Volunteer Biological Monitoring: Can It Accurately Assess the Ecological Condition of Streams?

Sarah R. Engel and J. Reese Voshell, Jr.

Abstract Government agencies have begun to use biological monitoring data collected by volunteers for official purposes, but questions have been raised regarding the validity of conclusions about ecological condition. We conducted a 2-yr study that assessed, modified, and validated the Virginia Save-Our-Streams (SOS) program, a popular volunteer monitoring program that emphasizes benthic macroinvertebrates. The study design consisted of sampling sites using accepted professional methods concurrently with volunteers using the SOS protocol. In addition, sites previously sampled by volunteers were re-sampled using professional methods. The numerical results from volunteer and professional samples were not correlated (r = 0.46) and at times produced different conclusions about ecological condition (65% agreement). The Virginia SOS protocol consistently overrated ecological condition. We determined that the reason for the inaccuracy was the simplistic numerical analysis in the volunteer protocol, which was based solely on presence of taxa. We developed a quantitative multimetric index that was appropriate for use by volunteers, and the SOS sampling protocol was modified to obtain counts of macroinvertebrates in the various taxa. The modified SOS protocol was then evaluated with a different set of concurrent samples taken by volunteers and professionals. The modified SOS protocol proved feasible for volunteers, and the new multimetric index correlated well with a professional multimetric index (r = 0.6923). The conclusions about ecological condition reached by the volunteer and professional protocols agreed very closely (96%). This study demonstrated that volunteer biological monitoring programs can provide reliable information about ecological condition, but every protocol needs to be validated by standard quantitative methods.

Volunteers have long been a vital source of labor in this country and they continue to make valuable contributions to the workforce. According to a Gallup poll conducted by Independent Sector (1999), 56% of the nation's citizens volunteered for an average of 3.5 h/wk during 1998. The volunteers' work had an estimated value of $225.9 billion in 1998 (Independent Sector 1999). Of these volunteers, 5.5% reported volunteering in the environmental realm during 1998 (Independent Sector 1999).

One way in which volunteers can contribute in the environmental domain is to participate in activities that monitor the ecological condition, sometimes called the "health," of aquatic environments. Aquatic environmental monitoring can involve measuring various physical or chemical factors, such as sediment accumulation on the stream bed or concentration of dissolved oxygen. Physical and chemical factors are often referred to collectively as water quality. In recent years, federal and state regulatory agencies have emphasized the need for biological monitoring of aquatic environments, in addition to water quality monitoring (Miller et al. 1988). Biological monitoring (also called biomonitoring or bioassessment) is defined as an evaluation of the condition of a water body using biological surveys and other direct measurements of the resident biota in surface waters (Matthews et al. 1982, Rosenberg and Resh 1993, Gibson et al. 1996). Monitoring is usually thought of as taking repeated measurements to keep track of something, whereas an assessment can be a one-time measurement.

Biological monitoring can be done with any living organisms, but benthic macroinvertebrate, fish, and periphyton (algal) assemblages are used most often, in that order. Benthic macroinvertebrates are those organisms that live on the bottom of aquatic environments, or on objects protruding above the bottom, and are large enough to see by eye without any magnification. Periphyton refers to the algae that live attached to firm substrates. Although complete studies may include all three assemblages, benthic macroinvertebrates are used most often for several reasons. First, benthic macroinvertebrates do not migrate very far, thereby ensuring exposure to a pollutant or stress reliably conveys local conditions. This reliable representation of ecological condition allows for comparison of sites that are in close proximity. Second, macroinvertebrate life stages are short enough that sensitive life stages will be affected by a stress, but long enough that any impairment is measurable in the assemblage. Benthic macroinvertebrates are found in even the smallest streams and have a wide range of sensitivity to all types of pollution and stress, allowing for monitoring in most conditions. Finally, sampling benthic macroinvertebrates is easy, cost effective, and does not permanently harm the local assemblage. Impairment can easily be detected by the trained monitor with even the simplest of identifications (Plafkin et al. 1989, Voshell et al. 1997).

Unfortunately, bioassessment with benthic macroinvertebrates can be very costly and time-consuming (Resh and Jackson 1993). Samples often contain a lot of sand, silt, and plant debris (leaves, roots, twigs, fine detritus) from which the organisms must be manually sorted. After sorting (also called "bug picking"), the numerous organisms must be identified and counted by means of a dissecting microscope. The identifications are made more difficult by the fact that most of the organisms are immature stages of aquatic insects, which are often very small and hard to identify. In the late 1980s, the U.S. Environmental Protection Agency (U.S. EPA) introduced Rapid Bioassessment Protocols to streamline the process, while retaining scientific accuracy, and these have been widely adopted by state regulatory agencies (Lenat and Barbour 1993, Barbour 1997, Barbour et al. 1999). However, most state environmental regulatory agencies have thousands of miles of streams with hundreds of sites to monitor, and only a few professional aquatic biologists to do the work that is mandated by the U.S. EPA. Although rapid assessment techniques allow aquatic biologists to accomplish more monitoring than they otherwise could, thousands of miles of streams remain unmonitored because of limited resources.

