North American Journal of Fisheries Management 30:565–587, 2010

DOI: 10.1577/M09-061.1

[Article]

A Comparison of the Performance and Compatibility of Protocols Used by Seven Monitoring Groups to Measure Stream Habitat in the Pacific Northwest

BRETT B. ROPER
U.S. Forest Service, Forestry Sciences Laboratory, 860 North 1200 East, Logan, Utah 84321, USA

JOHN M. BUFFINGTON
U.S. Forest Service, Rocky Mountain Research Station, 322 East Front Street, Suite 401, Boise, Idaho 83702, USA

STEPHEN BENNETT*
Watershed Sciences Department, Utah State University, Logan, Utah 84322, USA

STEVEN H. LANIGAN
Aquatic and Riparian Effectiveness Monitoring Program, 4077 Research Way, Corvallis, Oregon 97333, USA

ERIC ARCHER
U.S. Forest Service, Forestry Sciences Laboratory, 860 North 1200 East, Logan, Utah 84321, USA

SCOTT T. DOWNIE
California Department of Fish and Game, 1487 Sandy Prairie Court, Suite A, Fortuna, California 95540, USA

JOHN FAUSTINI
Department of Fisheries and Wildlife, Oregon State University, c/o U.S. Environmental Protection Agency, 200 Southwest 35th Street, Corvallis, Oregon 97333, USA

TRACY W. HILLMAN
BioAnalysts, 4725 North Cloverdale Road, Suite 102, Boise, Idaho 83713, USA

SHANNON HUBLER
Oregon Department of Environmental Quality, 1712 Southwest 11th Avenue, Portland, Oregon 97201, USA

KIM JONES
Oregon Department of Fish and Wildlife, 28655 Highway 34, Corvallis, Oregon 97333, USA

CHRIS JORDAN
National Marine Fisheries Service, Northwest Fisheries Science Center, c/o U.S. Environmental Protection Agency, 200 Southwest 35th Street, Corvallis, Oregon 97333, USA

PHILIP R. KAUFMANN
U.S. Environmental Protection Agency, 200 Southwest 35th Street, Corvallis, Oregon 97333, USA

GLENN MERRITT
Washington Department of Ecology, 300 Desmond Drive, Olympia, Washington 98504, USA

CHRIS MOYER
Aquatic and Riparian Effectiveness Monitoring Program, 4077 Research Way, Corvallis, Oregon 97333, USA

ALLEN PLEUS
Washington Department of Fish and Wildlife,
1111 Washington Street SE, 6th Floor, Olympia, Washington 98501, USA

Abstract.—To comply with legal mandates, meet local management objectives, or both, many federal, state, and tribal organizations have monitoring groups that assess stream habitat at different scales. This myriad of groups has difficulty sharing data and scaling up stream habitat assessments to regional or national levels because of differences in their goals and data collection methods. To assess the performance of and potential for data sharing among monitoring groups, we compared measurements made by seven monitoring groups in 12 stream reaches in northeastern Oregon. We evaluated (1) the consistency (repeatability) of the measurements within each group, (2) the ability of the measurements to reveal environmental heterogeneity, (3) the compatibility of the measurements among monitoring groups, and (4) the relationships of the measurements to values determined from more intensive sampling (detailed measurements used as a standard for accuracy and precision in this study). Overall, we found that some stream attributes were consistently measured both within and among groups. Furthermore, for all but one group there was a moderate correlation (0.50) between the group measurements and the intensive values for at least 50% of the channel attributes. However, none of the monitoring groups were able to achieve high consistency for all measured stream attributes, and few of the measured attributes had the potential for being shared among all groups. Given the high cost of stream habitat monitoring, we suggest directing more effort to developing approaches that will increase the consistency and compatibility of measured stream attributes so that they will have broader utility. Ultimately, local monitoring programs should consider incorporating regional and national objectives so that data can be scaled up and the returns to limited monitoring dollars can be maximized across spatial scales.

To meet management objectives and respond to environmental laws and regulations, many state, federal, and tribal agencies monitor the status and trend of stream habitat (Johnson et al. 2001; Whitacre et al. 2007). Physical characteristics of stream habitat are often monitored as a cost-effective surrogate for direct assessments of biological condition (Fausch et al. 1988; Budy and Schaller 2007). These data can also be used to assess watershed condition (Buffington et al. 2003) and degree of landscape disturbance (Woodsmith and Buffington 1996; Kershner et al. 2004). Understanding current stream conditions and how they change through time can be a critical first step to better understanding cause-and-effect relationships between measured stream attributes and the environmental processes that form and alter them. For example, evaluation of historic and long-term monitoring data has resulted in a better understanding of the effects of timber harvest on stream habitat and salmonid production in western North America (McIntosh et al. 1994; Hartman et al. 1996; Isaak and Thurow 2006; Smokorowski and Pratt 2007; Honea et al. 2009). Determining these cause-and-effect relationships is often recognized as a key factor for improving management of stream systems and implementing successful restoration programs (Bilby et al. 2004).

Many aquatic monitoring groups collect data on the status and trend of stream habitat at mesoscales associated with group-specific jurisdiction (e.g., state or management unit levels), but few collect data at broad enough scales and sufficient sampling intensity to evaluate regional or national conditions (Bernhardt et al. 2005; but see EPA 2006a for the exception). If data could be combined across multiple monitoring groups, it would enable larger-scale assessments and greatly increase the statistical power of regional and national assessments (Urquhart et al. 1998; Larsen et al. 2007; Whitacre et al. 2007). Examples of national and regional assessments that could benefit from being able to combine data from different monitoring programs include the U.S. Environmental Protection Agency's (EPA) assessment of surface waters (EPA 2006a) and the National Oceanic and Atmospheric Administration's (NOAA) effort to monitor the recovery of salmon *Oncorhynchus* spp. and steelhead *O. mykiss* in the Pacific Northwest (Crawford and Rumsey 2009). Both of these assessments have general objectives of conducting "baseline status and trend monitoring" and would benefit from increased sample sizes and more widespread sampling.

A review of attributes measured by monitoring groups reveals a large number of commonly measured attributes (Johnson et al. 2001). However, combining data across disparate monitoring groups can be difficult because of differences in group objectives, site selection processes, methods for measuring specific stream attributes (both in general terms and specific details of how and where), and the amount and type of training that monitoring crews receive (Bonar and Hubert 2002; Whitacre et al. 2007). Even when

* Corresponding author: bennett.ecological@gmail.com

Received April 24, 2009; accepted January 22, 2010
Published online May 13, 2010

monitoring groups have similar objectives and measure the same attributes, the measured values may be inherently different from one another because of the above differences. Nevertheless, the potential still exists to combine data across monitoring groups if the measurements within each group are consistent (repeatable) and are correlated to results from other groups. However, consistency and correlation do not guarantee accuracy of measurements, which also must be evaluated. Ideally, attribute measurements for status-and-trend monitoring should be consistent, precise, accurate, and capable of detecting environmental heterogeneity and change.

The goal of this paper is to assess the performance and compatibility of measurements obtained from seven monitoring groups that all use different monitoring protocols to assess stream habitat throughout the Pacific Northwest. This analysis expands on previous work defining acceptable levels of variability within stream habitat protocol data (Kaufmann et al. 1999; Whitacre et al. 2007). To address these issues, we examine (1) the consistency of the measurements within monitoring groups, (2) the ability of each monitoring protocol to detect environmental heterogeneity, (3) the compatibility of the measurements between monitoring groups, and (4) the relationships of the measurements to more intensive stream measurements that may better describe the true character of stream habitat (discussed further below). Understanding how the results of different monitoring programs are related to each other may foster improvement in the quality of stream habitat data, increase the sharing of data across monitoring groups, and increase statistical power to detect environmental trends (Larsen et al. 2007).

Study Sites

Data were collected from 12 streams in the John Day River basin in northeastern Oregon, which ranges in elevation from 80 m at the confluence with the Columbia River to over 2,754 m in the headwaters of the Strawberry Mountain Range (Figure 1; Table 1). This location was selected for several reasons: there was an ongoing collaborative agreement between different state and federal agencies in the state of Oregon; several of the groups had sample sites in the basins; and the logistics of organizing numerous groups were optimal (access, proximity of monitoring groups, and timing).

The John Day basin is located within the Blue Mountains ecoregion (Clarke and Bryce 1997), which encompasses a wide range of climates (semiarid to subalpine) and vegetation types (grassland, sagebrush [*Artemisia* spp.], and juniper [*Juniperus* spp.] at lower elevations to mixed fir [*Abies* spp.], spruce [*Picea* spp.], and pine [*Pinus* spp.] forests at higher elevations). The study sites are underlain by pre-Tertiary accreted island arc terrains, Cretaceous–Jurassic plutonic rocks, and Tertiary volcanics (Valier 1995; Clarke and Bryce 1997).

We used a composite list of randomly selected stream reaches produced by several of the monitoring groups to select our study reaches. We selected sites that were in fish-bearing, wadeable streams that represented a range of channel and habitat types: SP, PB, and PR channels (Figure 2; Montgomery and Buffington 1997), four replicates of each channel type comprising a range of channel complexity (simple, self-formed alluvial channels versus complex, wood-forced ones; Buffington and Montgomery 1999). The result was a set of stream reaches encompassing a range of channel size, slope, and morphology that could be used to detect differences in the performance, compatibility, and accuracy of different monitoring protocols (Table 1).

Methods

Habitat measurements were made at each of the 12 sites using monitoring protocols developed, or used, by seven monitoring groups. We define monitoring groups as independent groups that assess a suite of stream habitat attributes. Protocols are defined as the monitoring group's specific methodologies (including operational definitions, procedures, and training) used to evaluate a suite of attributes. The seven monitoring groups evaluated in this study were the U.S. Forest Service and Bureau of Land Management (USFS–BLM; aquatic and riparian effectiveness monitoring program [AREMP]; AREMP 2005), the California Department of Fish and Game (CDFG; Downie 2004), the U.S. Environmental Protection Agency (environmental monitoring assessment program [EMAP]; EPA 2006b), the Northwest Indian Fisheries Commission (NIFC; Pleus and Schuett-Hames 1998; Pleus et al. 1999; Schuett-Hames et al. 1999), the Oregon Department of Fish and Wildlife (ODFW; Moore et al. 1997), the USFS–BLM (biological opinion effectiveness monitoring program [PIBO]; Dugaw and coworkers, unpublished manual on monitoring streams and riparian areas [available: http://www.fs.fed.us/]), and the upper Columbia monitoring strategy (UC; T. W. Hillman, unpublished report on a monitoring strategy for the upper Columbia basin). The references cited here for each group refer specifically to their 2005 field methods, which were used during this study.

Field crews from each monitoring group sampled a suite of stream habitat attributes at each site following their program's protocols (see exception below), each

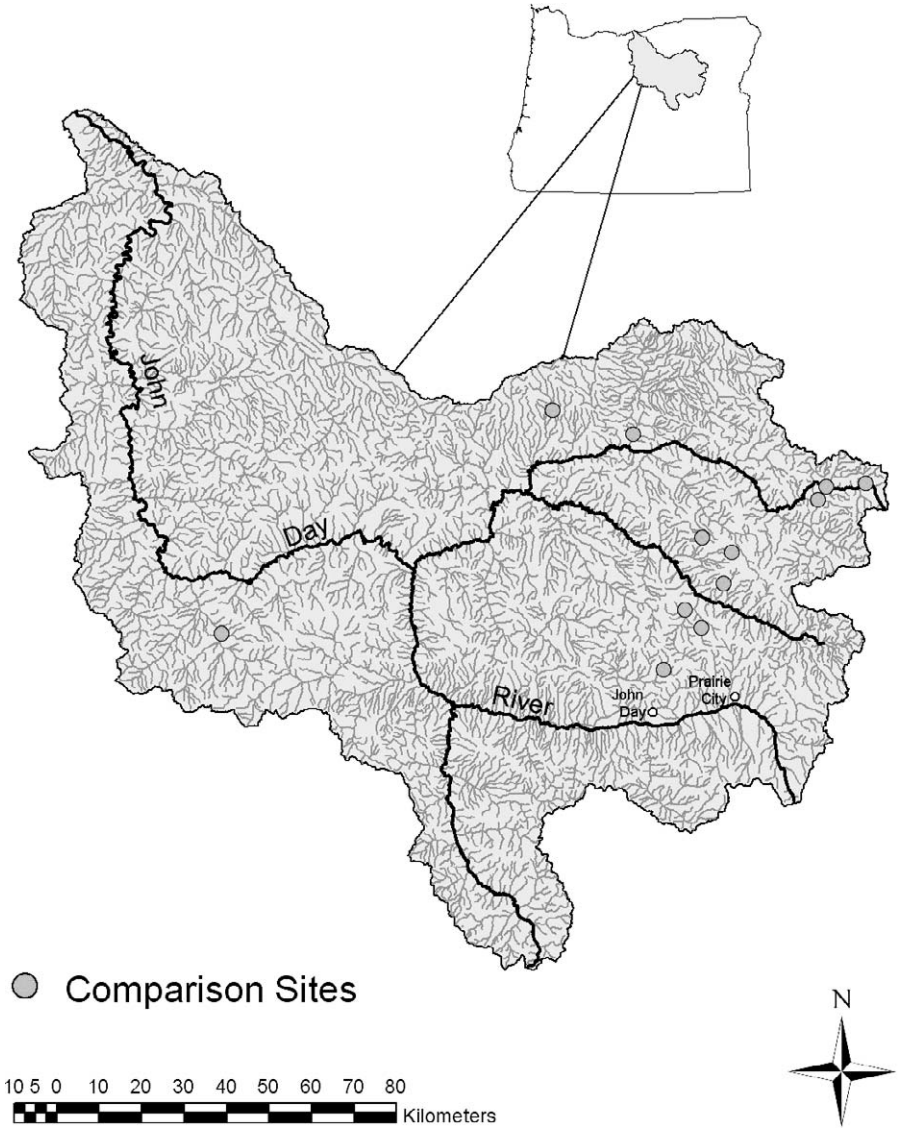

FIGURE 1.—Locations of the study sites within the John Day River basin (modified from Roper et al. 2008).

crew beginning at the same starting point and moving upstream a length defined by their protocol (reach lengths of 20–40 bank-full widths). Average reach lengths evaluated by the monitoring groups ranged from 150 to 388 m (overall average = 256 m; SD = 104 m). Three groups—CDFG, NIFC, and ODFW—evaluated a shorter length of stream than their protocols normally require (approximately 40 times bank-full width) to facilitate comparisons for this study. Modification of a group's standard protocol could lead to nonrepresentative results, but our goal was to compare results obtained over similar sampling domains (i.e., reaches that were 20–40 bank-full widths in length).

