**JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION**
**OCTOBER** **AMERICAN WATER RESOURCES ASSOCIATION** **2005**

# SOURCES OF VARIABILITY IN CONDUCTING PEBBLE COUNTS: THEIR POTENTIAL INFLUENCE ON THE RESULTS OF STREAM MONITORING PROGRAMS[1]

*Darren S Olsen, Brett B. Roper, Jeffrey L. Kershner, Richard Henderson, and Eric Archer*[2]

ABSTRACT: Pebble counts have been used for a variety of monitoring projects and are an important component of stream evaluation efforts throughout the United States. The utility of pebble counts as a monitoring tool is, however, based on the monitoring objectives and the assumption that data are collected with sufficient precision to meet those objectives. Depending upon the objective, sources of variability that can limit the precision of pebble count data include substrate heterogeneity at a site, differences in substrate among sample locations within a stream reach, substrate variability among streams, differences in when the substrate sample is collected, differences in how and where technicians pick up substrate particles, and how consistently technicians measure the intermediate axis of a selected particle. This study found that each of these sources of variability is of sufficient magnitude to affect results of monitoring projects. Therefore, actions such as observer training, increasing the number of pebbles measured, evaluating several riffles within a reach, evaluating permanent sites, and narrowing the time window during which pebble counts are conducted should be considered in order to minimize variability. The failure to account for sources of variability associated with pebble counts within the study design may result in failing to meet monitoring objectives.

(KEY TERMS: pebble count; variability; rivers and streams; sediment; monitoring; watershed management.)

Olsen, Darren S, Brett B. Roper, Jeffrey L. Kershner, Richard Henderson, and Eric 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 (JAWRA) 41(5):1225-1236.

## INTRODUCTION

The measurement of changes in stream substrate size is an important objective in many stream habitat monitoring programs. The pebble count (Wolman, 1954) is probably the most widely used to monitor the effects of sediment inputs from various land use activities (Platts *et al.,* 1983; MacDonald *et al.,* 1991; Kershner *et al.,* 2004a).

Pebble counts are currently being used in many large scale stream monitoring efforts in the United States (Kaufmann *et al.,* 1999; Kershner *et al.,* 2004a). Pebble counts have also been used to evaluate fine sediment deposition following land disturbance events (Potyondy and Hardy, 1994) and to assess the cumulative effects from a variety of land management activities (Bevenger and King, 1995; Schnackenberg and MacDonald, 1998).

The utility of pebble counts as a monitoring tool is partly based on the study question and the assumption that data are collected with sufficient precision to meet monitoring objectives. Sources of variability that can limit the precision of pebble counts can be associated with substrate heterogeneity at the measurement location (Wolcott and Church, 1991; Buffington and Montgomery, 1999a,b), differences in substrate among sample locations within the same stream reach (Kellerhals and Bray, 1971; Church and Kellerhals, 1978; Kondolf, 1997), differences in when the pebble counts were conducted (Larsen *et al.,* 2001), variation among streams (Roper *et al.,* 2002), differences in how and where observers pick up substrate particles (Marcus *et al.,* 1995; Bunte and Abt, 2001b), and how consistently observers measure a selected particle size (Hey and Thorne, 1983). Depending upon the specific sample design and monitoring question, each of these sources of variation has the potential to

---

[1] Paper No. 03126 of the *Journal of the American Water Resources Association* (JAWRA) (Copyright © 2005). **Discussions are open until April 1, 2006.**

[2] Respectively, Hydrologist (Olsen) and Aquatic Ecologists (Roper, Henderson, and Archer), U.S. Forest Service, Logan Forestry Sciences Laboratory, 860 North 1200 East, Logan, Utah 84321 (current address/Olsen: BIO-WEST, Inc., 1063 West 1400 North, Logan, Utah 84321); and Aquatic Ecologist (Kershner), U.S. Forest Service, Aquatic, Watershed and Earth Resources Department, Utah State University, 5210 University Hill, Logan, Utah 84322-5210 (E-Mail/Olsen: dolsen@bio-west.com).

preclude the detection of changes in particle size distribution.

Riffles are believed to be the most appropriate location to detect changes in particle size distribution because of their sensitivity to increased sediment supply (Parker and Klingeman, 1982; Dietrich *et al.,* 1989). Particle sizes become finer in "supply limited" riffles when increased sediment inputs overwhelm the bedload transport capacity (Dietrich *et al.,* 1989; Buffington and Montgomery, 1999b). However, repeated sampling from a single, fixed riffle to detect changes in particle size distribution can be problematic in meandering streams because the location and size of the monitoring site may migrate laterally and longitudinally over time (Knighton, 1998), especially in response to sediment pulses (Lisle, 1982; Benda and Dunne, 1997). While information collected at an individual riffle may be useful to evaluate differences at a site, more samples may be needed to characterize stream reaches or habitat conditions over larger areas.

Studies that attempt to determine changes at the stream reach or larger scales must evaluate the location and the number of sample units needed to characterize the area of interest. Investigators are often faced with questions regarding which riffle(s) should be sampled and how many riffles will be needed to characterize an entire reach (Larsen *et al.,* 2001, 2004). Large scale monitoring efforts may attempt to compare how particle size distributions change within streams and among multiple streams in order to assess the influence of land management. These efforts have shifted from using pebble counts within individual riffles to evaluating particle sizes at the reach scale for small streams (Bevenger and King, 1997; Kaufmann *et al.,* 1999; Kershner *et al.,* 2004a). Surface substrate is measured throughout a stream reach, and then reach wide metrics and distributions are reported. An advantage of this approach is that it can incorporate spatial variation within a reach if sample units expand, contract, or migrate through time (Lisle, 1982; Benda and Dunne, 1997; Madej, 2001). The disadvantage of this approach is that it may add variability associated with differences among riffles within a stream reach, differences among streams, and differences in the time of data collection during the sampling season (Urquhart *et al.,* 1998; Larsen *et al.,* 2001).

Several authors have identified additional sources of error that could confound the interpretation of pebble count data. The error associated with the selection and measurement of individual particles is an important source of variability in pebble count studies (Hey and Thorne, 1983; Marcus *et al.,* 1995; Bunte and Abt, 2001a,b). Temporal changes in particle size distribution during the base flow period have not been widely examined but could have significant influence on the ability to detect effects of land use changes (Larsen *et al.,* 2001). If changes do occur, the timing of repeated sampling must be adjusted to reduce error associated with seasonal variability.

