Pennsylvania DEP Multihabitat Stream Assessment Protocol

March 2007

Pennsylvania DEP Multihabitat Stream Assessment Protocol

March 2007

The United States Environmental Protection Agency's Rapid Bioassessment Protocols for use in Wadeable Streams and Rivers (Barbour et al.1999) describes two general approaches to assessing stream macroinvertebrate communities. These approaches are the "single, most productive habitat" approach and the "multihabitat" approach. The single, most productive habitat approach is typically used to assess streams where cobble substrate (riffle/run) is the predominant habitat. The multihabitat approach involves sampling a variety of habitat types instead of sampling a single habitat, such as cobble substrate in riffles and/or runs.

In April of 2002, the Pennsylvania DEP began developing a macroinvertebrate bioassessment protocol for assessing the Commonwealth's low-gradient streams. Low-gradient waterways consist of pool/glide channel morphology and naturally lack riffles. The multihabitat field and laboratory methods described in Barbour et al (1999) were used as a starting point for the project. Water chemistry, physical habitat, and aquatic macroinvertebrates were collected at 77 sampling sites in this study. The project goal was to identify practical and regionally appropriate field, laboratory, and data analysis procedures and to develop an index of biological integrity that accurately reflects the ecological conditions of Pennsylvania's low-gradient streams.

Reference and Stressed Sites

The abiotic conditions of all sample sites were analyzed to determine if the sites should be divided into different bioregions. None of the abiotic conditions investigated provided justification for dividing the sites into different bioregions, therefore, all sample sites will be held to the same criterion when determining if they are reference, non-reference, or stressed. Appendix A-1 contains a map of the sample sites. The 77 sample sites were categorized as reference, non-reference, or stressed based on 15 parameters. All 15 parameters were used as reference site criteria. For sites to be considered reference sites, they had to meet the criteria values of all 15 parameters (Appendix A-2). The first 14 parameters in Appendix A-2 were used to determine if the site was stressed. Sites were considered stressed if they failed any one of the 14 stressed criteria values. For example, if a site had a value of 4.8 mg/l for dissolved oxygen, the site would be considered stressed regardless of other parameter values.

Field Methods

All chemical water quality, physical habitat, and aquatic macroinvertebrate data is collected from a sample reach approximately 100 meters in length. During development of the protocol, water temperature, pH, dissolved oxygen, and conductivity were measured in the field and a chemical sample was collected from each reach for laboratory analysis. This sample was collected under base flow (non-stormwater runoff) conditions.

Field	Lab	
Temperature	pH	Total Organic Carbon
Dissolved Oxygen	Alkalinity	Chloride
pH	Nitrate-N	Sulfate
Conductivity	Total Phosphorus	Iron

Total phosphorus and total organic carbon samples are preserved with 10% sulfuric acid and samples analyzed for metals are preserved with concentrated nitric acid to a pH <2. All samples are kept on ice and delivered to the DEP laboratory in Harrisburg, PA within 48 hours of collection.

Physical habitat is documented using the EPA Glide/Pool Prevalence Habitat Assessment Field Data Sheet (Barbour et al. 1999). This evaluation divides the habitat of the stream and its adjacent land use into ten parameters. Each parameter is scored on a scale of 0 to 20, with a higher score indicating better conditions. Depending on the score, a parameter can fall into one of four categories: Poor, Marginal, Suboptimal, and Optimal.

For the purpose of this protocol, only nine of the ten parameters are used. Channel Sinuosity (indicated as Habitat Parameter 7 in Appendix B-1) is not used because the range of sinuosity as defined in the data sheet is not applicable to Pennsylvania streams. Even the State's most sinuous streams will have low values using this definition. Thus, total habitat site scores can range from 0-180, with 180 being a perfect score (Appendix B-1).

The majority of macroinvertebrate samples were collected from October to May. A small number of samples were collected outside of this period to test the seasonal variability of the protocol. Seasonal variability analysis results are discussed on page 6 and 7.

Aquatic macroinvertebrate samples are collected using a multihabitat sample collection method modified from that described in Barbour et al (1999). Organisms are collected from five different habitat types within the sample reach. The habitat types and explanations of sampling techniques are described in Appendix B-2. A total of 10 "jabs" are collected within each sample reach. Each jab consists of a 30-inch-long sweep of a 0.3-meter wide area, using a D-frame dip net (500 micron mesh). At least two jabs are made in each of the habitat types present within the sample reach.