Each monitoring group evaluated a stream reach using a minimum of two independent crews. All crews conducted surveys during summer low flow (July 15 to September 12, 2005) and were instructed to "tread lightly" to minimize impact to the channel parameters being measured over the course of the site visits by each of the monitoring groups and their crews. Visual inspection of channel characteristics before and after the study showed little impact of the crews on the measured parameters. Crews completed measurements

TABLE 1.—Characteristics of the 12 streams sampled in the John Day basin. The values for gradient, bank-full width (BFW), and the bank-full width-to-depth ratio (W:D) are averages across all monitoring protocols and field crews. The values for sinuosity, residual pool depth (RPD), and median particle size (D_{50}) are averages across all of the groups that collected these attributes. Large woody debris (LWD) is defined as pieces having a length of 3 m or more and a diameter of 0.1 m or more and includes data from four groups (AREMP, EMAP, PIBO, and ODFW; see text).

Stream	Channel type[a]	Stream order	Elevation (m)	Gradient (%)	Sinuosity	BFW (m)	W:D	RPD (cm)	D_{50} (mm)	LWD/100 m
Bridge	PB	4	655	1.17	1.26	4.5	17.4	21	21	0
Camas	PB	5	847	1.26	1.03	15.7	32.5	24	96	1
Potamus	PB	2	1,295	2.45	1.10	8.9	40.3	22	69	12
Tinker	PB	1	1,411	2.72	1.17	2.5	15.6	17	18	11
Big	PR	1	1,850	1.33	1.42	3.7	14.2	33	4	49
Crane	PR	2	1,630	1.25	1.47	4.2	21.9	38	6	26
Trail	PR	3	1,581	1.78	1.38	6.8	24.5	31	47	34
West Fork Lick	PR	2	1,298	3.34	1.29	3.9	17.7	24	27	12
Crawfish	SP	2	1,816	5.4	1.13	6.5	18.4	33	81	32
Indian	SP	1	1,813	5.8	1.15	4.3	22.5	18	21	48
Myrtle	SP	1	1,444	9.05	1.12	3.2	17.0	15	29	27
Whiskey	SP	2	1,213	6.72	1.10	3.0	16.9	20	40	4

[a] PB = plane bed, PR = pool–riffle, and SP = step pool.

at each stream in a single day, and all reaches were worked on by only one crew at a time except when precluded by logistics. Of the 236 total site visits conducted for this study, two crews were at the same site on the same day only 13 times (<6% of the time).

Crews were selected from each group based on availability and logistics, and not on experience; as such, results from each crew are considered typical for a given monitoring group. The sampling objective was to maximize the total number of crews each group used and randomize their sampling effort across the sampling time frame. Logistical constraints, however, led to differences in the number of unique crews each monitoring group used as well as the time period within which each group took to complete all sampling (i.e., the total number of days from the first to last day of sampling; AREMP = 6 crews/7 d to sample all sites, CDFG = 3 crews/15 d, EPA = 3 crews/27 d, NIFC = 3 crews/25 d, ODFW = 2 crews/37 d, PIBO = 6 crews/33 d, and UC = 3 crews/10 d).

We present data on a selection of 10 physical attributes that were measured by the majority of the monitoring groups. These attributes can be divided into four broad classes: (1) overall reach characteristics (channel gradient and sinuosity), (2) channel cross-section characteristics (mean bank-full width and width-to-depth ratio), (3) habitat composition (percent pools, pools/km, and mean residual pool depth [RPD]), and (4) bed material and channel roughness (median particle size [D_{50}], percent fines, and large woody debris [LWD]/100 m). We provide definitions of the above attributes and a summary of how each monitoring group collected these data in Appendix 1. Many of the groups use methods borrowed from one another, with modifications in some instances, but the approaches are largely variants on the same theme.

In addition to the data collected by the monitoring groups, intensive (i.e., more-detailed) measurements were conducted by staff from the U.S. Forest Service's Rocky Mountain Research Station (RMRS) in an effort to gain more accurate and precise estimates of attribute values compared with the rapid field techniques used

FIGURE 2.—Channel types examined: **(a)** pool–riffle (Crane Creek), **(b)** plane bed (Camas Creek), and **(c)** step pool (Crawfish Creek) (from Faux et al. 2009).

by the monitoring groups in this study. Previous studies have compared internal consistency within groups (e.g., Marcus et al. 1995; Roper and Scarnecchia 1995; Olsen et al. 2005) and compatibility across groups (Wang et al. 1996; Larsen et al. 2007; Whitacre et al. 2007), but not accuracy of measurements. Accuracy and precision may be an issue with the monitoring groups examined in this study as they employ rapid measurement techniques designed to allow sampling of one or more sites per day, resulting in fewer measurements with generally less-precise equipment than the intensive measurements conducted by RMRS (Appendix 1).

The RMRS crew measured attributes over reaches that were 40 bank-full widths in length, channel and flood plain topography being surveyed with a total station (874–2,159 points surveyed per site; 0.4–3.7 points/m^2; 21–57 points per square bank-full width). Cross sections were spaced every half bank-full width along the channel (81 cross sections per site), and the bed material was systematically sampled using a grid-by-number pebble count (Kellerhals and Bray 1971; 10 equally spaced grains per cross section, 810 particles per reach), grains being measured with a gravelometer (e.g., Bunte and Abt 2001). At each site, a longitudinal profile of the channel center line was surveyed with the total station, and the number, position, and function of LWD was inventoried (Robison and Beschta 1990; Montgomery et al. 1995), LWD defined as having a length of 1 m or more and a diameter of 0.1 m or more (Swanson et al. 1976). Pools were visually identified as bowl-shaped depressions (having a topographic head, bottom and tail), residual depths being determined from total station measurements of pool bottom and riffle crest elevations. The average width and length of each pool were measured based on channel morphology and topographic breaks in slope rather than on wetted geometry. Pools of all size were measured, without truncating the size distribution for requisite pool dimensions, and were classified as either self-formed or forced by external obstructions (Montgomery et al. 1995; Buffington et al. 2002). These RMRS surveys required three people, working 4–9 d at each site, depending upon stream size and complexity. A single crew was used for all sites, and no repeat sampling was conducted. Because of time constraints, RMRS data were only collected at 7 of the 12 study reaches (PB = two streams, PR = two streams, and SP = three streams).

Overall, the RMRS crew used more precise instruments than the monitoring groups (Appendix 1): the total station yields millimeter- to centimeter-level precision for measuring stream gradient, channel geometry (width, depth), sinuosity, and RPDs; and the gravelometer reduces observer bias in identifying and measuring *b*-axis diameters of particles (Hey and Thorne 1983). Furthermore, the much higher sampling density of measurements conducted by RMRS provided more precise estimates of parameter values (mean, variance). Finally, the RMRS crew was generally more experienced, composed of graduate students and professionals trained in fluvial geomorphology, while the monitoring groups typically employ seasonal crews with more diverse backgrounds and less geomorphic training. For these reasons, the RMRS measurements were assumed to be of higher precision and accuracy and therefore used as a standard for comparison in this study.

Attribute evaluations.—To assess performance and the potential for data sharing among monitoring groups, we evaluated (1) consistency of measurements within a monitoring group, (2) ability to detect environmental heterogeneity among streams, (3) compatibility of measurements among monitoring groups, and (4) relation of measurements to values determined from the more intensive sampling. While statistical tests are used where appropriate in our analysis, most of our comparisons are evaluated in terms of threshold criteria. Furthermore, because many environmental variables have skewed distributions that often fit the lognormal distribution (Limpert et al. 2001) and are often log transformed prior to analysis, we evaluated how transformations might affect our conclusions based on these criteria. We present results of two attributes that are commonly log transformed (D_{50} and LWD/100 m) in both untransformed units (additive error) and in logarithmically transformed values (multiplicative error; Limpert et al. 2001). We used the natural log ($\log_e$) for all transformations and added 0.1 to all LWD values prior to transformation to remove zero values at sites where no LWD was observed.

Consistency of measurements within a monitoring group.—We assessed a monitoring group's consistency by calculating the root mean square error (RMSE = the square root of the among-crew variance [i.e., SD]) and the coefficient of variation (CV = [RMSE/mean] × 100) for each attribute measured by each monitoring group. The RMSE of a given channel characteristic represents the average deviation of crew measurements within a given monitoring group across all sites, and the CV is a dimensionless measure of variability scaled relative to the grand mean across all sites (Zar 1984; Ramsey et al. 1992; Kaufmann et al. 1999). Using these measures, if all crews within a monitoring group produce identical results at each site, both RMSE and CV would be 0. We used analysis of variance (ANOVA [each of the 12 streams as a block]) to

estimate the grand mean (mean value of the 12 streams averaged over the crew observations for each stream), RMSE, and CV for each of the attributes evaluated by each of the monitoring groups.

The exact value of RMSE defining high, moderate, and low consistency is expected to differ by attribute and by differences in protocols among monitoring groups. Use of RMSE as a measure of consistency is best done when the investigator understands the attribute of interest and how much change in the attribute is meaningful (Kaufmann et al. 1999). Since the use of RMSE as a criterion is dependent upon the situation, we specify the values we consider to represent high, moderate, and low consistency for each parameter examined in this study (Table 2). These RMSE criteria represent what we consider to be meaningful differences in the measured attributes; however, care should be used when applying these criteria to other situations.

In contrast to RMSE, CV is a normalized parameter that can be compared across attributes and groups. We defined a protocol as having high consistency when the CV was less than 20%. This value was chosen because when the CV is low it greatly reduces the number of samples necessary to detect changes (Zar 1984; Ramsey et al. 1992). We defined CV values between 20% and 35% as having moderate consistency. While the upper value (35%) is somewhat subjective, it was chosen because values within this range should facilitate classification (e.g., deciding which class a stream is in) but would be less reliable for comparing mean values across time or space without numerous samples. Finally, CVs greater than 35% were defined as having low consistency because the average difference among observers within a group is greater than one-third the mean. This would suggest that meaningful classification might be difficult (e.g., different observers within the same monitoring group could classify the same stream differently; Roper et al. 2008), making statistical comparisons in time or across locations extremely expensive or impossible due to sample size requirements. A caveat regarding the use of CV is that results can be misleading if regional values are applied to specific field applications because the local and regional means may differ (Kaufmann et al. 1999). Overall, values of RMSE and CV in this study resulted in similar estimates of consistency. When these two metrics differed, we used the value that suggested the greater consistency. We used the value that suggested the higher consistency for comparisons in this paper but suggest researchers decide this on a case-by-case basis depending on how important detection of change in a specific metric is to their particular study.

Adequacy of a monitoring group's protocol to detect environmental heterogeneity.—The ability of a protocol to detect environmental heterogeneity was evaluated using a signal-to-noise (S:N) ratio, which quantifies the difference among streams (signal) relative to the difference among individuals evaluating a stream (noise; Kaufmann et al. 1999). To determine the S:N ratio, a random-effects ANOVA model was used to decompose the total variance into that associated with differences among streams versus variation in crew observations at a stream (all error not due to the main effect of stream site is treated as observer variability; Roper et al. 2002). An S:N ratio of 1 indicates that the variation in an attribute among a set of streams is equal to the variation among observers in evaluating those streams. For reasons described in the next section, we characterize the likelihood of detecting environmental heterogeneity as high when S:N ratio is greater than 6.5, moderate when S:N ratio is between 2.5 and 6.5, and low when S:N ratio is less than 2.5.