While each of these components of variability has been evaluated separately in previous studies, any attempt to evaluate the relative contribution of all of these potential sources of variation is unknown. In this paper five sources of variability associated with pebble counts and the characterization of surface substrate are identified. Five questions are asked. First, how variable are different observers in measuring the same particles? Second, within a riffle, do individual observers consistently characterize the surface particle size distribution? Third, can the results from specific riffles be combined to give a consistent interpretation of reach scale particle size distribution? Fourth, within an individual riffle does the particle size distribution change within the low flow sampling period? Finally, how does each of these sources of variability compare to the variability found among streams?

## STUDY SITE AND METHODS

Several evaluations of pebble counts were conducted in 2000 and 2001 to isolate and quantify the variability associated with these measurements. These evaluations were part of a quality control program implemented for the U.S. Department of Agriculture (USDA) Forest Service and the U.S. Department of the Interior (USDI) Bureau of Land Management Interior Columbia River Basin monitoring program (Kershner *et al.,* 2004a).

Riffles were sampled in wadeable (< 15 m bankfull width) stream reaches (gradients < 3 percent) within the Upper Columbia River Basin (Figure 1). To minimize inclusion of finer textural patches on streambanks, only the portion of the riffle within the active channel was sampled. The active channel was defined as the portion of the stream channel below bankfull but excluded the streambank (defined as a vertical shift in stream channel morphology accompanied by rapid fining of substrate) and areas greater than 50 percent vegetated, such as islands and inactive bars. Each riffle was divided into four transects using a grid sampling scheme to account for within-riffle variability while minimizing sampling variability (Wolcott and Church, 1991; Wohl *et al.,* 1996). Pebbles were collected along transects using the heel-to-toe technique (Wolman, 1954; Leopold, 1970). Particles were selected by the observers reaching down to the tips of their boots, averting their gaze from the bottom of the stream until their finger came into contact with a pebble, and then looking down and picking up that pebble (Leopold, 1970). The intermediate axis of each particle (Bunte and Abt, 2001a) was measured using a ruler. All particles were measured to the nearest millimeter except those particles less than 4 mm. Particles less than 4 mm were assigned to a size class of up to 4 mm and given a value of 4 mm for analysis purposes.

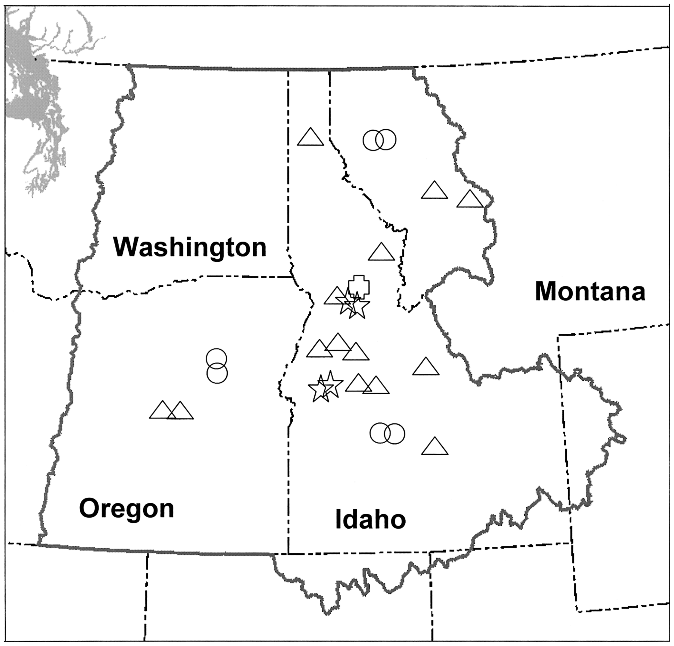

Figure 1. Sample Streams Within the Interior Columbia River Basin (solid line). The dashed lines are state boundaries. The stars are the four streams from which 100 pebbles were sampled and measured with a straight ruler. The open cross is the stream where seven observers individually sampled at least 100 pebbles each in the same riffles. The triangles represent the 14 streams from which at least 100 pebbles were sampled from eight consecutive riffles. The circles represent the six streams in which four riffles were sampled during six time periods within the summer.

Each individual received two days of training on conducting pebble counts prior to the study. Training included identification of riffles, delineation of the active channel within the riffle, particle selection, and particle measurement. Observers then sampled several test streams. Their results were evaluated and their techniques refined by field supervisors to further increase consistency.

Three metrics were derived from each sampled riffle: the median particle size ($D_{50}$), the particle size at which 16 percent of the material was smaller ($D_{16}$), and the particle size at which 84 percent of the material was smaller ($D_{84}$). These response variables were chosen because they represent a central measure as well as a measure of both tails of the cumulative distribution of particle sizes.

Two sources of variability were evaluated in this study; those due to differences among observers and those due to differences within and among sites.

### *Observer Variability*

Variability among observers can occur in two ways: differences in how an individual measures the intermediate axis of a particle (Hey and Thorne, 1983; Marcus *et al.,* 1995); and differences among observers in determining how, when, and where to select a pebble to be measured (Marcus *et al.,* 1995; Bunte and Abt, 2001a).

**Particle Measurement Variability.** Measurement variability was assessed by having multiple technicians measure the intermediate axis of 400 particles. The sample consisted of 100 particles selected by a single technician from riffles within four streams (Figure 1). At each stream the 100 particles were numerically identified and put on a table next to the stream. Each of the particles was then measured by multiple technicians who independently measured (by ruler) and recorded the diameter (in millimeters) of the intermediate axis. A total of 11 technicians were used, but not all technicians were available to sample all streams. Ten technicians sampled Study Stream One; Study Stream Two was evaluated with nine of the 10 technicians used in Study Stream One, and six individuals (five were part of the original 10) were used to sample the final two streams. Due to errors of omission, approximately 0.5 percent of the pebbles were not measured by all observers at a stream.

A random effects analysis of variance model (PROC MIXED of SAS,) (SAS Institute, 2001) was used to partition variance associated with differences among the particles in the sample and the variability in how observers measured them. Estimates of variance were evaluated by treating both particles and observers as a random effect within the model (Littell *et al.,* 1996). Estimates of the variance among pebbles and observers are additive and can be used to calculate the proportion of the variation due to each component (Montgomery, 1984; Roper *et al.,* 2002). In addition to partitioning variance, the differences in cumulative frequency distributions were compared among the 9 observers who measured the same 100 particles in each of two streams.

**Particle Selection Variability.** Particle selection variability among observers was determined by evaluating the particle selection of seven technicians. Each technician selected 50 particles from two consecutive riffles in a single stream reach that were then combined to make a sample size of 100 particles. The longitudinal and latitudinal boundary of the two riffles was defined so that all technicians sampled the same area. Each technician independently determined the transect locations and distance between samples to ensure that at least 50 rocks were collected in each riffle. After each particle was measured, it was placed as close as possible to the original position to minimize disturbance of the sample area. Evaluations were conducted on the same day.