Volunteers have recently organized and stepped forward to help fill the sampling gap. There are many volunteer programs in place across the country that are thought to be successful at collecting data at lower costs than professional surveys (Thomson 1987, Maas et al. 1991, Markusic 1991, Levy 1998). It has been assumed that, with proper training and adequate quality assurance/quality control plans, volunteers can collect quality data suitable for making regulatory decisions (Lee 1988, Mattson et al. 1994, Lathrop and Markowitz 1995, Sheehan 1998). An added benefit of using volunteers in a assessment program is the ability of a group of volunteers to sample multiple locations at one time (Maas et al. 1991), while professional aquatic biologists, who usually work alone or with one assistant, can only sample one site. In addition, volunteers often monitor the waters where they live or go for recreation, so they can watch for changing conditions and report them in a timely fashion (Livermore 1993). Professionals must monitor a large number of widely distributed sites, so they may only be able to visit a site once every few years, which greatly limits their ability to detect short-term changes in ecological condition.

There are many volunteer monitoring programs for streams across the United States (Cooke 1999). The Save-Our-Stream (SOS) program, administered by the Izaak Walton League of America, is one of the oldest and most popular volunteer biological monitoring programs. The SOS program, which involves many volunteers in the mid-Atlantic states, is representative of other programs and was the subject of the study that we report here on the effectiveness of volunteer monitoring in Virginia streams.

The U.S. EPA has decided that data from volunteers can and should be used in reports that are required from states on current environmental condition of water bodies, namely the 305(b) report (Lathrop and Markowitz 1995) and 303(d) list (U.S. Environmental Protection Agency 1999). The 305(b) reports the condition of waterways within a state (Barbour et al. 1999). The 303(d) list is an annual listing of impaired and threatened waterways in a state (U.S. Environmental Protection Agency 1999). For the 303(d) list, the U.S. EPA (1999) is very clear about the use of biological monitoring data from volunteers, requiring the use of all "existing and readily available data," including "data, information, and water quality problems reported by ... members of the public". When streams are listed as impaired in these reports, it triggers major action in the form of a total maximum daily load plan. A total maximum daily load is a plan of action to return the stream to a pre-impacted condition and get it removed from the impairment list. These plans require much effort, both in their design and implementation, and there are specific timelines for getting streams off the list. Therefore, it is important that the ecological condition of streams be accurately assessed.

As volunteer data are increasingly incorporated into important regulatory decisions that have far-reaching ramifications, concerns are being raised over the validity of using data from volunteers. Some of the primary reasons for concern are the level to which volunteers identify macroinvertebrates, the limitations of their collecting techniques, and the level of training the volunteers receive (Penrose and Call 1995). A 1993 conference, widely attended by representatives of universities, federal, state, and local governments, businesses, and volunteer groups, listed and ranked the barriers to volunteer biological monitoring efforts. Data concerns, including credibility, standardization, and quality assurance, were voted to be the top barrier, garnering 39.5% of the ranking vote (Godfrey 1994). Though volunteers in Virginia must go through a training and certification process, it is speculated that this process might not be rigorous enough to assure data of high enough quality for use in management decisions (Jay Gilliam, Coordinator of Virginia SOS, personal communication). Obviously, these concerns must be addressed before monitoring data from volunteers can be used with confidence.

To date, there have only been a few cursory studies comparing the results of volunteer biological monitoring to professional monitoring. In Connecticut, professionals resample one volunteer site per year in an effort to continually assess the quality of the volunteers' efforts (Ely 1997). After the initial assessment in 1992, the volunteer protocol was modified. The result was that in subsequent years there was more similarity in the decisions reached by volunteers and professionals (Ely 1997). A study in Washington found an excellent comparison between volunteers and professionals who were using the same sample collection methods (Ely 2000). Volunteer biological monitoring has not fared as well in other comparative studies. In North Carolina, previously untrained volunteers were able to identify higher quality streams, but were unable to differentiate the lower quality streams (Penrose and Call 1995). Sampling in Ohio indicated that volunteers were able to determine if streams were attaining their designated use category, but had a tendency to overrate water quality when compared to professionals sampling with the same methods (Dilley 1991). DeWalt (1999) also found that volunteers tended to overrate the environmental condition of Illinois streams.