TABLE 2.—Root mean square error values used as indicators of consistency (repeatability) among observers. The values chosen for high consistency were indicative of observer differences that would have small biological or physical consequences, while those chosen for low consistency would have substantial consequences. Abbreviations are defined in Table 1.

Attribute	Repeatability		
	High	Moderate	Low
Gradient (%)	<0.5	0.5 to 1.0	>1.0
Sinuosity	<0.2	0.2 to 0.5	>0.5
Bank-full width (m)	<1.0	1.0 to 1.75	>1.75
Bank-full width-to-depth ratio	<2.0	2.0 to 3.0	>3.0
Percent pools	<5.0	5.0 to 10.0	>10.0
Pools/km	<5.0	5.0 to 10.0	>10.0
RPD (cm)	<5.0	5.0 to 10.0	>10.0
D_{50} (mm)	<5.0	5.0 to 10.0	>10.0
Percent fines	<5.0	5.0 to 10.0	>10.0
LWD/100 m	<1.0	1.0 to 5.0	>5.0

Relationships among protocols for a given attribute: data crosswalks and sharing.—For monitoring groups to share data, the values for a measured attribute must be related to each other (i.e., correlated). Correlation requires that S:N ratios, as reflected in the following equation, be high (Faustini and Kaufmann 2007):

$$r^2_{\max} = \frac{\text{S:N}_1}{1 + \text{S:N}_1} \times \frac{\text{S:N}_2}{1 + \text{S:N}_2},$$

where $r^2_{\max}$ is the theoretical maximum coefficient of determination (r^2) between two protocols, and the numeric subscripts indicate the respective protocols. Based on the above equation, if protocols for the same attribute measured by two different monitoring groups had S:N ratios greater than 6.5, then $r^2_{\max}$ could be 0.75 or higher. As such, if S:N ratios are high, it becomes

possible to determine whether one protocol is highly correlated to another. In contrast, S:N ratios of ~2.5 would result in an r^2_{max} of only 0.5 (moderate correlation). When S:N ratios are less than 2.5, they are considered low because even if monitoring groups are measuring the same attribute, variation among observers within a group precludes detecting a relationship (low correlation). While the exact S:N thresholds are somewhat subjective, they meet our objective of providing criteria that assess the likelihood for monitoring groups to share data.

In addition to correlation, results obtained by each monitoring group should be accurate; correlation among groups does not guarantee accuracy of their measurements. Since it is difficult to know the true value of a given attribute, we evaluated compatibility of data among the monitoring groups using two approaches: (1) assessing whether attribute values were correlated between monitoring groups and across channel types (both in terms of the above S:N criteria and r^2_{max}-values), and (2) by comparing the results of each monitoring group to the RMRS data (intensive measurements that are used as a standard for accuracy in this study, as discussed above). (For the second approach, we considered correlations to be high when $r^2 > 0.75$, moderate when $0.5 < r^2 < 0.75$, and low when $r^2 < 0.5$.) We also calculated Cook's distance for each regression to determine if any stream had a significant effect on the relationship. As a rule of thumb, an observation has a heavy influence on a relationship when Cook's distance exceeds 1 (Dunn and Clark 1987).

To further compare results of the monitoring groups with each other, we evaluated whether mean estimates of a given attribute in a given channel type (PB, PR, and SP) were related. We used channel types for this comparison to minimize the influence of individual crew observations at a single stream (i.e., by using group means within streams). Replicates within channel types permitted estimation of both main effects (channel type and monitoring group) and interactions (see below).

Furthermore, channel attributes are expected to differ among channel types (e.g., Rosgen 1996; Buffington et al. 2003), and these differences should be detectable by each of the groups as part of their status-and-trend monitoring. Moreover, there is the potential for protocols to be biased by channel type; because of how a given protocol is defined or implemented, it may systematically over- or underestimate a given attribute. To examine these issues, we tested for a significant interaction ($P < 0.1$) using ANOVA, channel type and monitoring groups as the main effects and observers within a stream as a repeated measure (i.e., average of all crews for each stream). A significant interaction effect can be present if one group has a consistent measurement bias that varies with channel type. For example, if group A consistently measures bank-full width wider than group B in PB channel types but group A measures bank-full width narrower than group B in PR channel types, then this will result in an interaction effect. Alternatively, even if the same trend across channel types is observed for an attribute among all protocols (e.g., PB > PR > SP), a significant interaction can exist if the difference in mean among protocols changes among channel types. Examples of these types of interactions are presented in graphical form in Results.

Ideally, data could be shared among groups even if both main effects (channel type and monitoring group) were significantly different as long as there was no significant interaction. This result would suggest that the underlying attribute that the monitoring groups are measuring is different but correlated. When significant interactions were found, we graphed the resulting data to determine which monitoring groups exhibited patterns that differed from the others. If these graphs suggested such a result, we reran the analysis after excluding data from monitoring protocols that differed from the others to determine if the interaction was still significant.

Results

Consistency of Measurement within a Monitoring Group

Gradient and sinuosity were generally measured with high internal consistency (Table 3). Four of the six monitoring groups that measured gradient had RMSE values less than 0.5%, while the other two groups had values between 0.5% and approximately 1.0%. The four groups that measured gradient with RMSE less than 0.5% also had CV less than 20% (high consistency); the other two groups had CV less than 35% (moderate consistency for gradient). All four monitoring groups that measured sinuosity had RMSE less than 0.1 and CV less than 20%.

Consistency in measuring mean bank-full width and width-to-depth ratio was lower (Table 3). Three of the seven monitoring groups had values of RMSE less than 1 m and CV less than 20% for bank-full width. Two monitoring groups had CV values less than 20% for measurements of width-to-depth ratio, but none had RMSE less than 2. Consistency in measuring habitat composition was mixed; RPD was generally measured with high internal consistency (RMSE $\leq$ 5 cm and CV $<$ 20% for six out of seven monitoring groups), while measurements of percent pools and pools per kilometer were less consistent (none of the monitoring groups had a CV $<$ 20% or RMSE $<$ 5 for either of these attributes). The percent fines was generally measured with moderate to low consistency, while D_{50} was generally measured

TABLE 3.—Descriptive statistics for attribute data collected by individual monitoring groups for all channel types combined. Statistical abbreviations are as follows: RMSE = root mean square error; CV = coefficient of variation; and S:N = signal–noise ratio; NM = not measured. See Table 1 for other abbreviations and text for monitoring group acronyms.

			Monitoring group						
Attribute class	Attribute	Statistic	AREMP	CDFG	EMAP	NIFC	ODFW	PIBO	UC
Reach characteristics	Gradient (%)	Mean	3.35	3.41	3.60	NM	3.48	3.33	3.73
		RMSE	0.20	1.01	0.49	NM	0.76	0.24	0.39
		CV	5.9	29.5	13.7	NM	21.9	7.1	10.6
		S:N	188.2	4.9	28.7	NM	14.1	124.4	49.2
	Sinuosity	Mean	1.22	NM	1.19	NM	NM	1.25	1.22
		RMSE	0.04	NM	0.06	NM	NM	0.10	0.11
		CV	3.1	NM	5.1	NM	NM	8.3	8.8
		S:N	13.0	NM	5.5	NM	NM	1.0	2.4
Channel cross section	BFW (m)	Mean	7.40	6.10	5.27	5.90	6.16	4.57	4.01
		RMSE	1.63	1.48	1.89	0.83	2.58	0.33	0.57
		CV	22.0	24.3	35.9	14.0	41.8	7.3	14.2
		S:N	10.9	6.8	2.5	24.7	2.8	58.1	20.2
	W:D	Mean	15.45	19.80	14.26	19.65	18.09	18.63	27.28
		RMSE	2.94	5.68	4.30	3.93	3.10	3.97	7.89
		CV	19.0	28.7	30.1	20.0	17.1	21.3	28.9
		S:N	2.1	1.7	1.7	6.1	3.5	1.5	1.6
Habitat composition	Percent pools	Mean	37.83	8.37	10.28	23.59	20.99	29.44	21.52
		RMSE	8.26	6.22	8.30	5.53	7.12	12.91	11.01
		CV	21.8	74.2	80.7	23.4	33.9	43.8	51.2
		S:N	5.2	0.4	1.6	13.5	7.0	1.4	1.9
	Pools/km	Mean	60.57	14.13	86.36	32.47	26.43	61.42	42.22
		RMSE	25.82	10.92	19.42	19.75	9.25	27.04	18.89
		CV	42.6	77.3	22.5	60.8	35.0	44.0	44.7
		S:N	1.0	0.2	1.8	1.1	5.0	0.8	1.6
	RPD (cm)	Mean	18.24	33.55	14.46	32.60	32.90	20.98	21.60
		RMSE	3.06	18.42	2.48	4.54	5.52	2.67	3.50
		CV	16.8	54.9	17.4	13.9	16.8	12.7	16.2
		S:N	6.3	0.2	6.1	4.9	3.2	7.4	11.9
Channel roughness	D_{50} (mm)	Mean	39.25	NM	36.44	NM	NM	49.28	27.93
		RMSE	18.19	NM	24.29	NM	NM	13.80	14.20
		CV	46.3	NM	66.6	NM	NM	28.0	50.8
		S:N	2.4	NM	1.0	NM	NM	6.0	3.6
	$\log_e(D_{50})$	Mean	25.1	NM	22.89	NM	NM	36.8	14.34
		RMSE	1.66	NM	1.51	NM	NM	1.31	2.12
		CV	54.3	NM	43.0	NM	NM	27.2	87.5
		S:N	3.7	NM	6.9	NM	NM	9.4	2.3
	Percent fines	Mean	28.68	22.14	19.52	NM	20.45	18.83	29.90
		RMSE	7.96	14.21	6.45	NM	7.32	4.88	8.22
		CV	27.8	64.2	33.0	NM	35.8	25.9	27.5
		S:N	2.0	0.3	3.6	NM	3.5	7.1	1.6
	$\log_e$(LWD/100 m)	Mean	8.51	3.48	26.44	35.75	19.25	18.60	19.89
		RMSE	1.73	1.73	18.41	4.99	3.20	4.27	7.13
		CV	20.3	49.6	69.6	14.0	16.6	23.0	35.8
		S:N	9.9	1.7	0.9	44.1	32.9	13.6	4.9
	$\log_e$(LWD/100 m)	Mean	5.44	1.98	10.65	13.59	8.51	8.07	10.0
		RMSE	1.21	1.80	1.77	1.21	1.44	1.53	1.54
		CV	19.11	64.51	62.4	16.4	37.88	44.24	45.87
		S:N	53.3	4.4	10.8	87.1	24.5	19.4	13.6

with low consistency and estimates of LWD/100 m varied from high to low consistency (Table 3).

Although consistency of sediment metrics was generally low (D_{50}) to moderate (% fines), these metrics were likely sufficient to distinguish broad differences in D_{50} (i.e., differentiating sand-, gravel-, and cobble-bedded channels) and to distinguish critical thresholds for fine sediment (e.g., the 20% threshold for declining survival to emergence of salmonid embryos; Bjornn and Reiser 1991). The use of logarithmic transformations for D_{50} did not change results in terms of CV categories (low, moderate, high consistency). In contrast, logarithmic transformation of LWD/100 m had a negative effect on consistency estimates; two monitoring groups (ODFW and PIBO) went from moderate or high internal consistency (CV $< 35\%$) to low internal consistency (CV $> 35\%$).

Adequacy of a Monitoring Group's Protocol to Detect Environmental Heterogeneity

Three attributes (channel gradient, mean bank-full width, and $\log_e$[LWD/100 m]) had moderate (>2.5) or

TABLE 4.—Theoretical maximum correlation coefficients (r^2_{max}; Faustini and Kaufmann 2007) for attribute values measured by different pairs of monitoring groups for all channel types combined. The r^2_{max} = values are presented for the highest and lowest S:N results for each attribute. See text for a list of monitoring group and attribute acronyms.