The biologist first identifies which habitat types are present within the sample reach. A minimum surface area of approximately 0.46 m^2 is required for a given habitat type to be sampled. If the total number of jabs (10) is not evenly divisible by the number of habitat types present, the remaining jab(s) are distributed among the most extensive habitat type(s) in the reach. All jabs are combined into several 2-liter largemouth jars and preserved in ethyl alcohol. Typically, the combined 10 jabs will fill three to four 2-liter sample jars about 2/3 full with organic and inorganic material. Sample jars are topped-off with 95% ethanol to ensure adequate sample preservation.

Lab Methods

In the laboratory, each composited sample is placed into a 3.5" deep rectangular pan (measuring 14" long x 8" wide on the bottom of the pan) marked off into 28 four-square inch (2" x 2") grids. Using an illuminated magnifying lens, macroinvertebrates are picked from a minimum of four grids, selected at random, to generate a 200-organism (+/- 20%) sub-sample. Additional grids may be selected at random until the sub-sample is obtained. The organisms contained in the 200-organism sub-sample are identified to the lowest practical taxonomic level (usually genus). Some individuals collected will be immature and not exhibit the characteristics necessary for confident identification. If the individual cannot be confidently identified to the proper level, it should be discarded. All pupae are discarded. Certain groups are identified to a higher taxonomic level as follows:

Flatworms (Turbellaria) – Phylum Turbellaria
Segmented worms (Annelida), aquatic earthworms, & tubificids – Class Oligochaeta
Proboscis worms – Phylum Nemertea
Roundworms – Phylum Nematoda
Water mites – "Hydracarina" (an artificial taxonomic grouping of several mite superfamilies)
Midges – Family Chironimadae
Weevils – Family Curculionidae
Sand flies\no-see-ums – Ceratopogonidae
Decapoda, Gastropoda, and Pelecypoda are identified to family

A detailed explanation of the laboratory processing procedure is provided in Appendix C. Pollution tolerance values and functional feeding group information are listed in Appendix D.

Metrics Selection

The 200-organism sub-sample data, from 77 samples, was used to calculate values and produce box plots for an initial fifty metrics. Only "truly-aquatic" (hydropneustic) organisms included in the 200-organism sub-samples were used to generate these metric scores. By visually comparing box plots of all fifty metrics and choosing those that could discriminate between minimally disturbed reference and stressed sites, thirteen candidate metrics were selected. An explanation on interpreting box plots can be found in EPA's RBP manual (Barbour et al. 1999).

The discrimination efficiency (D.E.) of each candidate metric was calculated to better determine how well the metric could distinguish between a reference and stressed site. These values are listed in Table 1 below. The D.E. is the percentage of stressed samples whose scores do not overlap with the interquartile range of reference sample scores. The 25^{th} percentile of the total number of reference samples was used as the threshold for metrics that decrease with pollution. For these metrics, the following formula was used:

D.E. = (the # of stressed samples that fall below the 25th percentile value of the reference distribution / the total # of stressed samples) x 100

The 75th percentile of the total number of reference samples was used as the threshold for metrics that increase with pollution. For these metrics, the following formula was used:

D.E. = (the # of stressed samples that occur above the 75th percentile value of the reference distribution / the total # of stressed samples) x 100

Box plots depicting these two scenarios can be found in Appendix E-1. Those metrics with a D.E. less then 80 were eliminated because of their weak ability to discriminate. Trophic Diversity, % Tolerant Taxa, and % Intolerant Taxa (Hils<5) all had D.E.'s of 76 and were therefore dropped, leaving ten metrics.