Because of the disparity in the conclusions mentioned above and the importance of this issue in the environmental regulation of freshwater natural resources, we set out to conduct a thorough investigation of the effectiveness of volunteer biological monitoring with benthic macroinvertebrates in streams. The Virginia SOS Program was the subject of our study. Our objectives were as follows: (1) to compare the biological condition assessments made by volunteers in the Save-Our-Streams Program to those made by professionals, (2) to recommend modifications to improve the volunteer method should it not compare favorably with professional protocols, and (3) to validate a modified volunteer protocol, should one be necessary.

Materials and Methods

Study Design. We used two lines of investigation in our analysis of volunteer biological monitoring. The first, and primary one, involved concurrent sampling with certified Virginia SOS volunteers at 23 sites. All of these sites were on streams in the western part of Virginia. Most were in the Ridges and Valleys or Blue Ridge regions, but a few extended into the Piedmont and Coastal Plain. At all sites there was sufficient gradient for the bottom composition to be mostly rocky. Sites were on first to fifth order streams, and all sites were shallow enough to be waded. The concurrent sampling took place during the summer and fall of 1998. The 23 sites and volunteer groups were selected based on availability of the volunteers, proximity of sites to volunteers, and recommendation by Jay Gilliam, Coordinator of the Virginia SOS Program. We tried to ensure that the sites for concurrent sampling included a wide range of expected ecological conditions, from sites with little apparent human impact and best attainable ecological conditions, to sites where human activities were obviously causing changes in the streams and ecological conditions were likely to be impaired. This selection was done by reviewing previous SOS studies at each site, and visually observing each site and selecting those that would provide an overall mix of conditions.

At each concurrent sampling site, the volunteer group made an independent assessment according to the standard Virginia SOS protocol. The only interaction with us was an explanation of the study. We told them to proceed as they normally do. We tried to put them at ease by telling them that their performance would not be identified individually in the results, they were not being graded, and we were only evaluating the volunteer protocol, not individual volunteers. We observed the sampling techniques of the volunteers and recorded our observations. After they made their assessment we took possession of all volunteer samples and preserved them so that we could check their identifications. In addition, the debris left on the kick net, which they had sorted through and from which they had removed all necessary organisms, was retained and preserved to check for completeness of sample sorting by volunteers. Lastly, we took what we refer to as a professional sample in an undisturbed location at the same site to see if the volunteer monitoring reached the same conclusions as that done by professional aquatic biologists.

Our second line of investigation involved sampling at 122 sites where Virginia SOS volunteers had taken samples and made assessments of ecological conditions during the past 5 yr. We refer to these as historical samples. Data sheets for volunteer biological monitoring were compiled from Virginia SOS records, and the sites were located on 7.5-min USGS topographic quadrangles to the nearest 15 s. We visited those sites and took professional samples in the summer and fall of 1998. We recognized that there were weaknesses in this approach. There could have been changes in the ecological condition of the sites in the intervening time, either improvement or degradation. There was no opportunity to observe the volunteers taking their samples or to examine the contents of their sample. However, we were able to eliminate those sites that had obviously changed based on the habitat information the volunteers had recorded along with their macroinvertebrate data, so we decided that the large amount of available Virginia SOS data was potentially very useful. We therefore sampled those sites and considered the historical samples as supporting evidence.

Virginia Save-Our-Streams Protocol for Sampling and Data Analysis. The Virginia SOS sampling protocol consists of three individual kick net samples. The kick net measures 1 by 1 m, and the size of the mesh openings is approximately 1,500 μm. The net is held in a riffle by one volunteer and approximately 1 m^2 of substrate is sampled in front of the net by at least one other volunteer. Rocks are moved and rubbed on all sides by hand to remove all attached organisms, then the substrate is thoroughly disturbed by hand or rake. The net is returned to shore and spread on a sheet or board to catch organisms that crawl through the mesh. The macroinvertebrates in each individual sample are removed, or "picked," from the net and sorted by taxa into separate containers. The organisms are then identified in the field based on previous training and simple pictures (Kellogg 1994). The taxonomic level is mostly to order, with a few selected families. The presence of each taxonomic group is recorded, then the protocol dictates returning the organisms to the stream. The volunteers also complete a simple habitat assessment.

Data from each macroinvertebrate sample was used to calculate the SOS water quality rating score—a simple biotic index based on presence/ absence data. The macroinvertebrate taxa are divided into three sensitivity categories based on their tolerance of poor water conditions (sensitive, somewhat sensitive, and tolerant). The SOS water quality rating is calculated by multiplying the number of taxa present in each sensitivity category by a numerical tolerance value (sensitive = 3, somewhat sensitive = 2, tolerant = 1). The resulting numbers are added, then used to determine water quality (excellent >22, good = 17-22, fair = 11-16, poor <11). Each of the three kick-net samples is scored individually, and the sample with the highest score is considered to be the most accurate indication of the site's ecological condition.