Attribute	Rank	Group 1	Group 2	$S:N_1$	$S:N_2$	r^2_{max}
Gradient	Highest	AREMP	PIBO	188.2	124.4	0.987
	Lowest	ODFW	CDFG	14.1	4.9	0.776
Sinuosity	Highest	AREMP	EMAP	13.0	5.5	0.786
	Lowest	UC	PIBO	2.4	1.0	0.353
BFW	Highest	PIBO	NIFC	58.1	24.7	0.945
	Lowest	ODFW	EMAP	2.8	2.5	0.526
W:D	Highest	NIFC	ODFW	6.1	3.5	0.668
	Lowest	UC	PIBO	1.6	1.5	0.369
Percent pools	Highest	NIFC	ODFW	13.5	7.0	0.815
	Lowest	PIBO	CDFG	1.4	0.4	0.174
Pools/km	Highest	ODFW	EMAP	5.0	1.8	0.536
	Lowest	PIBO	CDFG	0.8	0.2	0.070
RPD	Highest	UC	PIBO	11.9	7.4	0.813
	Lowest	ODFW	CDFG	3.2	0.2	0.130
D_{50}	Highest	PIBO	UC	6.0	3.6	0.671
	Lowest	AREMP	EMAP	2.4	1.0	0.353
$\log_e(D_{50})$	Highest	PIBO	EMAP	9.4	6.9	0.789
	Lowest	AREMP	UC	3.7	2.3	0.548
Percent fines	Highest	PIBO	EMAP	7.1	3.6	0.686
	Lowest	UC	CDFG	1.6	0.3	0.124
LWD/100 m	Highest	NIFC	PIBO	44.1	32.9	0.949
	Lowest	CDFG	EMAP	1.7	0.9	0.276
$\log_e$(LWD/100 m)	Highest	NIFC	AREMP	87.1	53.3	0.970
	Lowest	EMAP	CDFG	10.8	4.4	0.745

high S:N ratios (>6.5) across all monitoring groups (Table 3). Two other attributes (RPD and $\log_e[D_{50}]$) had S:N ratios greater than 2.5 for the majority of the monitoring groups that measured them (Table 3). The remaining five (variables, sinuosity, width-to-depth ratio, percent pools, pools per kilometer, and percent fines) had low S:N ratios (<2.5) for at least 50% of the monitoring groups. Two of the seven monitoring groups (ODFW and NIFC) had S:N ratios greater than 2.5 for more than 80% of the channel attributes evaluated when transformed values of D_{50} and LWD/100 m were considered (Table 3). Two groups (AREMP and EMAP) had S:N ratios greater than 2.5 for 70% of the attributes. One monitoring group (PIBO) had S:N ratios greater than 2.5 for 60% of the measured attributes. Two groups (UC and CDFG) had S:N ratios greater than 2.5 for 50% or less of their measured attributes.

Relationships among Protocols for a Given Attribute: Data Crosswalks and Sharing

The potential to share data were highly dependent upon the attribute and the monitoring group. For example, measurements of channel gradient had very high S:N ratios (>10) for five of the six monitoring groups, suggesting that data for this attribute have a high potential of being shared (Tables 3, 4). In contrast, six of the seven groups that determined pools per km had an S:N ratio less than 2.5 (low likelihood of sharing), the remaining group having an S:N ratio of 5. The generally high S:N ratio (all but one group > 2.5) for measurements of gradient, bank-full width, RPD, $\log_e(D_{50})$ and $\log_e$(LWD/100 m) suggest that these attributes have the greatest potential for being shared among groups (Tables 3, 4). In contrast, values of width to depth, percent pools, and pools per kilometer may be difficult to share because of generally low S:N ratios (<2.5).

Protocols used by monitoring groups to measure channel gradient, sinuosity, pools per kilometer, and D_{50} had similar trends across channel types (no significant interactions; $P > 0.1$; see gradient and D_{50} for examples in Figure 3). Statistically significant interactions (indicating systematic differences among protocols) were found for bank-full width, width-to-depth ratio, percent pools, RPD, percent fines, and LWD. Because the characteristics of LWD and percent fines are defined differently by different monitoring groups (Appendix 1), it is not surprising to see a significant interaction for these attributes. Consequently, those data may be more difficult to share because the monitoring groups are not measuring the same underlying condition (i.e., different LWD size categories and different definitions of how and where fine sediment is measured; Appendix 1).

We found that significant interactions in several of the attributes (bank-full width, width-to-depth ratio, and percent pools) became nonsignificant when results

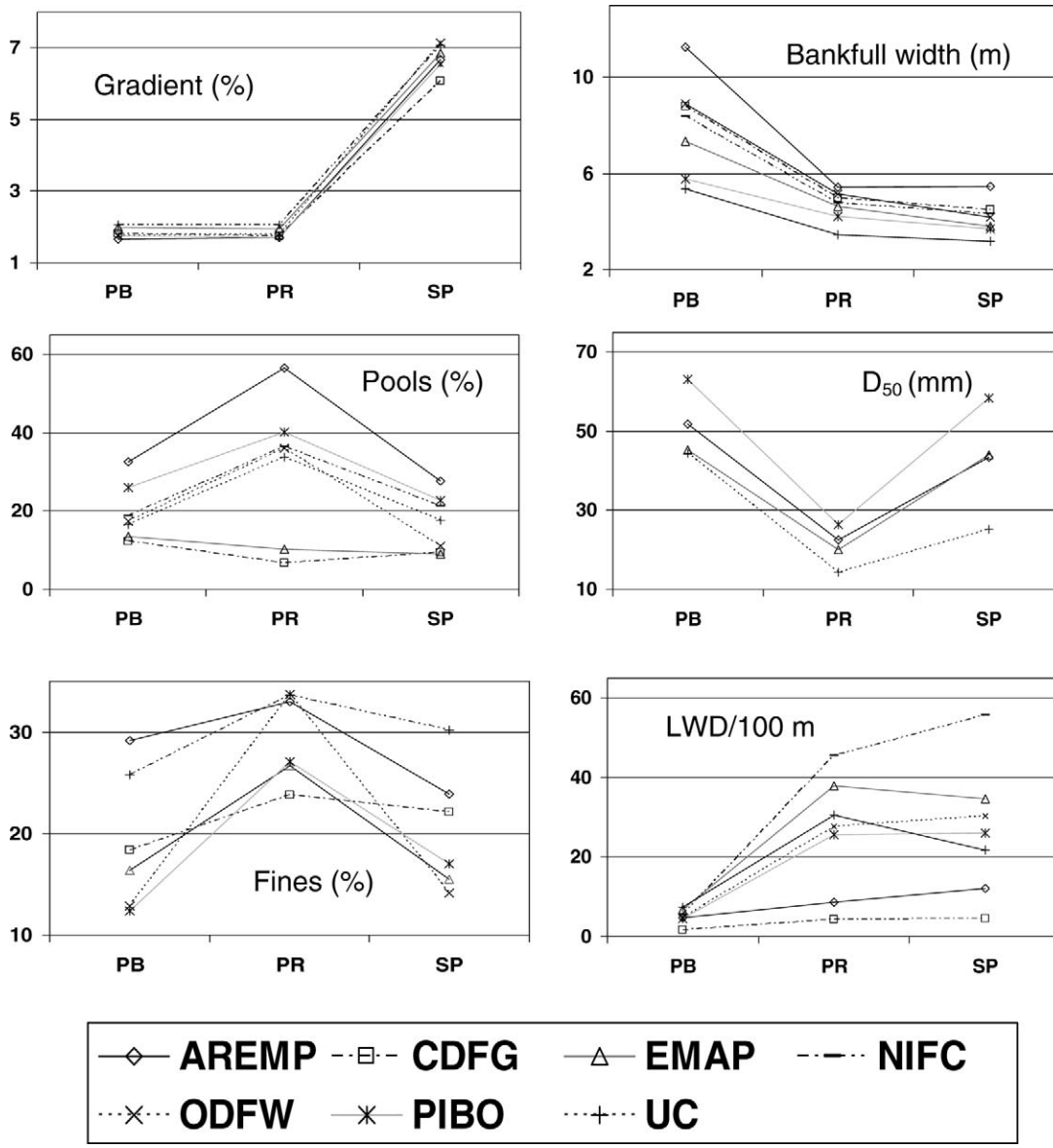

FIGURE 3.—Comparison of monitoring groups' results for six attributes averaged by channel type (plane bed [PB], pool–riffle [PR], and step pool [SP]). There are no interactions between the measurements for two attributes—channel gradient and median grain size (D_{50})—but significant interactions between those for the other attributes. See Methods for descriptions of the types of interactions and group acronyms.

from one or two monitoring groups were removed. Percent pools no longer had a significant interaction when CDFG and EMAP data were removed; this was because these two groups observed smaller percentages of pools in pool–riffle streams than in step-pool and plane-bed channels, while the remaining groups found that pool–riffle sites had a greater percentage of pools than step-pool and plane-bed streams (Figure 3). The width-to-depth ratio interaction was no longer significant when AREMP and CDFG data were removed; this was because these groups observed larger width-to-depth ratios in step-pool streams than in pool–riffle channels, while the remaining groups observed the reverse. The interaction in RPD was due to the EMAP and NIFC groups finding greater pool depths in plane-bed channels than in pool–riffle streams, while the

TABLE 5.—Coefficients of determination (r^2) between the values of the more intensively measured attributes and those obtained by individual monitoring groups for all channel types combined. See text for monitoring group and attribute acronyms.

Attribute	Monitoring group						
	AREMP	CDFG	EMAP	NIFC	ODFW	PIBO	UC
Gradient	0.99	0.98	0.99	NM	0.98	0.99	0.99
Sinuosity	0.93	NM	0.95	NM	NM	0.76	0.87
BFW	0.59	0.63	0.73	0.57	0.65	0.59	0.52
W:D	0.01	0.00	0.12	0.33	0.49	0.33	0.03
Percent pool	0.38	0.95	0.93	0.74	0.75	0.74	0.70
Pools/km	0.43	0.33	0.03	0.28	0.18	0.30	0.10
RPD	0.91	0.28	0.87	0.12	0.94	0.92	0.94
D_{50}	0.79	NM	0.87	NM	NM	0.92	0.73
$\log_e(D_{50})$	0.88	NM	0.87	NM	NM	0.93	0.86
Percent fines	0.40	0.07	0.72	NM	0.26	0.84	0.69
LWD/100 m	0.43	0.44	0.76	0.85	0.76	0.58	0.65
$\log_e$(LWD/100 m)	0.90	0.71	0.94	0.95	0.96	0.95	0.91

remaining groups found the opposite. Results for bank-full width differed from those above because the significant interaction arose from the magnitude of the differences among channel types rather than differences in their trends (Figure 3). Although there was no significant interaction for pools per kilometer, the high within-group variation in measuring this attribute results in low statistical power to detect differences among groups and poor potential for sharing this attribute (Tables 3, 4).

Comparison of measured attributes with results obtained by the RMRS team was similarly uneven (Table 5). We found that the monitoring groups' measurements of channel gradient, sinuosity, and D_{50} (both transformed and untransformed) were highly correlated with the RMRS measurements of those attributes ($r^2 > 0.75$). Measurements of bank-full width were moderately correlated with the RMRS values ($0.50 < r^2 < 0.75$), but this correlation would have been higher ($r^2 > 0.80$) if not for consistent differences between the monitoring groups and the RMRS measurements at two sites (Figure 4; Bridge and Crane creeks). Correlations between width-to-depth values measured by the monitoring groups and the RMRS team were uniformly low ($r^2 < 0.50$). There was a generally high correlation between monitoring group and RMRS measurements of percent pools for six of the seven monitoring groups. However, the relationship was dominated by Crane Creek, which had the largest abundance of pools (Cook's distance = 1–55 across the monitoring groups). Consequently, additional data collection over a broader range of conditions is warranted to further test the observed relationships. In contrast, the between-group correlation of pools per kilometer is uniformly poor.

We found that the strength of between-group relationships for measured wood loading (LWD/100 m) was partially dependent on whether the data were transformed or not. Five of seven groups had r^2 greater than 0.5 for untransformed values, but transformation resulted in all groups achieving r^2 greater than 0.5. Inspection of these relationships reveals that they are dominated by data from a single site. The relationships for untransformed LWD/100 m are driven by agreement across all groups that Bridge Creek has no LWD. When transformed, Bridge Creek became an outlier (Cook's distance > 8 for all monitoring groups) because its wood count was many units less than the other sites (e.g., $\log_e[0 + 0.1] = -2.3$ for 0 pieces of wood versus $\log_e[1 + 0.1] = 0.095$ for 1 piece/100 m). In contrast, the regression for the untransformed data had no heavily weighted observation (Cook's distance < 1 for all but one monitoring group). Overall, there was a moderate correlation ($r^2 > 0.50$) between monitoring groups and RMRS values for at least 50% of the channel attributes except for the CDFG group (Table 5).