Table 1. Discrimination Efficiencies of the Thirteen Candidate Metrics

Candidate Metrics	Discrimination Efficiency (D.E.)
EPT	100
Taxa Richness	94
# Of Caddisfly Taxa	94
# Intolerant Taxa (Hils<5)	94
# Of Mayfly Taxa	88
Shannon Diversity	88
Beck4	82
Beck3	82
% Taxa as EPT	82
% EPT	82
Trophic Diversity	76
% Tolerant Taxa	76
% Intolerant Taxa (Hils<5)	76

To eliminate redundant metrics that might measure similar attributes, Pearson correlation coefficients were calculated (Appendix E-2). If two metrics were highly correlated (r^2 >0.90) the most familiar, easiest to interpret, and/or higher D.E. metric was retained. This process eliminated two metrics: Beck3 and Number Intolerant Taxa (Hilsenhoff < 5). Beck3 was highly correlated (r^2=0.93) with the Beck4 metric. Beck4 had larger values and a tighter reference distribution and therefore was kept. Number Intolerant Taxa (Hilsenhoff < 5) was highly correlated with EPT (r^2=0.91); it had the lower D.E. and consequently was dropped.

Percent EPT was then eliminated to avoid having three EPT metrics; this would have created a heavy reliance on those taxa. Percent Taxa as EPT was found to produce high metric scores for streams that should be impaired because of low pH values. This can result from the inclusion of low pH tolerant stoneflies in the metric calculation. To

prevent the inapt assignment of attainment status to low pH streams, this metric was eliminated.
The remaining six metrics are the core metrics used to calculate the Total Biological Scores for this protocol.

EPT	Beck4
Taxa Richness	# Mayfly Taxa
Shannon Diversity	# Caddisfly Taxa

They are listed and explained in Appendix E-3. Box plots of the raw values for each metric are located in Appendix E-4.

Normalization of Metric Scores and Total Biological Score Calculation

All six core metrics decrease with increasing stress, and therefore were normalized to a scale of 0 to 100 based on the 95th percentile value (least squares estimate) of all samples (n = 77) using the following equation:

$$\text{Normalized Metric Score} = (\text{Observed Value} / 95^{th}\ \text{percentile}) \times 100$$

An example of how to calculate metric scores (observed value) and the Total Biological Score of two samples is shown in Appendix F.

Aquatic Life Use Benchmarks

Aquatic life use attainment status of a given sample reach is determined by comparing its Total Biological Score to a use attainment benchmark. If the Total Biological Score of the sample reach is less than the benchmark score, the sample reach is not attaining for aquatic life.

The 10th percentile of the Total Biological Scores of the reference site dataset (n=16) was used to set the aquatic life use benchmark. Appendix G supports using the 10th percentile value by showing the well defined separation of the Total Biological Scores of the reference and stressed sites.

Table 2. Aquatic Life Use (ALU) Benchmark

Multihabitat ALU Benchmark
55 (10th percentile)

Sites with Total Biological Scores scoring above the benchmark are attaining (Saw Creek, Appendix F) and sites with Total Biological Scores scoring below the benchmark are considered impaired for ALU (Wiconisco Creek, Appendix F).

Protocol Verification

The aquatic life use status (reference or stressed) of eighteen low gradient streams was predicted using the chemistry, habitat, and land use criteria listed in Appendix A-2. Ten of the streams were considered impaired and eight attaining, based solely on the abiotic conditions. Macroinvertebrate verification samples were then collected at those eighteen streams to test the accuracy of the field/lab methods and the reliability of the benchmark. The verification samples were collected between April 12th 2006 and May 31st 2006, using the same field/lab procedures described in Appendixes B and C. The Total Biological Scores for all 18 samples were calculated and the aquatic life use attainment status determined using the benchmark set in this protocol. Nine of the ten stressed sites were found to be impaired using the protocol benchmarks. Seven of the eight reference sites had Total Biological Scores exceeding the benchmark. Appendix H-1 lists the metric values and Total Biological Scores of the verification samples. An unnamed Tributary to South Branch Muddy Creek was the only reference sample that did not meet its predicted attainment status. Using this protocol, it had a Total Biological Score of 44, missing attainment status by 11 points. This resulted from the inclusion of a high number of stoneflies in the sub-sample. Randomly selecting more stoneflies would prevent the inclusion of other species in the sub-sample and therefore lower the metric score for Taxa Richness, # Of Mayfly Taxa, and # Of Caddisfly Taxa. This tributary is located in state forest and the macroinvertebrate list otherwise indicates attainment. Kitchen Run was the only stressed stream reach whose verification sample scored above the benchmark. The benchmark was only exceeded by two points. The top three genera in the sub-sample were Simulium, Prosimulium, and Chironomidae, making up 70% of the sub-sample. Four different Ephemeroptera taxa were identified, however, three of the genera contained only one organism. This would inflate the metric scores of EPT, Taxa Richness, and # of Mayfly Taxa. Also, eight of the eighteen taxa identified contained only one organism. This could mask the fact that the sample was dominated by pollution tolerant species.