Virginia SOS volunteers must be certified before contributing data to the program. The certification process includes lectures and hands-on practice sessions, followed by a test. The test includes sampling, which is observed and evaluated by a trainer, and identification of macroinvertebrate specimens, in which volunteers must score at least 84% to pass. Once the test has been passed, volunteers are then certified to monitor indefinitely.

Professional Protocol for Sampling and Data Analysis. The professional sampling method that we used was in accordance with the latest guidance from the U.S. EPA for rapid assessment protocols (Barbour et al. 1999). Our sample at each site consisted of a composite of four D-frame dip net subsamples. The dip net was 0.30 m long on its bottom flat side, and the mesh openings of the net were 500 μm. Two D-frame subsamples were collected from fast current (greater than ≈30 cm/s) in predominately cobble substrate, and two were collected from slow current (less than ≈30 cm/s) in predominately pebble substrate. For each individual subsample, we held the D-frame dip net in one location and disturbed the substrate immediately upstream of the net for 15 s in a square area equal to the width of the net frame (≈ 0.1 m^2). These samples were preserved in 95% ethanol for later analysis in the laboratory. While at each site, a habitat assessment was conducted according to the recommendations of the U.S. EPA's rapid assessment protocols (Barbour et al. 1999). The habitat assessment was solely used to compare to the volunteer habitat assessments to determine suitability of each site.

In the laboratory, all macroinvertebrates were sorted from the debris, identified to the genus level, and counted. The purpose of the professional samples was to make a statistical comparison between the results of the volunteer samples and the results of the professional samples at the same sites. The only numerical value that volunteers calculated for their samples was the SOS water quality rating score. For comparison with volunteer samples, we calculated one individual biotic index and one multimetric index from the professional sample data. The biotic index for the professional samples was the modified Hilsenhoff biotic index (HBI), calculated by the following equation (Hilsenhoff 1987):

$$= \Sigma (x_i t_i / n),$$

where, x_i = the number of individuals in the ith taxon, t_i = tolerance score of the ith taxon, and n = total abundance. Most of the fauna was identified to the genus level for the HBI. Modified refers to the tolerance values being adjusted for the fauna in Virginia, because the HBI was originally developed for Wisconsin streams.

The multimetric index that we calculated for the professional sample data was developed for streams in the mid-Atlantic highlands (Smith and Voshell 1997, Voshell et al. 1997). The macroinvertebrate aggregated index for streams (MAIS) was developed according to the framework proposed by Barbour et al. (1995) and Barbour et al. (1996). The MAIS score is calculated from the values of nine individual metrics, all of which are based on family-level identifications: % 5 dominant taxa, modified Hilsenhoff biotic index, % haptobenthos (those organisms needing clean, firm substrate), EPT index (Ephemeroptera, Plecoptera, and Trichoptera), #

Ephemeroptera taxa, % Ephemeroptera, Simpson diversity index, # intolerant taxa, and % scrapers. We used Microsoft Excel 97 SR-2 for storing data and calculating the HBI and MAIS.

Statistical Analyses. For the concurrent samples, the initial statistical analysis consisted of comparing the results of the volunteer samples with the results of the professional samples at the same sites, assuming that the professional samples yielded the correct results. We used Pearson product-moment correlation analysis (Sall and Lehman 1996) to compare the volunteer SOS water quality rating score to two numerical values calculated from the professional samples (MAIS and modified HBI scores). We determined *a priori* that the *r*-value should be ≥0.70 for the volunteer and professional results to be considered correlated. This criterion was based on the coefficient of determination (r^2), which is the amount of variation in the data that is explained by their correlation. For biological field studies, 50% of the variation should be explained, which equates to an *r*-value of 0.70 (Sokal and Rohlf 1969, Moore and McCabe 1993, Ramsey and Schafer 1997). We also tested how well a line fit the data to determine if the correlation was significant ($\propto = 0.05$). There was a potential problem that the two sets of results could be highly correlated, but the volunteer results could still lead to the wrong conclusion about ecological condition. To resolve that question, we used classification analysis (Sall and Lehman 1996) to compare the ecological condition category (acceptable or unacceptable conditions) determined by the SOS water quality rating score to that determined by the MAIS score. Both the SOS water quality rating score and the MAIS score place streams in one of four ecological condition categories. For the SOS water quality rating score, these are excellent, good, fair, and poor, whereas, for the MAIS the categories are designated very good, good, poor, and very poor. In both categorization schemes, the upper two categories are considered to represent acceptable ecological conditions and the lower two unacceptable conditions. We made an *a priori* decision that volunteer assessments of ecological condition (acceptable, unacceptable) should agree with professional assessments at 86% of the sites to be considered satisfactory. This criterion was determined from a chi-squared test in which the numbers of acceptable versus unacceptable determinations of ecological condition from the professional samples were the expected frequency and the numbers of acceptable versus unacceptable determinations of ecological condition from the volunteer samples were the observed frequency ($\propto = 0.05$). Volunteer determinations must agree with professional determinations at 20 of the 23 sites (86%) in order for the observed frequency to be not significantly different from the expected frequency. McNemar's test ($\propto = 0.05$) was used to determine if the proportion of sites showing agreement in ecological condition between SOS water quality rating score and the MAIS score was different from that showing disagreement (Stokes et al. 1995).