Summary

In almost every case, there were one or more groups that either measured an attribute with less consistency or seemed to be measuring a slightly different underlying environmental condition than the other groups (Table 5). However, the study results indicate that measurements were relatively consistent within each monitoring group (moderate to high consistency) and that there is a moderate to high likelihood for sharing data if monitoring groups alter their protocols for attributes that had significant interactions. Overall, four of the seven monitoring groups measured at least 80% of the evaluated attributes with moderate or high consistency (CV < 35%, high-to-moderate RMSE, or both; AREMP, 80% of the evaluated attributes; NIFC, 83.3%; ODFW, 88%; and PIBO, 80% [note that these summaries include the best of the raw or log-transformed values]). Two of the remaining monitoring

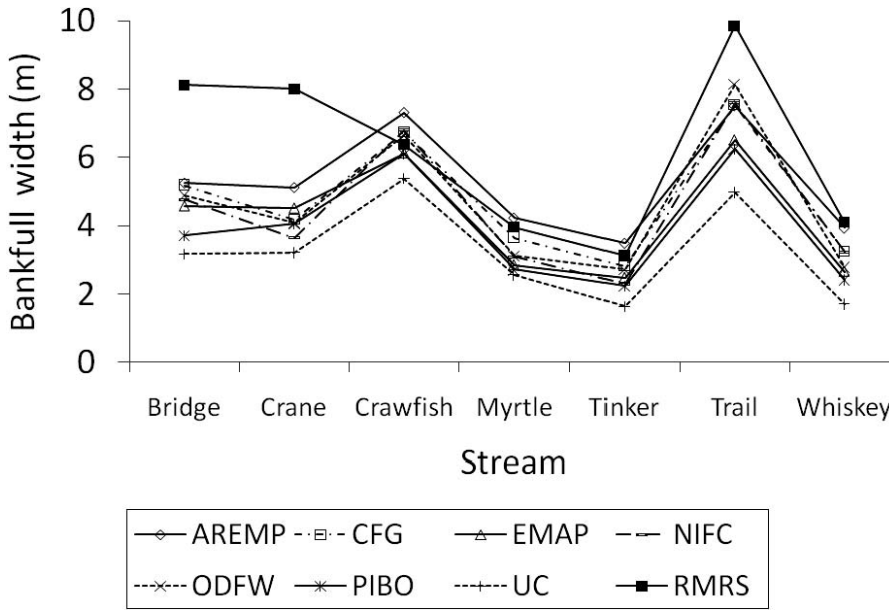

FIGURE 4.—Comparison of mean bank-full width as determined by the seven monitoring groups and via the more intensive data collection (RMRS) at 7 of the 12 study sites.

groups (EMAP, UC) measured the majority of the attributes with moderate or high consistency, while CDFG measured 50% of the attributes with moderate or high consistency (Table 6).

Two attributes, channel gradient and sinuosity, were consistently measured within monitoring groups (low-to-moderate CV and RMSE) and produced values that were correlated both among monitoring groups and with the RMRS data (high r^2; Tables 4, 5). Gradient also had generally high environmental heterogeneity (S:N $>$ 6.5), but sinuosity varied over a relatively narrow range, leading to low S:N values ($<$2.5) for two of the four protocols that measured it. Data on a third attribute (RPD) also was measured with high consistency, had high S:N ratios, and could likely be shared among a majority of the groups. Evaluations of LWD/

TABLE 6.—Overall assessment of the performance of monitoring groups, scored as high (H), moderate (M), or low (L); NM = not measured. Three performance characteristics are scored for each monitoring group: internal consistency (CV, RMSE, or both), environmental heterogeneity (S:N), and likelihood for sharing data (correlations with intensive data); H, M, and L values are defined in the text for each of these parameters. See text for a list of monitoring protocol and attribute acronyms.

	Monitoring group						
Attribute	AREMP	CDFG	EMAP	NIFC	ODFW	PIBO	UC
Gradient	H/H/H	M/M/H	H/H/H	NM	M/H/H	H/H/H	H/H/H
Sinuosity	H/H/H	NM	H/M/H	NM	NM	H/L/H	H/L/H
BWW	M/H/M	M/H/M	L/M/M	H/H/M	L/M/M	H/H/M	H/H/M
W:D	H/L/L	M/L/L	M/L/L	H/M/L	H/M/L	M/L/L	M/L/L
Percent pool	M/M/L	M/L/H	M/L/H	M/H/M	M/H/H	L/L/M	L/L/M
Pools/km	L/L/L	L/L/L	M/L/L	L/L/L	M/M/L	L/L/L	L/L/L
RPD	H/M/H	L/L/L	H/M/H	H/M/L	H/M/H	H/H/H	H/H/H
D_{50}	L/L/H	NM	L/L/H	NM	NM	M/M/H	L/M/H
$\log_e(D_{50})$	L/M/H	NM	L/H/H	NM	NM	M/H/H	L/L/H
Percent fines	M/L/L	L/L/L	M/M/M	NM	M/M/L	H/H/H	M/L/M
LWD/100 m	M/H/L	L/L/L	L/L/H	H/H/H	H/H/H	M/H/M	L/L/M
$\log_e$(LWD/100 m)	H/H/H	L/M/M	L/H/H	H/H/H	L/H/H	L/H/H	L/H/H
Percent H or M[a]	80/70/60	44/38/50	70/70/80	83/83/50	88/100/63	80/60/80	60/50/80
Number of attributes measured	10	8	10	6	8	10	10

[a] Percentage of scores that were H or M. If the scores for $\log_e$ transformed and nontransformed variables differed, the higher value (e.g., H or M rather than L) was used.

100 m had moderate-to-high internal consistency but would be difficult to share because of differences in sampling protocols (Appendix 1) and resultant differences in measured values (Figure 3).

Bank-full width, D_{50} (both transformed and untransformed), percent pools, and percent fines were generally measured with less consistency within each group and typically had smaller S:N ratios than the previously listed attributes, but had values that were moderately or highly correlated with each other and to the RMRS data. Width-to-depth ratios and pools per kilometer were generally inconsistently measured, had low environmental heterogeneity, and were weakly correlated to the RMRS data.

Overall, we found that five monitoring groups (AREMP, EMAP, ODFW, PIBO, and UC) measured two-thirds or more of the attributes with high internal consistency, had moderate S:N ratios, and produced results that were at least moderately related to those of the other groups and the RMRS data. One group (NIFC) measured attributes with high internal consistency but had fewer attributes related to the results of the other groups or to the RMRS data and collected fewer of the commonly evaluated stream attributes. California Department of Fish and Game measured attributes with lower average internal consistency and were not as strongly correlated to the results of the other monitoring groups.

Discussion

Our comparison of seven stream habitat monitoring groups from the Pacific Northwest suggests considerable variability in each group's ability to consistently and accurately measure some stream attributes. Reasons for differences in observer measurements have been studied elsewhere and include differences in the duration of training (Hannaford et al. 1997; Whitacre et al. 2007; Heitke et al. 2008), level of experience (Wohl et al. 1996; Heitke et al. 2008), operational definitions for the attribute of interest (Kaufmann et al. 1999; Heitke et al. 2008; Roper et al. 2008), intensity of measurements (Wolman 1954; Wang et al. 1996; Robison 1997), when and where the attribute is measured (Olsen et al. 2005; Roper et al. 2008), and characteristics of the sampled stream reach (Whitacre et al. 2007).

Regardless of the exact reason for poor performance of a monitoring group (e.g., low repeatability, poor accuracy, etc.), we argue improvement can only occur if groups and protocols are regularly evaluated and training and oversight are thorough and ongoing. Assessment of groups and protocols may identify weaknesses that can then be remedied. For example, in an earlier study by Whitacre et al. (2007), it was found that PIBO had a higher internal consistency in measuring stream gradient than AREMP. This difference occurred even though AREMP used an instrument with greater precision (total station versus hand level). Whitacre et al. (2007) speculated that the presence of dense riparian vegetation caused the total station to be moved multiple times, resulting in greater cumulative survey errors by AREMP. Following this study, AREMP altered its protocols for measuring stream gradient and increased training. Results of the current comparison indicate that both PIBO and AREMP now measure gradient with high consistency. Comparison of results among monitoring groups, such as the one above, not only provides feedback on what is possible but can also provide incentive for improving a monitoring group's protocols.

Two concerns identified prior to implementing our study were how declining stream flows during the summer sampling period and different numbers of crews would affect study results. We found no strong evidence supporting either of these concerns. Changes in flow have the potential to cause variability within and between monitoring groups whose protocols rely on wetted channel dimensions. With the exception of RMRS, all of the groups examined in this study use wetted dimensions to some degree in their measurements (Appendix 1). Stream gradient, percent pools, and pools per kilometer depend on wetted dimensions for all groups, as does percent fines (except in the UC protocol). Three additional parameters (reach length, sinuosity, and D_{50}) depend on wetted dimensions for the EMAP protocol. To document changes in flow during the study period, staff gauges were installed at each site. The maximum change in reach-average width and depth was 17% and 33%, respectively, on average across the sites during the study period (excluding Tinker Creek, which went dry). The corresponding changes in stream gradient were small (0.8% on average across sites), but the effects of changing flow on the above-listed parameters are not easily quantified. Examination of the within- and between-group variability over the study period may offer some insight into the potential effects of declining flows. Both PIBO and NIFC maintained high consistency within their groups even though observers sampled over an entire month during which stream flows declined. In contrast, CDFG crews had lower repeatability even though they sampled over a 2-week period. One reason why AREMP may have been able to determine pool attributes more consistently than PIBO was because base flows were more stable over the shorter time frame they sampled (7 versus 33 d). However, this would not explain why ODFW was able to consistently determine pool attributes despite sampling over a 37-d

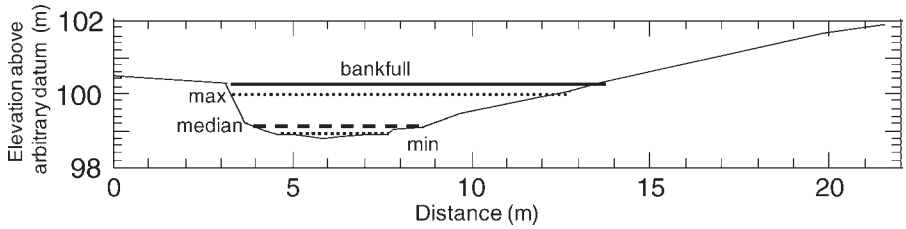

FIGURE 5.—One of five cross sections at Bridge Creek that was measured by all monitoring groups and surveyed by the RMRS team. The line labeled "bank-full" was determined from the RMRS data and field indicators. The maximum, minimum, and median bank-full values were determined from the values reported by the monitoring groups. Most of the groups tended to underestimate the bank-full width at this cross section (the median value is closer to the minimum) because vegetation obscured the bank-full morphology.

time frame. Overall, these observations suggest that declining flows did not have a strong influence on internal consistency, but further examination of this issue is warranted. Sampling over a shorter period reduces within-group variability due to changing flows but does not account for between-group differences that may result from groups sampling different flows at different times of the study period.

The number of crews used (two to six) also seemed to have little relationship with consistency. The monitoring groups that sent more crews (AREMP and PIBO) may have been slightly more consistent, but ODFW only had two crews and were also consistent (Table 3). A possible reason for greater consistency with a larger number of crews is that anomalous results may be damped. For example, if a single crew provides a slightly different evaluation of a stream attribute, their results will have less impact if it is one of six crews rather than one of three crews. Certainly, the changes in base flow and the number of observers had some impact on the comparisons made in this study, but they do not appear to be significant.

Our results also anecdotally suggest that additional training may be required when crews are unfamiliar with the channel types they are sampling. All monitoring programs other than the CDFG had previous experience sampling streams similar to those present in the John Day River basin. The lack of experience in monitoring these systems may partially explain the lack of consistency within this program.

Comparing monitoring group results with more intensive site measurements (e.g., RMRS data) provided additional insight into the types of common errors associated with rapid monitoring protocols, such as those evaluated in this study. For example, in all but two streams there was a high correlation between bank-full width measurements made by the monitoring groups and those determined from the RMRS data (Figure 4; Bridge and Crane creeks). This difference can be explained by examining cross-section transects common to all of the field crews (five transects were established in each stream reach where all monitoring groups, including the RMRS team, measured the exact same cross section).

Bank-full width is clearly visible in the cross section surveyed by the RMRS team at Bridge Creek (Figure 5), but monitoring groups consistently underestimated bank-full width at this location. Based on photographs taken at the site and subsequent field visits at different times of year, it became apparent that the observed differences in bank-full width measurements were due to dense seasonal riparian vegetation that obscured the bank morphology during the summer sampling period, causing field crews to underestimate channel dimensions. Comparisons of this sort highlight the value of periodically evaluating protocols and having quality assurance–quality control (QA–QC) plans. This was accomplished in this study by obtaining intensive measurements, establishing common sampling locations for calibrating and comparing the results of monitoring groups, interpreting discrepancies, and identifying areas for additional training (e.g., identifying bank-full in brushy streams). Overall, the correlations between many of the attributes measured by the monitoring groups and the RMRS data were high (Table 5). This finding is encouraging considering that the monitoring group evaluations were conducted over different stream lengths (same starting point, but different ending locations), reach length potentially having a significant impact when comparing results among and within monitoring groups (Whitacre et al. 2007).

We found that transforming data could affect estimates of internal consistency, S:N ratios, and relationships with more intensive measurements. Comparisons between transformed and untransformed data for both D_{50} and LWD/100 m suggest that the effects of transformation on CV and S:N ratios were mixed (Table 3), the exact effect dependent on the distribution of data across streams and among observers. However,

transformations consistently had a positive effect on the correlation between measurements made by monitoring groups and the RMRS data collection. For D_{50}, this improvement was due, in part, to the fact that the underlying data distribution was lognormal, even though the observer error distribution was not. In contrast, the improved correlation with transformation of the LWD/100 m was due to the influence of a single site (Bridge Creek) and its lack of LWD.