Overall the benchmark was 88% affective at identifying ALU attainment and 90% affective in determining ALU impairment. These percentages are very high, indicating the benchmark is accurate in determining the Aquatic Life Use of a sample reach. Appendix H-2 contains box plots of the verification samples verses the reference and stressed sites. These eighteen samples verify the methodology described in this protocol and justify the placement of the aquatic life use benchmark.

Method and Annual Variability

Between April 23rd and May 30th 2003, aquatic macroinvertebrate samples were collected from adjacent stream reaches on three different streams. These paired-samples were used to document method variability. The standard deviation of the Total Biological Scores, calculated as the root mean squared error in an ANOVA, was 10.9. The standard deviation indicates the average variation of the Total Biological Scores in a paired sample. A standard deviation of zero would indicate the sample pairs received the same Total Biological Score. The 90% confidence interval calculated from the standard deviations was 14 for one sample and 9.7 if two samples are collected. This is relatively

high variability but it may be an overestimate because it was based upon only three paired comparisons. As a rule, variability measures decline as the sample size increases. The annual variability discussed below also indicates this standard deviation based on the three pairs may be an overestimate.

A similar analysis was conducted using paired-sample data collected from four sample reaches during October-May. Two of these reaches were re-sampled one year later, and the remaining reaches were re-sampled two years after the initial data collection effort. The standard deviation (calculated in the same manner described above) of the four sample pairs was used to document long-term variability. The standard deviation of the annual pairs was 6.6 indicating less variability than the paired samples. The 90% confidence interval calculated from the standard deviations was 8.1 for one sample and 5.9 if two samples are collected. This is a more acceptable range of variability.
The 3 paired and 4 annual samples all had scores above the attainment benchmark no matter which repeated sample was used in the comparison. This is an indication that at least in this instance the variability was not great enough to effect the attainment/impairment decisions. It would have been a concern if one repeated sample showed attainment and the other impairment creating a lack of consistency. The success of the verification effort is another indication the variability is not creating inconsistencies in attainment/impairment decisions. The Department will continue to refine the variability estimates with additional surveys in spring 2007. The variability results are summarized in Appendix H-3.

Conclusion

As stated earlier, the project goal was to apply practical and regionally appropriate field, laboratory, and data analysis procedures to the development of an index of biological integrity that accurately reflects the ecological conditions of Pennsylvania's low-gradient streams. Seventy-seven samples collected statewide from low gradient streams, between October and May, were used in developing this protocol. Data analyses did not show any natural differences between the statewide sites that would justify creating separate assessment categories. Therefore, all sites were held to the same criteria when discriminating between reference and stressed sites.

The method used to collect macroinvertebrate samples is modified from the steps described in the EPA document Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers (Barbour et al 1999). Using a D-frame dip net, ten jabs were distributed between five possible habitat types in each sample reach. The jabs were combined and taken to the laboratory for macroinvertebrate identification. A 200-organism +/-20% sub-sample was identified to the genus level or to the lowest confident taxonomic level.

Six core metrics were chosen from an initial list of fifty metrics, based on how well the metric could distinguish between reference and stressed sites. The resulting six metrics are:

EPT	Beck4	# Of Caddisfly Taxa
Taxa Richness	Shannon Diversity	# Of Mayfly Taxa

Metric scores were then normalized and summed for each sample to produce a Total Biological Score. By visually comparing box plots of the Total Biological Scores of the reference and stressed sites, the 10^{th} percentile value (55) of the reference sites was chosen as the aquatic life use benchmark. This value has an extremely high D.E. of 94. The placement of the benchmark was confirmed by the success of the verification and variability analyses. Although the intra site variability was high, the annual variability was low indicating the protocol can be successfully repeated for low gradient streams. This benchmark of 55 is used as the threshold in determining aquatic life use attainment status for low gradient streams.