If volunteers made different assessments of ecological condition, it would be necessary to determine if those differences were caused by the SOS water quality rating score *per se* or the volunteer protocol for acquiring the data (sampling, identifying, counting). To answer this question, we calculated the SOS water quality rating score using data from the professional samples. Values for the SOS water-quality rating scores for the two types of samples (volunteer and professional) were compared with a paired *t*-test (Sall and Lehman 1996). The SOS water quality rating scores calculated from the professional samples were compared to the professional MAIS and modified HBI scores by the same statistical techniques used for the original volunteer SOS water quality rating scores (correlation, classification analyses, and McNemar's test).

It was possible to do additional statistical analysis for the concurrent samples because the volunteer samples were preserved and retained along with the professional samples. We re-identified and counted the organisms in the sample that the volunteers removed from the net, which we referred to as the actual volunteer sample. In addition, we identified and counted any organisms that the volunteers left undiscovered on the net and combined the results with the actual volunteer sample. We referred to this corrected sample as the potential volunteer sample. In order to determine if samples collected by the Virginia SOS volunteers were adequate, we calculated MAIS and modified HBI scores from the actual and potential volunteer samples and used correlation analysis to compare them individually to MAIS and modified HBI scores from the professional samples. A correlation, classification analysis, and McNemar's test of the MAIS and modified HBI scores from the actual volunteer samples to potential volunteer samples was also completed in order to determine if volunteers were collecting enough organisms to make an accurate assessment of ecological condition. The final evaluations of the concurrent samples were paired *t*-tests comparing the actual and potential volunteer samples to each other and to the equivalent professional sample scores.

Similar analyses were done for the historical samples, except that there were not as many possibilities without having the original samples or raw data. Our analysis of the historical data included a correlation, classification analysis, and McNemar's test to compare the volunteer water quality rating score to the professional MAIS score. We also calculated an SOS water quality rating score using the professional data, which we compared to the volunteer water quality rating score. Having only presence/ absence data from volunteer original samples prohibited further analyses.

Results

Analysis of the Existing Virginia SOS Protocol. The SOS water quality rating score determined from volunteer samples correlated only moderately with the MAIS score and with the HBI value determined from professional samples (Fig. 1). This lack of agreement was consistent for the concurrent samples and the historical samples. In all cases the *r*-values were below our target of 0.70. The relation-

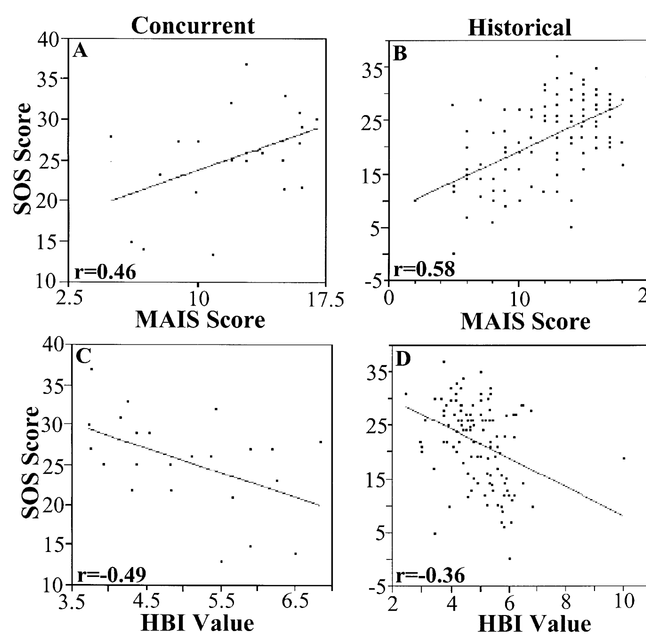


Fig. 1. Results of Pearson product-moment correlation analysis comparing Virginia (USA) Save-Our-Streams (SOS) water quality rating scores for volunteer samples to multimetric index (macroinvertebrate aggregated index for streams - MAIS) scores and Hilsenhoff biotic index (HBI) values for professional samples. (A and C) Volunteer samples taken concurrently in 1998. (B and D) Volunteer samples taken during the preceding 5 yr and professional samples taken at the same sites in 1998.

ship between the SOS water quality rating score and the HBI has a negative slope because HBI values decrease as ecological condition increases.

