream habitat attributes. North American Fisheries Management. 16: 340–347.

Zar, J. H. 1996. Biological analysis. Englewood Cliffs, NJ: Prentice Hall.



The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona
Fort Collins, Colorado*
Boise, Idaho
Moscow, Idaho
Bozeman, Montana
Missoula, Montana
Lincoln, Nebraska
Reno, Nevada
Albuquerque, New Mexico
Rapid City, South Dakota
Logan, Utah
Ogden, Utah
Provo, Utah
Laramie, Wyoming

*Station Headquarters, Natural Resources Research Center,
2150 Centre Avenue, Building A, Fort Collins, CO 80526

Federal Recycling Program  Printed on Recycled Paper