A random effects analysis of variance model (PROC MIXED of SAS) (SAS Institute, 2001) was again used to partition variance associated with differences among the particles in the sample and the variability in how observers selected and measured them. Estimates of variance were evaluated with observers as a random effect and the intermediate particle size as the response variable. The model error term is the variation among the $D_{50}$ sizes of the particles within the sampled riffles. The above model also was run with observers as a fixed effect to determine if statistical differences existed among observers (Marcus *et al.,* 1995).

### *Site Variability*

Twenty streams (gradient < 3 percent) were sampled to evaluate variability in particle size distributions within a reach and between monitoring sites. Three components of site variability were evaluated: differences among streams, difference among riffles within a reach, and differences within the same riffles throughout the summer sampling season.

Streams were divided into two groups to evaluate different components of site variability. In the first group, 14 streams were sampled once during the summer low flow period. Eight consecutive riffles were identified, and a grid-based pebble count was conducted in which at least 100 particles were measured within each riffle. The second group consisted of six streams. In each of these streams at least 50 pebbles were collected and measured in each of four consecutive riffles. At the time of the initial survey, the upstream and downstream extent of each sampled riffle was flagged so it could be identified in later visits. Pebble counts were conducted in the same riffles every two weeks throughout the summer (late June to late August). Each reach was visited a total of six times. Each of the 20 surveyed stream reaches was evaluated by a crew of two technicians. The revisits were conducted by the same crew throughout the summer.

Pebble counts were conducted using the following protocols. Technicians identified the upstream and downstream extent of the riffle as well as the dimension of the active channel. Four equally spaced transects were sampled with a minimum of either 13 or 25 pebbles selected from each transect depending upon whether the goal for that riffle was a sample of 50 or 100 particles. In the six streams that were revisited during the summer, the measured pebbles were placed back in the original position to minimize disturbance of the sample area.

Analysis of variance (PROC MIXED of SAS®) (SAS Institute, 2001) was used to partition total variation in the average riffle $D_{50}$, $D_{16}$, and $D_{84}$ particle sizes into one of four sources: stream, riffle within a stream, differences within a riffle through the summer sampling window, and sampling error due to unexplained variation in results derived from return visits by the same technicians. Because of the unbalanced sample design, data from all streams were not used in the estimation of all variance components. Streams and riffles were treated as random effects, riffles were nested within a stream reach, and time within the summer season was treated as a repeated measure in this model. Compound symmetry was evaluated for the covariance structure of the repeated measure because the correlation was reduced between time periods, and this covariance structure better explained the data than an equal variance model (smaller Akaike's Information Criterion values) (Burnham and Anderson, 2002). A random effects model was used instead of a fixed effect model (Church and Kellerhals, 1978) because of the interest in estimating variance components, not in determining if significant differences exist among streams, among riffles within streams, or within a riffle through time.

Analysis of $D_{50}$, $D_{16}$, and $D_{84}$ particle sizes was conducted on untransformed data. The assumption of statistical normality was justified by the central limit theorem, not the underlying distribution of the individual values (Urquhart *et al.,* 1998). Tests for normality failed to reject the null hypothesis for $D_{50}$ and indicated a slight departure from normality for $D_{16}$ ($p < 0.05$, Shapiro-Wilks = 0.92) and $D_{84}$ ($p < 0.05$, Shapiro-Wilks = 0.96). The analysis of residuals indicated that approximately 10 of the almost 250 sampled riffles had higher values than would be expected given a normal distribution. These small departures from normality likely had a limited effect on the results (Zar, 1996).

One component of site variability could be that the numbers of pebbles collected are insufficient to adequately describe the particle distribution. To evaluate this component of sampling error, an analysis of how often an unbiased selection of pebbles resulted in precise estimates of pebble size distribution metrics was performed. Since there are no direct simple ways to estimate the accuracy of $D_{50}$, $D_{16}$, and $D_{84}$ (most standard statistical approaches evaluate means) (see Rice and Church, 1996), six of the 14 streams where the same observers measured at least 800 particles were selected. Six streams were selected because the surface substrate represented a gradient of fine grained to coarse grained particle sizes (Figure 2). For each stream, 1,000 independent samples were drawn with replacement (Efron and Tibshirani, 1993) from the total sample of 800 particles. This procedure was repeated for each of the following sample sizes: 25, 50, 100, 200, 300, and 500 pebbles. After a sample was drawn, the $D_{50}$, $D_{16}$, and $D_{84}$ were calculated and evaluated to determine whether the calculated metrics were within 10 percent of the metric value calculated from the 800 rocks sampled from that stream. A sufficient number of particles had been evaluated when the sample values derived from pebble counts were within 10 percent of the true values in greater than 900 of the 1,000 simulations. This level of precision, if it could be achieved, would be effective in reducing sampling error.

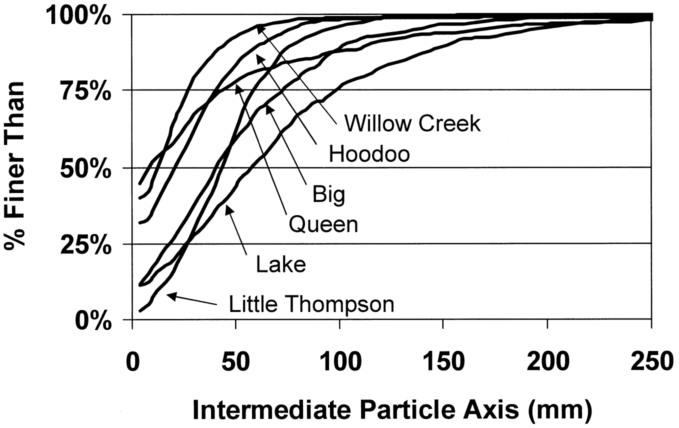

Figure 2. The Distribution of Pebbles in the Six Streams in Which Monte Carlo Simulations Were Conducted.

## RESULTS

### *Observer Variability*

**Particle Measurement Variability.** The intermediate diameters of the 400 particles ranged from 5 mm to 150 mm. The $D_{16}$, $D_{50}$, and $D_{84}$ were 15, 32, and 59 mm, respectively. Almost all (98.5 percent) of the variance was attributable to differences among the sizes of particles within the sample, while the remaining 1.5 percent of the variability was associated with differences in measurements among observers. The average variation (one standard deviation) among observers when measuring a particle with an intermediate axis diameter of 37.7 mm (the arithmetic mean of all 400 particles) was 2.9 mm.