The use of logarithmic transformation often has biological or physical basis and has the advantages of offering a convenient approach for interpreting multiplicative effects and for producing confidence intervals that are always positive (Limpert et al. 2001). Use of this transformation can therefore aid our understanding of the system. This is not true of all transformations (such as ranks used for nonparametric tests or arcsine square root transformations), which may address statistical concerns but do not improve data interpretation (Johnson 1995). The possibility that different monitoring groups might have different error structures (e.g., if one group bins pebble size with the phi scale [Krumbein 1936] while another measures to the nearest millimeter) complicates the ability to join data because one must account for differences in both means and the error structure.

From a Blue Mountains ecoregion perspective, the relationship among monitoring groups and the RMRS data suggests the possibility of being able to combine (share) data from attributes with high internal consistency (CV $<$ 20%, relatively low RMSE values, or both), high S:N ratios ($>$6.5), and high correlation ($r^2 > 0.75$) with the intensive measurements (RMRS). In order to combine these data, it first would be necessary to clearly define the target population of interest (e.g., fish-bearing streams) and provide consistent representation of the distribution of that target population across the landscape (e.g., a digital map of the stream network and potential habitat; Buffington et al. 2004) so that weights could be determined for each reach sampled by a monitoring group (Larsen et al. 2007).

The simplest way to analyze the combined data would be to treat monitoring groups as a block effect in a larger ANOVA design. Because of the difference in overall means among monitoring groups, this approach may not be helpful in assessing status, but it would help to evaluate trend since correlated protocols should show the same change over time. The second approach would be to combine data using the rank order of the reach (percentile) within a sample of stream reaches. This approach would be bolstered if specific stream reaches were measured by all monitoring groups so as to serve as comparison–correlation sites. Such an approach might permit the construction of cumulative frequency histograms using conditional probability analysis (Paul and McDonald 2005).

Both of these approaches will be hampered where interactions occur among values measured by different monitoring groups. A significant interaction will influence how a stream attribute is perceived to change through time. This can affect the rank order of stream attributes among monitoring groups (i.e., how consistently a monitoring group measures an attribute compared with other groups) and will prevent data crosswalks. Therefore, combining data among monitoring groups, though conceptually straightforward, should be done with care and, in some cases, may not be feasible.

Conclusions and Recommendations

For each of the 10 attributes we evaluated, at least one monitoring group was able to simultaneously achieve at least moderate internal consistency (CV$<$ 35% and low-to-moderate RMSE) and moderate detection of environmental heterogeneity (S:N $>$ 2.5). However, none of the monitoring groups were able to achieve this standard for all their measured parameters, suggesting that there is room for improvement in all the monitoring groups evaluated in this study, both for internal program success and the possibility of sharing data and scaling up results to regional and national levels.

This study was conducted on a limited number of streams, over a limited area, and over a protracted sampling period that could have influenced the results due to changes in streamflow. We also recognize that the criteria used here for evaluating protocol performance and compatibility will not necessarily fit every situation or management objective. However, some sort of criteria are needed for evaluating data collected by monitoring groups; poorly measured attributes add little to our understanding of streams and provide fodder for articles questioning the need for, or validity of, large-scale monitoring groups (GAO 2000, 2004; Stem et al. 2005; Nichols and Williams 2007). To ensure that data are efficiently collected, we suggest that agencies and organizations either adopt standards such as those used in this analysis or develop other meaningful criteria for aquatic habitat data, and implement an annual QA–QC program. There is always a trade-off between rigor at a site and the benefits gained from visiting more sites. However, greater measurement precision and consistency increase the likelihood that data can be combined across groups, in turn increasing the number of monitoring sites available for combined analysis.

We recognize that the loss of legacy data are a primary concern for monitoring groups when consid-

ering new or modified protocols (Bonar and Hubert 2002). The power of using protocols that capture legacy data are well demonstrated by McIntosh et al. (1994; historic changes in pool size and frequency in the Columbia basin) and Rodgers et al. (2005; use of basin surveys to augment random surveys in an assessment of carrying capacity for juvenile coho salmon *Oncorhynchus kisutch* in Oregon coastal drainages). However, if legacy data cannot be integrated into regional assessments or do not allow trend detection because of low measurement repeatability, there is little justification for spending limited resources to continue acquiring these data.

Our results, in combination with an earlier study (Whitacre et al. 2007), suggest that major improvement in monitoring group precision and consistency can occur without changing protocols. To do this it is first necessary to identify inconsistent protocols. Once a procedure within a protocol has been identified as being inconsistent (e.g., by evaluating how different observers make observations at a variety of selected locations; Figure 5), the protocol can be altered by either clarifying the operational definition or by providing additional training focused on aspects of the protocol that have been applied inconsistently.

The compelling reason to improve current protocols is the growing need to determine status and trends of stream habitat at a regional and national scale while being fiscally responsible (GAO 2000, 2004; EPA 2006a). There has been progress on (1) sampling designs that foster regional and national estimates of aquatic conditions (Urquhart et al. 1998; Larsen et al. 2001), (2) methods for combining data based on different sampling designs (Larsen et al. 2007), and (3) understanding which stream habitat attributes should be measured (MacDonald et al. 1991; Bauer and Ralph 2001; Montgomery and MacDonald 2002), but there has been little progress toward ensuring that monitoring groups incorporate QA–QC as part of their monitoring procedures. From an accountability standpoint, we think it is incumbent on managers of stream monitoring groups to be able to demonstrate that long-term stream habitat monitoring efforts provide a cost-effective assessment of habitat trends (Lovett et al. 2007). This goal will be best achieved by continually having monitoring groups assess data quality and by increasing coordination among monitoring organizations to improve data quality, reduce redundancy, and promote data sharing.

Acknowledgments

We thank the Independent Scientific Advisory Board (Northwest Power and Conservation Council) for study plan comments, and we thank two anonymous reviewers and the associate editor for constructive comments that improved the manuscript. The University of Idaho provided field equipment that was ably used by Patrick Kormos, Darek Elverud, Russ Nelson, Brian Ragan, and Kathy Seyedbagheri. This project was a collaborative effort of the Pacific Northwest Aquatic Monitoring Partnership and was funded by US Forest Service, Environmental Protection Agency, Bureau of Land Management, NOAA Fisheries, U.S. Fish and Wildlife Service, Bonneville Power Administration, the states of Washington, Oregon, and California, and the Northwest Indian Fisheries Commission.

References

AREMP (Aquatic and Riparian Effectiveness Monitoring Program). 2005. Field protocol manual. U.S. Forest Service, and Bureau of Land Management, Corvallis, Oregon. Available: reo.gov/. (August 2009).

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

Bernhardt, E. S., M. A. Palmer, J. D. Allan, G. Alexander, K. Barnas, S. Brooks, J. Carr, S. Clayton, C. Dahm, J. Follstad-Shah, D. Galat, S. Gloss, P. Goodwin, D. Hart, B. Hassett, R. Jenkinson, S. Katz, G. M. Kondolf, P. S. Lake, R. Lave, J. L. Meyer, T. K. O'Donnell, L. Pagano, B. Powell, and E. Sudduth. 2005. Synthesizing U.S. river restoration efforts. Science 308:636–637.

Bilby, R. E., W. J. Ehinger, C. Jordan, K. Krueger, M. McHenry, T. Quinn, G. Pess, D. Poon, D. Seiler, and G. Volkhardt. 2004. Evaluating watershed response to land management and restoration actions: intensively monitored watersheds (IMW) progress report. Prepared for the Washington Salmon Recovery Funding Board, Olympia.

Bisson, P. A., J. L. Nielsen, R. A. Palmason, and L. E. Grove. 1982. A system of naming habitat types in small streams, with examples of habitat utilization by salmonids during low streamflow. Pages 62–73 *in* N. B. Armantrout, editor. Proceedings of a symposium on acquisition and utilization of aquatic habitat inventory information. American Fisheries Society, Western Division, Portland, Oregon.

Bjornn, T., and D. Reiser. 1991. Habitat requirements of salmonids in streams. Pages 83–138 *in* W. R. Meehan, editor. Influences of forest and rangeland management on salmonid fishes and their habitats. American Fisheries Society, Symposium 19, Bethesda, Maryland.

Bonar, S. A., and W. A. Hubert. 2002. Standard sampling of inland fish: benefits, challenges, and a call for action. Fisheries 27(3):10–16.

Budy, P., and H. Schaller. 2007. Evaluating tributary restoration potential for Pacific salmon recovery. Ecological Applications 17:1068–1086.

Buffington, J. M., T. E. Lisle, R. D. Woodsmith, and S. Hilton. 2002. Controls on the size and occurrence of pools in coarse-grained forest rivers. River Research and Applications 18(6):507–531.

Buffington, J. M., and D. R. Montgomery. 1999. Effects of

hydraulic roughness on surface textures of gravel-bed rivers. Water Resources Research 35:3507–3522.

Buffington, J. M., D. R. Montgomery, and H. M. Greenberg. 2004. Basin-scale availability of salmonid spawning gravel as influenced by channel type and hydraulic roughness in mountain catchments. Canadian Journal of Fisheries and Aquatic Sciences 61:2085–2096.

Buffington, J. M., R. D. Woodsmith, D. B. Booth, and D. R. Montgomery. 2003. Fluvial processes in Puget Sound rivers and the Pacific Northwest. Pages 46–78 *in* D. R. Montgomery, S. M. Bolton, and D. B. Booth, editors. Restoration of Puget Sound rivers. University of Washington Press, Seattle.

Bunte, K., and S. R. Abt. 2001. Sampling surface and subsurface particle size distributions in wadable gravel- and cobble-bed streams for analyses in sediment transport, hydraulics, and streambed monitoring. U.S. Forest Service General Technical Report RMRS-GTR-74.

Clarke, S. E., and S. A. Bryce, editors. 1997. Hierarchical subdivisions of the Columbia Plateau and Blue Mountains ecoregions, Oregon and Washington. U.S. Forest Service General Technical Report PNW-GTR-395.

Crawford, B. A., and S. Rumsey. 2009. Draft guidance for monitoring recovery of Pacific Northwest salmon and steelhead listed under the federal Endangered Species Act (Idaho, Oregon, and Washington). National Marine Fisheries Service, Seattle. Available: nwr.noaa.gov/. (August 2009).

Downie, S. T. 2004. Stream channel measurement methods for core attributes, 2004. California Department of Fish and Game Fortuna. Available: coastalwatersheds.ca.gov. (August 2009).

Dunn, O. J., and V. A. Clark. 1987. Applied statistics: analysis of variance and regression. Wiley, New York.

EPA (U.S. Environmental Protection Agency). 2006a. Wadeable stream assessment: a collaborative survey of the nation's streams. EPA, EPA 841-B-06-002, Washington, D.C.

EPA (U.S. Environmental Protection Agency). 2006b. Environmental monitoring and assessment program: surface waters western pilot study—field operations manual for wadeable streams. EPA, EPA/620/R-06/003, Washington, D.C.

Fausch, K. D., C. L. Hawkes, and M. G. Parsons. 1988. Models that predict standing crop of stream fish from habitat variables: 1950–1985. U.S. Forest Service General Technical Report PNW-GTR-213.

Faustini, J. M., and P. R. Kaufmann. 2007. Adequacy of visually classified particle count statistics from regional stream habitat surveys. Journal of the American Water Resources Association 43:1293–131.

Faux, R., J. M. Buffington, G. Whitley, S. Lanigan, and B. Roper. 2009. Use of airborne near-infrared LiDAR for determining channel cross-section characteristics and monitoring aquatic habitat in Pacific Northwest rivers: a preliminary analysis. Pages 43–60 *in* J. M. Bayer and J. L. Schei, editors. Proceedings of the American Society for Photogrammetry and Remote Sensing. Pacific Northwest Aquatic Monitoring Partnership, Cook, Washington.

Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. Environmental Management 10:199–214.

GAO (U.S. Government Accountability Office). 2000. Water quality: key EPA and state decisions limited by inconsistent and incomplete data. Report to the Chairman, Subcommittee on Water Resources and Environment, Committee on Transportation and Infrastructure, U.S. House of Representatives. GAO, GAO/RCED-00-54, Washington, D.C.

GAO (U.S. Government Accountability Office). 2004. Watershed management: better coordination of data collection efforts needed to support key decisions. Report to the Chairman, Subcommittee on Water Resources and Environment, Committee on Transportation and Infrastructure, U.S. House of Representatives. GAO, GAO-04-382, Washington, D.C.

Hannaford, M. J., M. T. Barbour, and V. H. Resh. 1997. Training reduces observer variability in visual-based assessments of stream habitat. Journal of the North American Benthological Society 16:853–860.

Harrelson, C. C., C. L. Rawlins, and J. P. Potyondy. 1994. Stream channel reference sites: an illustrated guide to field technique. U.S. Forest Service General Technical Report GTR-RM-245.

Hartman, G. F., J. C. Scrivener, and M. J. Miles. 1996. Impacts of logging in Carnation Creek, a high-energy coastal stream in British Columbia, and their implication for restoring fish habitat. Canadian Journal of Fisheries and Aquatic Sciences 53(Supplement 1):237–251.

Heitke, J. D., R. C. Henderson, B. B. Roper, and E. K. Archer. 2008. Evaluating livestock grazing use with streambank alteration protocols; challenges and solutions. Rangeland Management and Ecology 61:647–655.