Classification analysis showed that, for some sampling locations, different conclusions about ecological condition would be reached by the SOS (volunteer) water quality ranking score and the MAIS (professional) score (Table 1). These disparities in classification in the concurrent samples were supported by those from the historical samples, with the rate of agreement between the conclusions from volunteer and professional samples below our criterion of 86% in both cases. Classification analysis also revealed a clear pattern in the disparities. In instances where the two protocols differed in their conclusions, the SOS score consistently concluded that ecological condition was acceptable, while the MAIS score concluded that ecological condition was unacceptable. McNemar's test revealed that the volunteer protocol significantly overrated water quality (Table 1). This means that the volunteer protocol would fail to detect degraded ecological conditions in some instances where they exist.

These results required that we address the question of whether the discrepancy in conclusions about ecological condition was caused by the data analysis done by volunteers (SOS water quality ranking score *per se*) or the volunteer protocol for acquiring the data (e.g., sampling, sorting, identifying, counting). We did this by interchanging the numerical results and sampling protocols, as explained in the methods.

Table 1. Classification analysis comparing the conclusions about ecological condition based on Macroinvertebrate Aggregated Index for Streams (MAIS) scores for professional samples to conclusions based on Virginia (USA) Save-Our-Streams (SOS) water quality rating scores for volunteer samples

	VA SOS water quality rating score			
	Concurrent samples		Historical samples	
	Acceptable	Unacceptable	Acceptable	Unacceptable
MAIS				
Acceptable	12	0	65	5
Unacceptable	8	3	25	27
% agreement	65%		75%	
McNemar test *P*-value	0.0082		0.0003	

Concurrent professional and volunteer samples were taken in 1998, as well as the professional samples for comparison to historical volunteer samples. Historical volunteer samples were taken during the preceding 5 yr. For the McNemar test, α = 0.05.

The results did not change when the SOS water quality rating score was recalculated for the professional samples. There was no significant difference between the SOS scores recalculated for the professional samples and the SOS scores originally calculated for the volunteer samples (t = 1.59, df = 22, P = 0.1271). There was a significant difference between the SOS scores recalculated for the professional samples and the SOS scores originally calculated for the volunteer samples for the historical samples (t = 5.55, df = 122, P < 0.0001). The SOS score calculated from professional samples did not correlate well with either the MAIS score or HBI value determined from professional samples (Fig. 2). This finding was consistent for the concurrent samples and the historical samples, all with *r*-values below our target of 0.70.

Classification analysis also showed that different conclusions about ecological condition were reached by the SOS score calculated from professional samples and the MAIS score calculated from professional samples (Table 2), both for the concurrent and historical samples. The rate of agreement between the conclusions was always below our criterion of 86%, the recalculated SOS score consistently

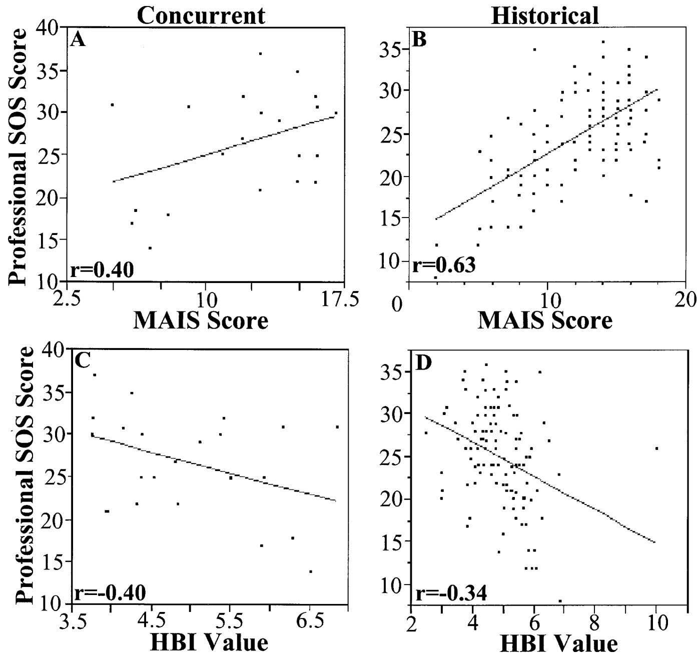


Fig. 2. Results of Pearson product-moment correlation analysis comparing Virginia (USA) Save-Our-Streams (SOS) water quality rating scores that we calculated for professional samples to multimertric index (macroinvertebrate aggregated index for streams, MAIS) scores and Hilsenhoff biotic index (HBI) values for professional samples. (A and C) Professional samples from concurrent sampling in 1998. (B and D) Professional samples taken in 1998 at sites previously sampled by volunteers.

overrated ecological condition, and McNemar's test indicated that the overrating of biological condition by the SOS protocol was significant (P < 0.05).