The size distribution was very similar for the 200 particles in the two streams where the same nine technicians measured the same particles (Figure 3). The maximum difference between the highest and lowest observed values was 1 mm for $D_{16}$, 2 mm for $D_{50}$, and 7 mm $D_{84}$.

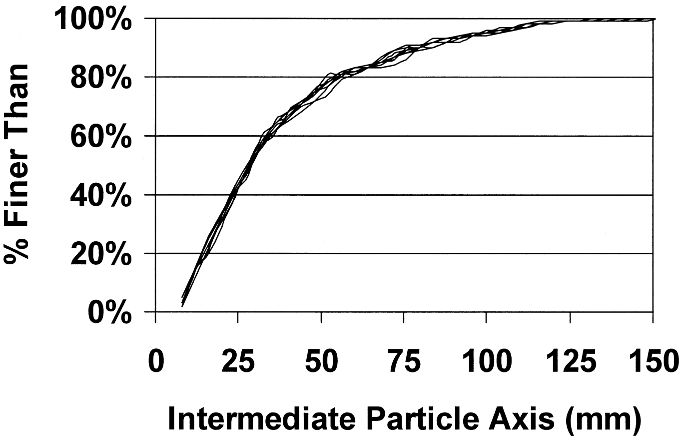

Figure 3. The Particle Size Distributions Obtained by Nine Observers Who Measured the Intermediated Axis of the Same 200 Particles.

**Particle Selection Variability.** The largest proportion of the variability associated with evaluating the same two riffles with seven observers was due to differences among pebbles within the riffle (99 percent) rather than differences among observers (1 percent). While all technicians did not arrive at the same derived value, the shapes of the cumulative distributions were similar (Figure 4). Estimates of $D_{50}$ ranged from 23 to 30 mm, $D_{16}$ ranged from 9 to 17 mm, and $D_{84}$ ranged from 40 to 56 mm.

The test for differences among the seven technicians found no experiment wide statistical differences ($p > 0.05$). However, there was a clear break between six of the observers' mean values, which ranged from 26.9 mm to 30.9 mm, and the last observer, whose mean was 34.3 mm. This could be random error or indicative of a bias in one observer.

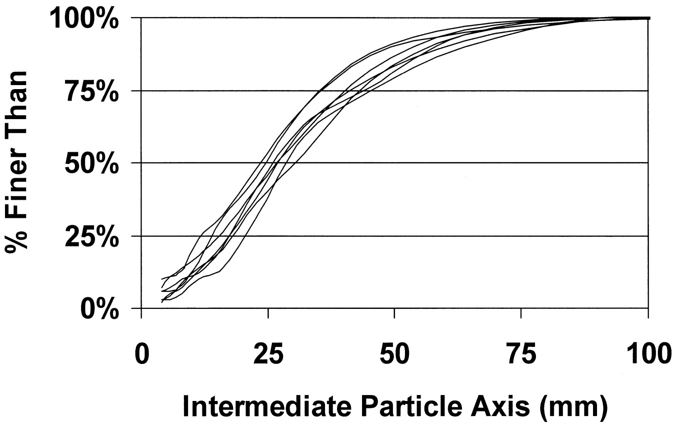

Figure 4. Cumulative Frequency Distribution of Particle Size for the Seven Observers Who Evaluated the Same Two Riffles.

*Site Variability*

**$D_{50}$.** The overall mean $D_{50}$ within the 20 streams sampled was 39.5 mm (Table 1). The overall variation in $D_{50}$ was high, with the coefficient of variation, calculated as (standard deviation/mean*100), equal to 59 percent. The largest percentage of the total variance (56.2 percent) was due to differences among streams, while 21.4 percent was due to differences among riffles within a stream, 8.3 percent was due to the timing of the survey during the summer, and the remaining 14.1 percent was sampling error. The average $D_{50}$ for the 20 streams ranged from 12.6 to 72.1 mm (Table 2). The greatest variation in $D_{50}$ among riffles within a single stream reach was from 33 to 105 mm. The greatest variation in $D_{50}$ within a single riffle during the sample window was from 32 to 78 mm.

**$D_{16}$.** The overall mean $D_{16}$ within the 20 sample streams was 10.8 mm (Table 1). The overall variation in $D_{16}$ was higher than for $D_{50}$, with a coefficient of variation of 88. The largest percentage of the total variation (45.6 percent) was due to differences among streams – 20.3 percent was due to differences among riffles within a stream, 9.5 percent was due to the stream being surveyed during the summer, and the remaining 24.6 percent was sampling error. The average $D_{16}$ for the 20 streams ranged from up to 4 to 28 mm (Table 3). The greatest variation in $D_{16}$ among riffles within a single stream reach was from 6 to 56 mm. The greatest variation in $D_{16}$ within a single riffle during the sample window was from 15 to 56 mm.

**$D_{84}$.** The overall mean $D_{84}$ within the 20 streams was 90.7 mm (Table 1). The overall variation in $D_{84}$ was lower than for $D_{50}$, with a coefficient of variation equal to 48. The largest source of variance was again due to differences among streams (59.6 percent), while 26.7 percent was due to differences among riffles within a stream, 2.6 percent was due to when the stream was surveyed during the summer, and the remaining 11.1 percent was sample error. The average $D_{84}$ for the 20 streams ranged from 36.8 to 156 mm (Table 4). The greatest variation in $D_{84}$ among riffles within a single reach was from 40 to 220 mm, while the greatest variation in $D_{84}$ within a single riffle during the sample window was from 102 to 180 mm.

TABLE 1. Results From Partitioning Variance to Evaluate Site Heterogeneity.

| | Metric | | |
|---|---|---|---|
| | $D_{50}$ | $D_{16}$ | $D_{84}$ |
| **Value** | | | |
| Mean (mm) | 39.5 | 10.8 | 90.7 |
| Standard Deviation (mm) | 23.2 | 9.5 | 43.4 |
| **Variance ($s^2$ in mm) Due to** | | | |
| Stream | 302.2 | 41.3 | 1,121.9 |
| Riffle Within Stream | 115.0 | 18.4 | 503.8 |
| Visit | 44.7 | 8.6 | 48.9 |
| Sample Error | 76.0 | 22.3 | 208.6 |
| Total | 537.9 | 90.6 | 1,883.2 |
| **Percent Variance ($s^2$) Due to** | | | |
| Stream | 56.2 | 45.6 | 59.6 |
| Riffle Within Stream | 21.4 | 20.3 | 26.7 |
| Visit | 8.3 | 9.5 | 2.6 |
| Sample Error | 14.1 | 24.6 | 11.1 |
| Total | 100.0 | 100.0 | 100.0 |

Notes: An estimate of stream variability is derived from 20 streams, estimates of riffle variability are based on differences between riffles within each stream, and visit variability is due to changes within a riffle throughout the summer. Variance is expressed both in units (mm) and in percent.