Hey, R. D., and C. R. Thorne. 1983. Accuracy of surface samples from gravel bed material. Journal of Hydraulic Engineering 109:842–851.

Honea, J. M., J. C. Jorgensen, M. M. McClure, T. D. Cooney, K. Engie, D. M. Holzer, and R. Hilborn. 2009. Evaluating habitat effects on population status: influence of habitat restoration on spring-run Chinook salmon. Freshwater Biology 54:1576–1592.

Isaak, D. J., and R. F. Thurow. 2006. Network-scale spatial and temporal variation in Chinook salmon (*Oncorhynchus tshawytscha*) redd distributions: patterns inferred from spatially continuous replicate surveys. Canadian Journal of Fisheries and Aquatic Sciences 63:285–296.

Johnson, D. H. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

Johnson, D. H., N. Pittman, E. Wilder, J. A. Silver, R. W. Plotnikoff, B. C. Mason, K. K. Jones, P. Roger, T. A. O'Neil, and C. Barrett. 2001. Inventory and monitoring of salmon habitat in the Pacific Northwest: directory and synthesis of protocols for management/research and volunteers in Washington, Oregon, Idaho, Montana, and British Columbia. Washington Department of Fish and Wildlife, Olympia.

Kaufmann, P. R., J. M. Faustini, D. P. Larsen, and M. A. Shirazi. 2008. A roughness-corrected index of relative bed stability for regional stream surveys. Geomorphology 99:150–170.

Kaufmann, P. R., E. G. Levine, E. G. Robinson, C. Seeliger,

and D. V. Peck. 1999. Quantifying physical habitat in wadeable streams. U.S. Environmental Protection Agency, EPA/620/R-99/003, Washington, D.C.

Kellerhals, R., and D. I. Bray. 1971. Sampling procedures for coarse fluvial sediments. Journal of the Hydraulics Division, American Society of Civil Engineers 97:1165–1180.

Kershner, J. L., B. B. Roper, N. Bouwes, R. Henderson, and E. Archer. 2004. An analysis of stream reach habitat conditions in reference and managed watersheds on some federal lands within the Columbia River basin. North American Journal of Fisheries Management 24:1363–1375.

Krumbein, W. C. 1936. Application of logarithmic moments to size frequency distributions of sediments. Journal of Sedimentary Petrology 6(1):35–47.

Larsen, D. P., T. M. Kincaid, S. E. Jacobs, and N. S. Urquhart. 2001. Designs for evaluating local and regional scale trends. BioScience 51:1069–1078.

Larsen, D. P., A. R. Olsen, S. H. Lanigan, C. Moyer, K. K. Jones, and T. M. Kincaid. 2007. Sound survey designs can facilitate integrating stream monitoring data across multiple programs. Journal of the American Water Resources Association 43:384–397.

Limpert, E., W. A. Stahel, and M. Abbt. 2001. Log-normal distributions across the sciences: keys and clues. BioScience 51:341–352.

Lisle, T. E., and S. Hilton. 1992. The volume of fine sediment in pools: an index of sediment supply in gravel-bed streams. Water Resources Bulletin 28(2):371–383.

Lovett, G. M., D. A. Burns, C. T. Driscoll, J. C. Jenkins, M. J. Mitchell, L. Rustad, J. B. Shanley, G. E. Likens, and R. Haeuber. 2007. Who needs environmental monitoring? Frontiers in Ecology and the Environment 5:253–260.

MacDonald, L. H., A. W. Smart, and R. C. Wissmar. 1991. Monitoring guidelines to evaluate effects of forestry activities on streams in the Pacific Northwest and Alaska. U.S. Environmental Protection Agency, EPA-910/9-001, Seattle.

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

McIntosh, B. A., J. R. Sedell, J. E. Smith, R. C. Wissmar, S. E. Clarke, G. H. Reeves, and L. A. Brown. 1994. Historical changes in fish habitat for select river basins of eastern Oregon and Washington. Northwest Science 68:36–53.

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

Montgomery, D. R., J. M. Buffington, R. D. Smith, K. M. Schmidt, and G. Pess. 1995. Pool spacing in forest channels. Water Resources Research 31:1097–1105.

Montgomery, D. R., and L. H. MacDonald. 2002. Diagnostic approach to stream channel assessment and monitoring. Journal of the American Water Resources Association 38:1–16.

Moore, K. M., K. K. Jones, and J. M. Dambacher. 1997. Methods for stream habitat surveys: aquatic inventories project. Oregon Department of Fish and Wildlife, Information Report 97-4, Corvallis. Available: nrimp.dfw.state.or.us/. (August 2009).

Nichols, J. D., and B. K. Williams. 2007. Monitoring for conservation. Trends in Ecology and Evolution 21:668–673.

Olsen, D. S., B. B. Roper, J. L. Kershner, R. Henderson, and E. Archer. 2005. Sources of variability in conducting pebble counts: their potential influence on the results of stream monitoring programs. Journal of the American Water Resources Association 41:1225–1236.

Paul, J. F., and M. E. McDonald. 2005. Development of empirical, geographically specific water quality criteria: a conditional probability analysis approach. Journal of the American Water Resources Association 41:1211–1223.

Peck, D. V., A. T. Herlihy, B. H. Hill, R. M. Hughes, P. R. Kaufmann, D. J. Klemm, J. M. Lazorchak, F. H. McCormick, S. A. Peterson, P. L. Ringold, T. Magee, and M. Cappaert. 2006. Environmental monitoring and assessment program: surface waters western pilot study—field operations manual for wadeable streams. U.S. Environmental Protection Agency, EPA/620/R-06/003, Washington, D.C.

Pleus, A., and D. Schuett-Hames. 1998. TFW monitoring program method manual for the reference point survey. Northwest Indian Fisheries Commission, TFW-AM9-98-002, Olympia, Washington. Available: nwifc.org/tfw/downloads.asp. (August 2009).

Pleus, A., D. Schuett-Hames, and L. Bullchild. 1999. TFW monitoring program method manual for the habitat unit survey. Northwest Indian Fisheries Commission, TFW-AM9-99-003, Olympia, Washington. Available: nwifc.org/tfw/downloads.asp. (August 2009).

Ramsey, S. C., M. Thompson, and M. Hale. 1992. Objective evaluation of precision requirements for geochemistry analysis using robust analysis of variance. Journal of Geochemical Exploration 44:23–36.

Robison, E. G. 1997. Reach-scale sampling metrics and longitudinal pattern adjustments of small streams. Doctoral dissertation. Oregon State University, Corvallis.

Robison, E. G., and R. L. Beschta. 1990. Coarse woody debris and channel morphology interactions for undisturbed streams in Southeast Alaska, USA. Earth Surface Processes and Landforms 15:149–156.

Rodgers, J. D., K. K. Jones, A. G. Talabere, C. H. Stein, and E. H. Gilbert. 2005. Oregon coast coho habitat assessment, 1998–2003. Oregon Department of Fish and Wildlife, OPSW-ODFW-2005-5, Salem.

Roper, B., J. Kershner, E. Archer, R. Henderson, and N. Bouwes. 2002. An evaluation of physical stream habitat attributes used to monitor streams. Journal of the American Water Resources Association 38:1637–1646.

Roper, B. B., J. M. Buffington, E. Archer, C. Moyer, and M. Ward. 2008. The role of observer variation in determining Rosgen channel types in northwestern Oregon mountain streams. Journal of the American Water Resources Association 42:418–427.

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

Rosgen, D. L. 1996. Applied river morphology. Wildland Hydrology, Pagosa Springs, Colorado.

Schuett-Hames, D., A. Pleus, J. Ward, M. Fox, and J. Light. 1999. TFW monitoring program method manual for the large woody debris survey. Northwest Indian Fisheries Commission, TFW-AM9-99-004, Olympia, Washington. Available: nwifc.org/tfw/downloads.asp. (August 2009).

Smokorowski, K. E., and T. C. Pratt. 2007. Effect of change in physical structure and cover on fish habitat in freshwater ecosystems—a review and meta-analysis. Environmental Reviews 15:15–41.

Stem, C., R. Margoluis, N. Salatsky, and M. Brown. 2005. Monitoring and evaluations in conservation: a review of trend and approaches. Conservation Biology 19:295–309.

Swanson, F. J., G. W. Lienkaemper, and J. R. Sedell. 1976. History, physical effects, and management implications of large organic debris in western Oregon streams. U.S. Forest Service General Technical Report PNW-GTR-56.

Urquhart, N. S., S. G. Paulsen, and D. P. Larsen. 1998. Monitoring for policy-relevant regional trends over time. Ecological Applications 8:249–257.

Valier, T. L. 1995. Petrology of pre-Tertiary igneous rocks in the Blue Mountains region of Oregon, Idaho, and Washington: implications for the geologic evolution of a complex island arc. Pages 125–209 *in* T. L. Vallier and H. C. Brooks, editors. Geology of the Blue Mountains region of Oregon, Idaho, and Washington: petrology and tectonic evolution of pre-Tertiary rocks of the Blue Mountains region. U.S. Geological Survey, Professional Paper 1438.

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

Whitacre, H. W., B. B. Roper, and J. L. Kershner. 2007. A comparison of protocols and observer precision for measuring physical stream attributes. Journal of the American Water Resources Association 43:923–937.

Wohl, E. E., D. J. Anthony, S. W. Madsen, and D. M. Thompson. 1996. A comparison of surface sampling methods for coarse fluvial sediments. Water Resources Research 32(10):3219–3226.

Wolman, M. G. 1954. A method of sampling coarse river-bed material. Transactions of the American Geophysical Union 36:951–956.

Woodsmith, R. D., and J. M. Buffington. 1996. Multivariate geomorphic analysis of forest streams: implications for assessment of land use impacts on channel conditions. Earth Surface Processes and Landforms 21:377–393.

Zar, J. H. 1984. Biostatistical analysis, 2nd edition. Simon and Schuster, Englewood Cliffs, New Jersey.

Appendix 1: Monitoring Group Protocols

TABLE A.1.—Summary of monitoring group protocols for measuring stream attributes examined in this study. Some protocols modified as noted in the descriptions presented here.

Attribute and monitoring group	Protocol procedure
Length	Length of stream evaluated by each crew. All lengths are measured along the thalweg except for those in EMAP and RMRS, which are measured along the channel center line.
AREMP	20 times the average bank-full width (BFW), which is estimated from five evenly spaced measurements about the first transect (cross section), the spacing of the measurements being equal to the BFW of that transect. The average BFW is binned into width-classes (reach length is determined as 20 times the upper end of the class); minimum reach length, 160 m; maximum, 480 m.
CDFG	Modified to ~40 times BFW for this study.
EMAP	40 times the wetted stream width (based on five measurements evenly spaced within ~10 channel widths distance near the center of the reach). The minimum reach length is 150 m.
NIFC	Lengths measured along channel center line with a hip chain.
ODFW	Modified to ~40 times BFW for this study.
PIBO	Same as AREMP, but BFW is estimated from five evenly spaced measurements (every 16 m) over the first 64 m of stream length.
UC	20 times the mean BFW, but not less than 150 m or greater than 500 m.
USFS Rocky Mountain Research Station (RMRS)	40 times the average BFW (estimated from 11 evenly spaced measurements [every 20 m] over the first 200 m of stream length).
Gradient	Gradient is calculated by dividing the elevation change along the reach by the total reach length (as defined above).
AREMP	A total station is used to determine the water surface elevation (to the nearest centimeter) at each end of the reach, gradient being calculated as the difference in elevation divided by reach length. Elevations are measured twice during two separate passes through the reach, and if they differ by more than 10%, a third set of measurements is made and the closest two values averaged.
CDFG	Water surface gradient is measured between the bottom and top of the reach using a tripod-mounted, self-leveling level and a stadia rod. The elevation difference is divided by reach length measured along the thalweg.
EMAP	A clinometer is used to measure water surface elevations at 11 equally spaced transects, the intertransect slope values being averaged to produce a mean reach value. Intertransect slope is based on straight-line distances between the center of each transect. Supplemental slope measurements are taken between transects to avoid sighting across bends. Peck et al. (2006) describe acceptable alternative procedures for EMAP, including use of laser or hydrostatic levels.
NIFC	Not measured.

TABLE A.1.—Continued.