We then took the opposite approach by recalculating professional scores and values for the volunteer samples retained from concurrent sampling and comparing them to the original volunteer samples. We did this separately for the actual volunteer samples and the potential volunteer samples, the latter of which included the organisms they overlooked in the samples (Table 3). Paired *t*-tests revealed that a majority of the professional numerical scores and values recalculated for volunteer samples did not differ significantly from the original calculations for professional samples (P > 0.05). The exception was the HBI value for the actual volunteer samples (P = 0.0006). Comparisons of the actual volunteer samples to the potential volunteer samples with paired *t*-tests yielded similar results

Table 2. Classification analysis comparing the conclusions about ecological condition based on Macroinvertebrate Aggregated Index for Streams (MAIS) scores for professional samples to those based on Virginia (USA) Save-Our-Streams (SOS) water quality rating scores that we calculated for professional samples

	VA SOS water quality rating score			
	Concurrent samples		Historical samples	
	Acceptable	Unacceptable	Acceptable	Unacceptable
MAIS				
Acceptable	13	0	70	0
Unacceptable	9	1	42	10
% agreement	61%		65%	
McNemar test *P*-value	0.0027		<0.0001	

All professional samples were taken in 1998. Historical volunteer samples were taken during the preceding 5 yr. For the McNemar test, α = 0.05.

Table 3. Summary of results from correlation analyses and *t*-tests comparing Macroinvertebrate Aggregated Index for Streams (MAIS) scores and Hilsenhoff Biotic Index (HBI) values for professional samples to MAIS scores and HBI values calculated from volunteer samples

Comparisons	Correlation (*r*-values)	*t*-Test (*P*-values)
MAIS for professional samples compared with MAIS for actual volunteer samples	0.8561	0.7265
HBI for professional samples compared with HBI for actual volunteer samples	0.8361	0.0006
MAIS for professional samples compared with MAIS for potential volunteer samples	0.8033	0.2747
HBI for professional samples compared with HBI for potential volunteer samples	0.8787	0.1694
MAIS for actual volunteer samples compared with MAIS for potential volunteer samples	0.8712	0.0829
HBI for actual volunteer samples compared with HBI for potential volunteer samples	0.8936	0.0056

All data were from 1998 concurrent samples. Actual samples consisted of the organisms that volunteers removed from the sampling net. Potential samples consisted of the actual samples plus the organisms that we found remaining on the net after the volunteers finished sorting.

(Table 3). The MAIS score was not significantly different between actual and potential volunteer samples ($P > 0.05$), whereas the HBI score was significantly different ($P = 0.0056$). We found that all of the possible comparisons of recalculated professional scores and values for volunteer samples were highly correlated with those from the original professional samples, with *r*-values well above our target of 0.70 (Table 3). There was also strong correlation between the actual and potential volunteer samples for the MAIS scores and HBI values (Table 3).

It is not clear why there is a discrepancy with the HBI results. The most likely explanation is that volunteers were consistently overlooking some taxa that were facultative or tolerant. Facultative macroinvertebrates occur in environments with conditions ranging from pristine to moderate levels of disturbance, and often occur in high numbers under conditions of moderate disturbance. The small midges (Chironomidae), which are rated as facultative, were overlooked most often.

Classification analysis of decisions about ecological condition based on recalculated MAIS values for volunteer samples and the original professional samples indicated that the same conclusions were reached from both samples (Tables 4 and 5). This was true for both the actual and potential volunteer samples. The rates of agreement exceeded our target of 86%, and McNemar's test indicated that the volunteer protocol did not significantly over or under-rate ecological condition ($P > 0.05$). The disagreements were more evenly split between acceptable and unacceptable ecological conditions.

Table 4. Classification analysis comparing the conclusions about ecological condition based on Macroinvertebrate Aggregated Index for Streams (MAIS) scores for professional samples to those based on MAIS scores that we calculated for volunteer samples

	Actual volunteer MAIS		Potential volunteer MAIS	
	Acceptable	Unacceptable	Acceptable	Unacceptable
Professional MAIS				
Acceptable	11	2	13	0
Unacceptable	1	9	2	8
% agreement	87%		91%	
McNemar test *P*-value	0.5637		0.1573	

All data were from 1998 concurrent samples. Actual samples consisted of the organisms that the volunteers removed from the net. Potential samples consisted of the actual samples plus the organisms that we found remaining on the net after the volunteers finished sorting. For the McNemar test, ∝ = 0.05.

Table 5. Classification analysis comparing the conclusions about ecological condition based on Macroinvertebrate Aggregated Index for Streams (MAIS) scores that we calculated for potential volunteer samples to those based on MAIS scores that we calculated for actual volunteer samples

	Actual volunteer MAIS	
	Acceptable	Unacceptable
Potential Volunteer MAIS		
Acceptable	11	4
Unacceptable	1	7
% agreement	78%	
McNemar test *P*-value	0.1797	

All data were from 1998 concurrent samples. Actual samples consisted of the organisms that the volunteers removed from the net. Potential samples consisted of the actual samples plus the organisms that we found remaining on the net after the volunteers finished sorting. For the McNemar test, ∝ = 0.05.