**Sampling Error Related to the Number of Pebbles Collected.** Sample sizes required to be consistently within 10 percent of a true value varied dramatically among metrics and streams. The most consistently determined value was $D_{16}$ in streams where the particle size was within the size class up to 4 mm. In these streams (Hoodoo, Queen, and Willow Creeks) the $D_{16}$ could be measured accurately (±10 percent in at least 900 of the 1,000 simulations) with samples of 25 particles (Table 5). In contrast, the $D_{16}$ was the most difficult to accurately estimate when

TABLE 2. Estimated Stream Values for $D_{50}$ (mm) in the 20 Evaluated Streams.

| Stream | Average | Among Riffles | | Within Riffles | |
|---|---|---|---|---|---|
| | | Minimum | Maximum | Minimum | Maximum |
| Bench | 26.1 | ≤ 4 | 36 | 23 | 35 |
| Cape | 44.2 | 27 | 65 | 31 | 56 |
| East Fork Meadowbrook | 32.8 | ≤ 4 | 50 | ≤ 4 | 40 |
| Indian | 63.9 | 32 | 97 | 32 | 78 |
| Martin | 72.1 | 44 | 113 | 70 | 113 |
| Rock | 70.9 | 33 | 105 | 46 | 91 |
| Big | 40.6 | 30 | 40 | Not Repeated | |
| Boulder | 43.6 | 20 | 63 | Not Repeated | |
| Cougar | 49.0 | ≤ 4 | 63 | Not Repeated | |
| Goose | 48.3 | 23 | 86 | Not Repeated | |
| Hoodoo | 20.1 | 10 | 32 | Not Repeated | |
| Lake | 60.0 | 48 | 77 | Not Repeated | |
| Little Thompson | 44.3 | 33 | 52 | Not Repeated | |
| Morgan | 16.3 | 11 | 23 | Not Repeated | |
| Peterson | 20.5 | ≤ 4 | 31 | Not Repeated | |
| Queen | 12.6 | ≤ 4 | 29 | Not Repeated | |
| Roback | 19.6 | 6 | 65 | Not Repeated | |
| Salmon | 41.9 | 28 | 78 | Not Repeated | |
| WF Smith | 68.3 | 58 | 88 | Not Repeated | |
| Willow | 13.5 | 5 | 24 | Not Repeated | |

Notes: The average value is the overall estimate for the stream. The minimum and maximum values among riffles within a stream and with in the most variable single riffle though time are presented.

TABLE 3. Estimated Stream Value for $D_{16}$ (mm) in the 20 Evaluated Streams.

| Stream | Average | Among Riffles | | Within Riffles | |
|---|---|---|---|---|---|
| | | Minimum | Maximum | Minimum | Maximum |
| Bench | 6.9 | ≤ 4 | 15 | ≤ 4 | 13 |
| Cape | 7.0 | ≤ 4 | 19 | ≤ 4 | 19 |
| East Fork Meadowbrook | 4.1 | ≤ 4 | 5 | ≤ 4 | 5 |
| Indian | 20.0 | ≤ 4 | 37 | 16 | 37 |
| Martin | 22.9 | 6 | 56 | 15 | 56 |
| Rock | 17.5 | ≤ 4 | 36 | 10 | 21 |
| Big | 10.9 | ≤ 4 | 31 | Not Repeated | |
| Boulder | 5.6 | ≤ 4 | 12 | Not Repeated | |
| Cougar | 10.3 | ≤ 4 | 23 | Not Repeated | |
| Goose | 13.4 | 5 | 23 | Not Repeated | |
| Hoodoo | 6.1 | ≤ 4 | 13 | Not Repeated | |
| Lake | 14.1 | ≤ 4 | 23 | Not Repeated | |
| Little Thompson | 20.3 | 14 | 25 | Not Repeated | |
| Morgan | 4.8 | ≤ 4 | 7 | Not Repeated | |
| Peterson | 7.1 | ≤ 4 | 14 | Not Repeated | |
| Queen | ≤ 4.0 | ≤ 4 | ≤ 4 | Not Repeated | |
| Roback | ≤ 4.0 | ≤ 4 | ≤ 4 | Not Repeated | |
| Salmon | 12.3 | 8 | 18 | Not Repeated | |
| WF Smith | 28.0 | 14 | 35 | Not Repeated | |
| Willow | ≤ 4.0 | ≤ 4 | ≤ 4 | Not Repeated | |

Notes: The average value is the overall estimate for the stream. The minimum and maximum values among riffles within a stream and with in the most variable single riffle though time are presented.

TABLE 4. Estimated Stream Value for $D_{84}$ (mm) in the 20 Evaluated Streams.

| Stream | Average | Among Riffles | | Within Riffles | |
|---|---|---|---|---|---|
| | | Minimum | Maximum | Minimum | Maximum |
| Bench | 48.9 | 13 | 77 | 13 | 50 |
| Cape | 98.0 | 60 | 130 | 64 | 119 |
| East Fork Meadowbrook | 89.2 | 67 | 118 | 67 | 118 |
| Indian | 124.1 | 72 | 200 | 130 | 200 |
| Martin | 146.8 | 113 | 191 | 113 | 191 |
| Rock | 156.0 | 102 | 180 | 102 | 180 |
| Big | 89.4 | 68 | 104 | Not Repeated | |
| Boulder | 123.5 | 40 | 220 | Not Repeated | |
| Cougar | 103.0 | 29 | 126 | Not Repeated | |
| Goose | 106.7 | 47 | 172 | Not Repeated | |
| Hoodoo | 51.8 | 29 | 70 | Not Repeated | |
| Lake | 131.3 | 100 | 160 | Not Repeated | |
| Little Thompson | 68.9 | 58 | 77 | Not Repeated | |
| Morgan | 36.8 | 25 | 51 | Not Repeated | |
| Peterson | 38.9 | 25 | 57 | Not Repeated | |
| Queen | 85.0 | 35 | 143 | Not Repeated | |
| Roback | 68.9 | 13 | 150 | Not Repeated | |
| Salmon | 101.8 | 75 | 136 | Not Repeated | |
| WF Smith | 122.4 | 90 | 190 | Not Repeated | |
| Willow | 37.4 | 26 | 58 | Not Repeated | |

Notes: The average value is the overall estimate for the stream. The minimum and maximum values among riffles within a stream and within the most variable single riffle though time are presented.

particle sizes were greater than 4 mm (Big, Lake) (Table 5). Samples of more than 500 particles were needed to be within 10 percent of the true value of the $D_{16}$ in more than 90 percent of the bootstrap simulations. The $D_{50}$ was estimated accurately with 500 particles or less in half the streams evaluated. In general, except when the $D_{16}$ particle size was in the size class of up to 4 mm, estimates of $D_{84}$ were most likely to be accurate, estimates of $D_{50}$ were intermediate, and $D_{16}$ was least accurate when smaller numbers of particles were selected. There was one stream in the sample in which the true $D_{50}$ could be consistently estimated with 100 pebbles. This stream (Little Thompson) also had the most normal particle size distribution.