Attribute and monitoring group	Protocol procedure
ODFW	Water surface slope is measured with a clinometer at every habitat unit (each reach is comprised of at least 20 habitat units). The values are weighted by unit length and averaged across the reach length.
PIBO	Same as AREMP, but elevations are measured with an automatic level.
UC	Same as EMAP, but elevations are measured with a hand level and incremental slopes are determined between 21 equally spaced transects, straight-line distances being measured between thalweg crossings of each transect.
RMRS	Gradient is determined from a center line profile of the streambed surveyed with a total station and fit by linear regression. The survey is conducted to capture all major topographic breaks (rather than surveying at fixed intervals); the number of points per profile ranged from 118 to 532 across the study sites (3–12 points per channel width of stream length).
Sinuosity	Sinuosity is determined by dividing the reach length by the straight-line distance of the reach.
AREMP	Calculated as the thalweg length of the reach divided by the straight-line distance between the top and bottom of the reach.
CDFG	Not measured.
EMAP	Calculated as the sum of straight-line distances between the center points of 11 equally spaced transects divided by the straight-line distance of the reach. The straight-line distance is determined by trigonometry using compass bearings between transects (see Kaufmann et al. 1999).
NIFC	Not measured.
ODFW	Not measured.
PIBO	Same as AREMP.
UC	Not normally measured, but determined for this study as the sum of straight-line distances between thalweg crossings of 21 equally spaced transects divided by the straight-line distance of the reach.
RMRS	Center line length of the reach divided by the straight-line distance between each end of the reach.
Percent pools	Percent of reach length comprised of pool habitat, determined as the sum of pool lengths divided by reach length or the sum of pool surface area divided by total area of the reach (NIFC, RMRS). The definitions below largely focus on how pools are defined–identified and measured.
AREMP	Pools are defined as depressions in the streambed that are concave in profile, laterally and longitudinally. Pools are bounded by a head crest (upstream break in streambed slope) and a tail crest (downstream break in streambed slope). Only main-channel pools where the thalweg runs through the pool, and not backwater pools, are considered. Pools must span at least 90% of the wetted channel width at any one location within the pool. Pool length, measured along the thalweg from the head to the pool tail crest, must be greater than its width. Pool width is measured perpendicular to the thalweg at the widest point of the pool. Maximum pool depth must be at least 1.5 times the maximum depth of the pool tail crest.
CDFG	Pools are visually identified as "slower and less-turbulent" flow, usually deep compared with other parts of the channel. Pool length must be greater than its width, and width must be at least 60% of the wetted channel width. Pool length is measured along the thalweg, and percent of pools is the length of pools divided by the thalweg length.
EMAP	Pools are visually identified as "still water" with low velocity and smooth, glassy surfaces and are usually deep compared with other parts of the channel; they must be at least as wide or as long as the wetted channel width. Pool type is classified using definitions modified from Bisson et al. (1982) and Frissell et al. (1986). Calculation is percent of reach length.
NIFC	Pools are sections of stream channel with a closed topographical depression. Percent pools are calculated as the total wetted surface area of pools divided by the total wetted surface area of the reach. Pool lengths connect the longest dimensions of the unit upon which perpendicular width measurements are taken at a standard frequency determined by the length. A minimum pool depth criterion and wetted surface area is applied based on the mean bank-full width of the reach.
ODFW	Pools are visually identified by "slow water" and must be longer than they are wide (based on wetted dimensions). Pool length is measured from head crest to tail crest.
PIBO	Same as AREMP, but pools must span more than 50% of the wetted channel width.
UC	A pool must span more than half the wetted width, include the thalweg, be longer than it is wide (except in the case of plunge pools), and the maximum depth must be at least 1.5 times the tail depth. Pool length is measured from head crest to tail crest.
RMRS	Pools are visually identified as bowl-shaped depressions having a topographic head, bottom, and tail. Pools of all size are measured, without truncating the size distribution for requisite pool dimensions. "Pocket pools" formed by a single ring of coarse grains are not considered. The average topographic width and length of each pool are measured with a tape based on pool morphology and topographic breaks in the slope. Pool surface area is determined from these measurements and classification of pool shape (ellipse, rectangle, or triangle). Percent pools is determined as the sum of pool surface area divided by total bed area of the channel.
Residual pool depth	Residual pool depth is typically calculated by subtracting the pool tail depth from the maximum pool depth, both measured on the thalweg (Lisle and Hilton 1992).
AREMP	Averaged across all pools identified by this protocol.
CDFG	Averaged across all pools identified by this protocol.
EMAP	A computational routine based on reach average slope is applied to identify local riffle crests and residual depths along longitudinal profiles of thalweg depth from which residual pools and their dimensions are calculated (Kaufmann et al. 1999). Mean residual depth from this protocol is conceptually different from that of the other protocols, which all calculate the mean value of maximum RPDs in the reach. The EMAP protocol value is an expression of reachwide mean residual depth, the sum of all residual depths over the reach, including zero values (not just the maximum residual depths of pools) divided by the total number of depth measurements in the reach.

TABLE A.1.—Continued.

Attribute and monitoring group	Protocol procedure
NIFC	Calculated by subtracting the downstream "riffle crest" depth from the maximum pool depth for each pool. Residual pool depths are recorded for units that do not meet minimum criteria; they are classified as riffles.
ODFW	Averaged across all pools identified by this protocol.
PIBO	Averaged across all pools identified by this protocol.
UC	Averaged across all pools identified by this protocol; measured to the nearest 0.01 m.
RMRS	Elevations of pool bottoms and tails measured with a total station. Residual depth averaged across all pools identified by this protocol.
Pools/kilometer	Number of pools divided by reach length expressed in kilometer. Reach length and the methods for identifying and measuring pools as defined above (see Length and Percent Pools, respectively). Only pools with residual depth greater than or equal to 50 cm are counted by EMAP. Oregon Department of Fish and Wildlife sums all pools in primary and secondary channels and divides by primary channel length (in kilometers).
AREMP	Number of pools divided by reach length expressed in kilometers.
CDFG	Number of pools in the total reach is summed and divided by the reach length as measured along the thalweg to determine pools/kilometers.
EMAP	Number of pools with residual depth of 50 cm in the sampled reach multiplied by 1,000 m/reach length.
NIFC	Number of pools divided by reach length expressed in kilometers.
ODFW	Calculated as the number of pools in primary and secondary channels divided by the primary channel length, then standardized to 1 km.
PIBO	Number of pools divided by reach length expressed in kilometers.
UC	Estimated as the number of pools (as defined by the protocol) within a sampling reach times 1,000, divided by the length (m) of the sampling reach.
RMRS	
Bank-full width	Width of the channel at bank-full discharge, which is the flow at which the water just begins to spill onto the active flood plain. All protocols used bank-full indicators as described by Harrelson et al. (1994).
AREMP	Average value of 11 equally spaced transects surveyed with a total station.
CDFG	Average value of five approximately equally spaced transects at pool tail crests.
EMAP	Average value of 11 equally spaced transects measured with tape or stadia rod.
NIFC	An average value at 100-m intervals or, if less than 400-m stream length, an average value of five equally spaced transects.
ODFW	Average value of five equally spaced transects.
PIBO	Average value of 21 equally spaced transects measured with a tape.
UC	Average value of 21 equally spaced transects. Widths were measured to the nearest 0.1 m.
RMRS	Average value of 81 equally spaced transects surveyed with a total station.
Width-to-depth ratio	The width-to-depth ratio (W:D) is defined as the average bank-full width (BFW) divided by the average bank-full depth (BFD).
AREMP	Measured in unconstrained, relatively narrow, planar channel units (riffle–run–rapid), the location of which varies by channel type: (1) in pool–riffle (C, E, F) streams, measured in the riffle section where thalweg "crosses over" between successive pools; (2) in plane-bed (B) streams, measured in narrowest "rapid" section; and (3) in step-pool (A, G) streams, measured in the "run" section between step and pool head (Rosgen 1996). Transect surveyed with a total station in first downstream occurrence of the above locations. A minimum of 10 equally spaced measurements, left and right wetted edges, and thalweg are taken for the depth measurements.
CDFG	Determined from the same five transects used to measure BFW. At each transect, 20 equidistant measurements of BFD are averaged. The width-to-depth ratio is calculated as the average BFW of the five transects divided by the average BFD of the transects. Measurements are made with a tape and stadia rod.
EMAP	Determined from the same 11 transects used to measure BFW; measured with tape and stadia rod.
NIFC	Determined from the same transects used to measure BFW. Average of 10 evenly spaced BFD measurements across the BFW.
ODFW	Determined from the same five transects used to measure BFW.
PIBO	Determined from four transects measured at the widest point of the first four riffles–runs. Transects measured with tape and stadia rod.
UC	Determined from the 21 transects used to measure BFW. Mean BFD modified from normal method. Determined here from thalweg measurements using Kaufmann et al.'s (1999) method, assuming a triangular cross section; see Kaufmann et al. (2008) for further discussion.
RMRS	Average value of 81 equally spaced transects surveyed with a total station. The transect surveys are conducted to capture all major topographic breaks (rather than surveying at fixed intervals); the average number of points measured within the bank-full extent of each transect ranged from 9 to 21 across the study sites (average point spacing of 5–13% of BFW).
D_{50}	The median grain size (D_{50}) determined from Wolman (1954) pebble counts of intermediate grain diameters.
AREMP	Grid-by-number pebble count (Kellerhals and Bray 1971), grids being composed of 21 equally spaced transects, five particles per transect (105 total), sampled at equal increments across bed and banks within bank-full limit. Particle size measured with a ruler to the nearest millimeter.
CDFG	Grid-by-number pebble count, grids being composed of five equally spaced transects, 20 particles per transect (100 total), sampled across bed and banks within the bank-full limit.
EMAP	Same as AREMP, but particle sizes visually estimated into size-classes and sampling limited to wetted channel.
NIFC	Not measured.
ODFW	Not measured.
PIBO	Same as AREMP, but bank particles are excluded from calculation of substrate size.
UC	Same as EMAP, but sampled across bed and banks of bank-full channel.

TABLE A.1.—Continued.

Attribute and monitoring group	Protocol procedure
RMRS	Grid-by-number pebble count, grids being composed of 81 equally spaced transects, 10 particles per transect (810 total), sampled at equal increments across the active bed of each transect. Particle sizes measured with a gravelometer (one-half phi scale), sizes less than 2 mm binned into a single class.
Percent fines	Percent of the reach or specific portion of the streambed covered by fine sediment (sizes <6 or 2 mm, depending on specific protocol).
AREMP	Grid sampling (14 × 14 in) of 50 particles in three equally spaced locations along the wetted boundaries of each pool tail, fines defined as sizes less than 2 mm (measured with a ruler). Values are averaged for the first 10 pools at the site, those values then being averaged for the reach.
CDFG	Same as AREMP.
EMAP	Determined from above reach-average D_{50} pebble counts, fines defined as particles less than 2 mm.
NIFC	Not measured.
ODFW	Visually estimated in the wetted area of each habitat unit and defined as silt, organics, and sand less than 2 mm. The value reported for the reach is the average across all habitat units.
PIBO	Same as AREMP, but with fines defined as sizes less than 2 and less than 6 mm. Measurements are taken in the first 10 pools, three grids per pool, at 25, 50, and 75% the distance across the wetted channel.
UC	Same as EMAP, but sampled across bed and banks of bank-full channel.
RMRS	Determined from above reach-average D_{50} pebble counts, fines being defined as particles less than 2 mm.
LWD/100m	Defined as the total number of pieces of LWD divided by the reach length (m) and multiplied by 100.
AREMP	Large woody debris is defined as wood greater than or equal to 3 m long and greater than or equal to 0.3 m in diameter at breast height one-third of the way up from the base or largest end. Length and diameter are visually estimated and validated by periodic measurements (done for first 10 pieces in the reach and every fifth piece thereafter for sites with ≤100 pieces, or every 10th piece for sites with >100 pieces). Each piece fully–partially within the bank-full channel or suspended above it (Robison and Beschta 1990) is estimated for size, tallied, and classified as single pieces, accumulations (two to four pieces) or logjams (greater than five pieces).
CDFG	Large woody debris is defined as logs greater than or equal to 2 m long and greater than or equal to 0.3 m in diameter, or root wads with a trunk diameter greater than or equal to 0.3 m and root bole intact. All such pieces within the BFW are counted.
EMAP	Large woody debris is defined as wood greater than or equal to 1.5 m long and with a small-end diameter greater than or equal to 0.1 m. All such pieces within the bank-full channel or suspended above it are measured and binned in length- and large-end diameter-classes.
NIFC	Dead pieces of wood are counted if they are greater than 2 m long and greater than 0.1 m in diameter with at least 0.1 m of its length, and extend into the bank-full channel or are suspended above it.
ODFW	All dead wood greater than 3 m long and greater than 0.15 m in diameter that are within the bank-full channel or suspended above it are counted.
PIBO	Same as AREMP, but LWD is defined as pieces greater than or equal to 1 m in length and greater than or equal to 0.1 m in diameter at breast height measured at one-third the distance from the base, and all pieces are counted individually (i.e., logjams are not considered single units). For analysis in this paper only, pieces of LWD greater than 3 m long were included.
UC	Large woody debris is defined as wood greater than 1.0 m long and greater than 0.1 m in diameter. All such pieces within the bank-full channel (Robison and Beschta's [1990] zones 1–2) are measured and binned in three length- and diameter-classes.
RMRS	Large woody debris is defined as wood greater than or equal to 1 m long and greater than or equal to 0.1 m in diameter (Swanson et al. 1976). Sampled in four equally-spaced sections (each having a length of 10 BFWs), wood being classified into zones according to Robison and Beschta (1990) and function according to Montgomery et al. (1995).