Results of these simulations suggest that particle size distributions from a single unbiased observer could vary dramatically because, in general, 100 pebbles are not sufficient to produce results repeatable within 10 percent of the true value.

## DISCUSSION

Variability associated with the measurement of stream substrate was highest among streams, followed by differences among riffles in the same stream, sampling error, and temporal differences during the summer sampling period. Each of these environmental sources of variability was greater than those due to differences among observers. These findings suggest that environmental heterogeneity is a greater source of variability than differences among observers when the monitoring objective is to compare differences among streams or riffles or to monitor change over time. This should not be interpreted to mean that substrate monitoring programs can ignore observer variation. For example, in projects where the objective is to evaluate change within a specific riffle over short timeframes, conclusions will be affected primarily by sources of variation associated with particle selection and measurement. In these studies, variability due to environmental heterogeneity is minimized because sampling is usually conducted at a permanent site (Roper *et al.,* 2003).

This study and others (Marcus *et al.,* 1995; Bunte and Abt, 2001b) suggest that when the project objective is to monitor changes at a specific location, it is difficult to reduce differences in pebble count metrics among observers below 10 to 15 percent. One suggested solution has been to use a single technician to conduct pebble counts so that observer variability will be minimized (Wohl *et al.,* 1996). While this approach may reduce errors associated with observer measurement in short term, small scale studies, this approach

TABLE 5. Number of Monte Carlo Simulations Out of 1,000 That Were Within 10 Percent of the True Population Value.

| SS | Stream | | | | | |
|---|---|---|---|---|---|---|
| | **Big (41)** | **Hoodoo (22)** | **Lake (59)** | **Lake Thompson (44)** | **Queen (9)** | **Willow (13)** |
| | | | **$D_{50}$** | | | |
| 25 | 404 | 273 | 289 | 576 | 29 | 206 |
| 50 | 500 | 339 | 415 | 744 | 75 | 280 |
| 100 | 699 | 431 | 583 | 912 | 105 | 383 |
| 200 | 862 | 621 | 729 | 980 | 135 | 501 |
| 300 | 933 | 685 | 800 | 984 | 150 | 652 |
| 500 | 979 | 838 | 907 | 1,000 | 154 | 751 |
| **SS** | **Stream** | | | | | |
| | **Big (9)** | **Hoodoo (≤ 4)** | **Lake (14)** | **Lake Thompson (20)** | **Queen (≤ 4)** | **Willow (≤ 4)** |
| | | | **$D_{16}$** | | | |
| 25 | 114 | 935 | 106 | 191 | 987 | 995 |
| 50 | 127 | 988 | 114 | 300 | 1,000 | 1,000 |
| 100 | 57 | 1,000 | 132 | 456 | 1,000 | 1,000 |
| 200 | 198 | 1,000 | 172 | 624 | 1,000 | 1,000 |
| 300 | 223 | 1,000 | 190 | 731 | 1,000 | 1,000 |
| 500 | 246 | 1,000 | 205 | 846 | 1,000 | 1,000 |
| **SS** | **Stream** | | | | | |
| | **Big (91)** | **Hoodoo (52)** | **Lake (127)** | **Lake Thompson (69)** | **Queen (77)** | **Willow (36)** |
| | | | **$D_{84}$** | | | |
| 25 | 457 | 358 | 366 | 566 | 128 | 302 |
| 50 | 613 | 496 | 505 | 715 | 188 | 418 |
| 100 | 790 | 668 | 659 | 848 | 289 | 523 |
| 200 | 928 | 835 | 819 | 956 | 396 | 672 |
| 300 | 973 | 885 | 890 | 989 | 457 | 762 |
| 500 | 993 | 972 | 975 | 998 | 588 | 873 |

Notes: Sample size (SS) is the number of particles sampled at random from the population of 800+ particles. For example, if 1,000 independent samples of 25 particles were chosen at random from those 800 particles sampled in Big Creek, 404 would be ±10 percent of the $D_{50}$ value of 41. The value in parentheses is the derived metric for that stream. These values may deviate from the previous table because they were derived from all measured particles, while the previous values are based on an average of either 8 or 16 riffle values.

is likely to be impractical if the scale of the study is large or continues over a long period.

Standardized training will increase data comparability if multiple observers are used to conduct monitoring studies (Wohl *et al.,* 1996). Following two days of training, little evidence of a bias was found among observers in either selecting or measuring pebbles. The concern that observers undersample small particles relative to their distribution (Wolman, 1954; Fripp and Diplas, 1993; Bunte and Abt, 2001b) and undersample large particles relative to their volumetric contribution (Leopold, 1970; Diplas and Sutherland, 1988; Bunte and Abt, 2001b) can be addressed by using consistent, repeatable methods and providing adequate training.

One source of variation that had little effect on variability regardless of the monitoring objective was the difference among how observers measured the intermediate axis of a particle with a ruler. In this study, the differences among observers measuring the same pebbles were minimal, suggesting that the precision gained by using a large size class template may not be justified. In the measurement component of this study, the average particle size of the 400 measured particles was 37.7 mm, and the average observer error (±1 standard deviation) was within 3 mm of this value. In contrast, the use of a template based on 0.5 Φ sizes would place this particle in the size class between 32 and 45 mm. This is a range of 13 mm, almost one-third the average particle size of this study and far larger than the average observer error. While there are graphical methods to estimate specific percentiles of interest when using size classes (Yang, 2003), these estimates are only as good as the assumption of a log-normal distribution on which they are based. A second disadvantage of 0.5 $\Phi$ size class template is that it results in fewer classes than are suggested for statistical analysis (30 to 300) (Sokal and Rohlf, 1995). Therefore, it is suggested that pebbles be measured with a ruler to the nearest millimeter or be evaluated using a 0.25 $\Phi$ or smaller size class template (see Hey and Thorne, 1983).

Large differences were observed in the particle size distribution among riffles within the same low gradient stream reach. Specifically, the largest $D_{50}$ was 4.9 times greater than the smallest $D_{50}$ in a single reach. While such variation in particle size distributions over small spatial scales is not surprising (Wolcott and Church, 1991), it has important implications for monitoring studies where investigators are interested in comparing changes in particle size distribution at the reach scale (Buffington and Montgomery, 1999b) or are unsure they can return to exactly the same riffle through time. The large variability among riffles in the same reach suggests the need to sample all riffles within the reach and to sample a larger number of particles from each of those riffles. The number of potential sample riffles with a reach that is 20 times bankfull width in length is generally low (< 6 in this study). Given the variability in the particle distribution among riffles in the same reach, it seems reasonable that monitoring programs interested in reach scale values sample as many riffles as possible to eliminate uncertainty associated with subsampling at the reach scale.

Variability at the reach scale could also be reduced by sampling more particles. A large part of the total error was due to the sampling error associated with collecting 100 or fewer particles from each riffle (Rice and Church, 1996; Bunte and Abt, 2001b). Monte Carlo simulations indicated that this number of pebbles would rarely be sufficient to estimate particle size metrics with any precision. Increasing the total number of pebbles sampled at a riffle from 100 to 300 can greatly increase precision in estimating $D_{50}$ or $D_{84}$ particle sizes (Rice and Church, 1996; Green, 2003). An even larger sample of particles will be needed to estimate the $D_{16}$ if no part of the particle size distribution falls into a size class $\leq 4$ mm. Randomly selecting a large number of particles should also help increase accuracy, as this larger sample could (if the observers are properly trained) more clearly define the true distribution. Practical limitations, however, suggest that there are negligible gains in precision after 400 particles have been collected (Rice and Church, 1996).

This study found considerable variability in particle size distributions among streams. This may influence the ability to understand how particle size distributions change through space and/or time as a result of management practices. Sampling and analysis strategies that may be useful in reducing this variability include stratification (Montgomery and MacDonald, 2002), sampling at permanent sites (Roper *et al.,* 2003), and the use of analysis of covariance to control for concomitant independent variables (Roper *et al.,* 2002; Kershner *et al.,* 2004b). These strategies will only work if methods used to describe the pebble distribution within a stream reach can be consistently repeated by different observers.

A final change in sampling protocol that would increase comparability among streams if the monitoring objective is to evaluate particle size distributions over several years would be to start sampling later in the summer. The biggest difference within the base flow sampling window occurred between the first sampling period (the last two weeks of June) and the second sampling period (the first two weeks of July, after high spring runoff flows). In this time frame, the average $D_{16}$ particle size increased from 7.9 mm to 11.3 mm. After this time period, the average weekly $D_{16}$ value varied between 10.0 and 13.5 mm with no consistent pattern. This suggests that the winnowing of fines, which likely caused the increase in $D_{16}$, occurs early in the summer during the receding limb of the annual hydrograph. If sampling started after July 1, at least within the study area of the Upper Columbia River Basin, total variability could likely be reduced. However, the investigator must weigh the tradeoff between shortening the sampling period in order to reduce a minor component of the total variation and reducing the overall sample size by limiting the number of weeks available to sample. Within season variation will vary regionally. Due to climatic, geologic, and land use influences, it is suggested that investigators working in different physiographic regions evaluate the duration and timing within sampling seasons so as to minimize within season variability in pebble counts.

## CONCLUSION

There was considerable variability associated with the application of pebble counts to stream habitat monitoring. Depending upon the sample design and desired level of precision, each source of variability could affect the ability to meet monitoring objectives. For projects evaluating bed material particle size distributions at a single site, variability associated with different observers could preclude detecting changes of less than 15 percent. In contrast, studies evaluating changes in bed material particle size distributions at a regional scale must be more concerned with variability among streams and among riffles within streams than with differences among trained observers.

The sources of variability associated with pebble counts at the scale of the project must be accounted for within the sample design to ensure that monitoring objectives are met. Depending upon the specific objective, actions that can reduce variability include observer training, increasing the number of pebbles measured, evaluating several riffles within a reach, evaluating more reaches, and narrowing the time window during which pebble counts are conducted. Failure to account for specific sources of variability associated with a specific sample design can lead to a failure in meeting monitoring objectives (Bunte and Abt 2001b; Roper *et al.,* 2002; Larsen *et al.,* 2004).

## ACKNOWLEDGMENTS

The authors thank the numerous individuals who collected the data used in this study. Regions 1, 4, and 6 of the USDA Forest Service and the USDI Bureau of Land Management state offices of Oregon/Washington and Idaho provided funding for this project. The authors thank Jack Schmidt for an early review of this paper, and the paper was greatly improved by the comments of four anonymous reviewers. Darren Olsen's involvement with this project was for partial fulfillment of his Ph.D. Dissertation at Utah State University.

## LITERATURE CITED

Benda, L.E. and T. Dunne, 1997. Stochastic Forcing of Sediment Routing and Storage in Channel Networks. Water Resources Research 33:2865-2880.

Bevenger, G.S. and R.M. King, 1995. A Pebble Count Procedure for Assessing Watershed Cumulative Effects. Research Paper RM-RP-319, U.S. Dept. of Agriculture, Forest Service, Rocky Mountain Forest and Range Experimental Station, Fort Collins, Colorado, 17 pp.

Buffington, J.M. and D.R. Montgomery, 1999a. A Procedure for Classifying Textural Facies in Gravel-Bed Rivers. Water Resources Research 35:1903-1914.

Buffington, J.M. and D.R. Montgomery, 1999b. Effects of Hydraulic Roughness on Surface Textures of Gravel-Bed Rivers. Water Resources Research 35:3507-3522.

Bunte, K. and S.R. Apt, 2001a. Sampling Surface and Subsurface Particle-Size Distributions in Wadable Gravel- and Cobble-Bed Streams for Analysis in Sediment Transport, Hydraulics, and Streambed Monitoring. General Technical Report RMRS-GTR-74, U.S. Dept. of Agriculture, Forest Service, Rocky Mountain Forest and Range Experimental Station, Fort Collins, Colorado, 428 pp.

Bunte, K. and S.R. Abt, 2001b. Sampling Frame for Improving Pebble Count Accuracy in Coarse Gravel-Bed Streams. Journal of the American Water Resources Association (JAWRA) 37(4):1001-1014.

Burnham, K.B. and D.R. Anderson, 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York, New York.

Church, M. and R. Kellerhals, 1978. On the Statistics of Grain Size Variation Along a Gravel River. Can. J. Earth Sci. 15:1151-1160.

Dietrich, W.E., J.E. Kirchner, H. Ikeda, and F. Iseya, 1989. Sediment Supply and the Development of the Coarse Surface Layer in Gravel-Bed Rivers. Nature 340:215-217.

Diplas, P. and A.J. Sutherland, 1988. Sampling Techniques for Gravel Sized Sediments. Journal of Hydraulic Engineering 114:484-501.

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman and Hill, London, United Kingdom.

Fripp, J. and P. Diplas, 1993. Surface Sampling in Gravel Streams. Journal of Hydraulic Engineering 119(4):473-489.

Green, J.C., 2003. The Precision of Sampling Grain-Size Percentiles Using the Wolman Method. Earth Surface Processes and Landforms 28:979-991.

Hey, R.D. and C.R. Thorne, 1983. Accuracy of Surface Samples From Gravel Bed Material. ASCE Journal of Hydraulic Engineering 109:842-851.

Kaufmann, P.R., P. Levine, E.G. Robison, C. Seeliger, and D.V. Peck, 1999. Quantifying Physical Habitat in Wadable Streams. EPA/620/R-99/003, U.S. Environmental Protection Agency, Washington D.C.

Kellerhals, R. and D.I. Bray, 1971. Sampling Procedures for Coarse Fluvial Sediments. Journal of Hydraulic Division ASCE 97(HY8):1165-1180.

Kershner, J.L., E.K. Archer, M. Coles-Ritchie, E.R. Cowley, R.C. Henderson, K. Kratz, C.M. Quimby, D.L. Turner, L.C. Ulmer, and M.R. Vinson, 2004a. Guide to Effective Monitoring of Aquatic and Riparian Resources. Gen. Tech. Rep. RMRS-GTR-121, U.S. Department of Agriculture, Rocky Mountain Research Station, Fort Collins, Colorado, 57 pp.

Kershner, J.L., B. Roper, N. Bouwes, R.C. Henderson, and E.A. Archer, 2004b. A Comparison of Stream Reach Habitat Characteristics in Managed and Reference Watersheds on Federal Lands Within the Columbia Basin. North American Journal of Fisheries Management 24:1353-1365.

Knighton, D., 1998. Fluvial Forms and Processes: A New Perspective. Arnold Publishing, London, United Kingdom.

Kondolf, G.M., 1997. Application of the Pebble Count: Notes on Purpose, Method, and Variants. Journal of the American Water Resources Association (JAWRA) 33(1):79-87.

Larsen, D.P., P.R. Kaufman, T.M. Kincaid, and N.S. Urquhart, 2004. Detecting Persistent Change in the Habitat of Salmon Bearing Streams in the Pacific Northwest. Canadian Journal of Fish and Aquatic Sciences 61:283-291.

Larsen, D.P., T.M. Kincaid, S.E. Jacobs, and N.S. Urquhart, 2001. Designs for Evaluation Local and Regional Trends. Bioscience 51:1069-1078.

Leopold, L.B., 1970. An Improved Method for Size Distribution in Stream-Bed Gravel. Water Resources Research 6:1357-1366.

Lisle, T.E., 1982. Effects of Aggradation and Degradation on Riffle-Pool Morphology in Natural Gravel Channels, Northwestern California. Water Resources Research 18:1643-1651.

Littell, R.C., G.A. Miliken, W.W. Stroup, and R.D. Wolfinge, 1996. SAS® System for Mixed Models. SAS Institute Inc. Cary, North Carolina, 633 pp.

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. EPA-910/9-001, U.S. Environmental Protection Agency, Region 10, Seattle Washington.

Madej, M.A., 2001. Development of Channel Organization and Roughness Following Sediment Pulses in Single-Thread, Gravel Bed Rivers. Water Resources Research 37:2259-2272.

Marcus, W.A., S.C. Ladd, J.A. Stoughton, and J.W. Stock, 1995. Pebble Counts and the Role of User-Dependent Bias in Documenting Sediment Size Distribution. Water Resources Research 31:2625-2631.

Montgomery, D.C., 1984. Design and Analysis of Experiments. John Wiley and Sons, New York, New York.

Montgomery, D.R. and L.H. MacDonald, 2002. Diagnostic Approach to Stream Channel Assessment and Monitoring. Journal of the American Water Resources Association (JAWRA) 38(1):1-16.

Parker, G. and P.C. Klingeman, 1982. On Why Gravel Bed Streams Are Paved. Water Resources Research 18:1409-1423.

Platts, W.S., W.F. Megahan, and G.W. Minshall, 1983. Methods for Evaluating Stream, Riparian, and Biotic Conditions. U.S. Department of Agriculture, Forest Service, General Technical Report INT-138, Ogden Utah, 70 pp.

Potyondy, J.P. and T. Hardy, 1994. Use of Pebble Counts to Evaluate Fine Sediment Increase in Stream Channels. Water Resources Bulletin 30(3):509-520.

Rice, S. and M. Church, 1996. Sampling Surfical Fluvial Gravels: The Precision of Size Distribution Percentile Estimates. Journal of Sedimentary Research 66:654-665.

Roper, B.B., J.L. 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 (JAWRA) 38(6):1637-1646.

Roper, B.B., J.L. Kershner, and R.C. Henderson, 2003. The Value of Using Permanent Sites When Evaluating Steam Attributes at the Reach Scale. Journal of Freshwater Ecology 18:585-592.

SAS Institute, 2001. SAS Version 8.2. SAS Institute, Inc., Cary, North Carolina.

Schnackenberg, E.S. and L.H. MacDonald, 1998. Detecting Cumulative Effects on the Routt National Forest, Colorado. Journal of the American Water Resources Association (JAWRA) 34(5):1163-1177.

Sokal, R.R. and J. Rohlf, 1995. Biometry: The Principles and Practice of Statistics in Biological Research. W.H. Freeman and Company, New York, New York.

Urquhart, N.S., S.G. Paulsen, and D.P. Larsen, 1998. Monitoring for Policy-Relevant Regional Trends Over Time. Ecological Applications 8:246-257.

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 32:3219-3226.

Wolcott, J. and M. Church, 1991. Strategies for Sampling Spatially Heterogeneous Phenomena: The Example of River Gravels. Journal of Sediment. Petrol. 61:534-543.

Wolman, M.G., 1954. A Method of Sampling Coarse River-Bed Material. Transactions of the American Geophysical Union 35(6):951-956.

Yang, C.T. 2003. Sediment Transport; Theory and Practice. Krieger Publishing Company, Malabar, Florida, 396 pp.

Zar, J.H., 1996. Biological Analysis (Third Edition). Prentice Hall, Englewood Cliffs, New Jersey, 662 pp.