Based on these analyses, we concluded that the SOS water quality ranking score, rather than volunteer collection methods or identification, was responsible for the discrepancy in conclusions about ecological condition made by volunteers and professional aquatic biologists. The SOS score did not adequately distinguish impaired ecological conditions, even when applied to samples taken by professionals. Conversely, the samples taken by volunteers were adequate for judging impaired ecological conditions, as long as professional numerical measures were calculated for the samples. The SOS score is not rigorous enough to distinguish impaired ecological conditions because it is based solely on the presence or absence of taxa, without consideration of their abundance. By using only the presence or absence of the different kinds of macroinvertebrates, the SOS score omits important information about the overall assemblage of organisms in streams. Because relative abundances are not included in the SOS score, all kinds of macroinvertebrates found in a sample are weighted equally. Thus, a single organism of a pollution-sensitive kind in a sample has the same weight of evidence as hundreds of organisms belong to a pollution-tolerant kind. One of the principles of community ecology is that in relatively stable, natural, undisturbed communities, there tend to be many kinds of organisms present with few individuals of most kinds. If a community is disturbed, either by natural events or human activities, the resulting community usually comes to have fewer kinds of organisms with many more individuals of the remaining kinds. Species that are adapted to exist in the changed conditions flourish in the absence of competition now released by the disappearance of the sensitive species. A simple metric based on the presence or absence of taxa will miss the changes that reflect this principle of community ecology, which is also one of the cornerstones of biological monitoring. The Virginia SOS protocol has an option for estimating categories of abundance for individual taxa: 0 to 10, 11 to 99, and >100 organisms. However, this option was only intended for use when the water quality rating score was very close to the cutoffs for the various categories of ecological condition. Although the volunteers we observed did estimate the abundances of individual macroinvertebrate kinds, we never saw these abundances considered in the final determination of ecological condition, and the SOS protocol provides no instructions on when or how to do this.

The volunteer protocol for acquiring data (sampling, sorting, identifying, and counting) did not appear to be part of the explanation for why volunteers and professionals reached different conclusions about ecological condition. Observations we recorded during the concurrent sampling events indicated that all the volunteers were adhering to the sampling and sorting methods in the Save-Our-Streams protocol. The area in front of the net was always thoroughly sampled down into the substrate, and all nets were placed on a sheet or board upon removal from the stream so that any organisms that might crawl through the net would be discovered. The volunteers were thorough in their sorting of both the net and the sheet below, stopping only when 100 organisms of a taxon were collected or no new taxa were found. The volunteers correctly identified the majority of the taxa. Their only taxonomic problem was differentiating flatworms from leeches. However, both of these macroinvertebrates were in the same pollution tolerance category for calculating the SOS water quality rating score, so this misidentification did not produce erroneous conclusions about ecological condition. The volunteers accurately placed all taxa into their correct category of estimated abundance.

Modification of the Virginia SOS Protocol. Our next step was to determine if the Virginia SOS protocol could be modified to bring the conclusions about ecological condition made by volunteers into close agreement with the conclusions made by professional aquatic biologists. If this proved to be possible, then it would be appropriate to use volunteer data for some regulatory purposes. There were two conspicuous modifications that would be likely to improve the accuracy of the Virginia SOS protocol: (1) identify all of the macroinvertebrates to lower taxonomic levels (at least to family) and (2) develop numerical analyses based on actual counts of the different kinds of macroinvertebrates contained in the samples.

In addition to improved accuracy, there was an overriding consideration that any modification of the Virginia SOS protocol must remain feasible for the volunteers who presently participate in the program. After discussions with volunteers and careful thought, we dismissed the idea of having volunteers identify macroinvertebrates to lower taxonomic levels. Being able to correctly identify all macroinvertebrates to at least the family level usually requires a college-level course or considerable experience, or both. Although groups in other states have had success with family-level identifications in volunteer programs (IDNR 1998), it has been accomplished by a division of labor among the volunteers. Most of the volunteers do field work to collect macroinvertebrate samples, which are preserved and shipped to a few select volunteers with taxonomic expertise. The macroinvertebrates are later identified in a laboratory setting, where the numerical results are also analyzed and conclusions about ecological condition are reached. As a result of our extensive interactions with Virginia SOS volunteers, we were aware that they wanted to be involved with the entire process and they wanted to continue to get immediate decisions on the ecological condition of sites before leaving. We were convinced that attempting family-level identifications in the field with volunteers would introduce excessive error into the program. In addition, we knew from our experience with macroinvertebrate monitoring that undesirable backlogs of preserved samples often occur. Thus, we decided to continue to use mostly order-level identifications that could be done in the field.

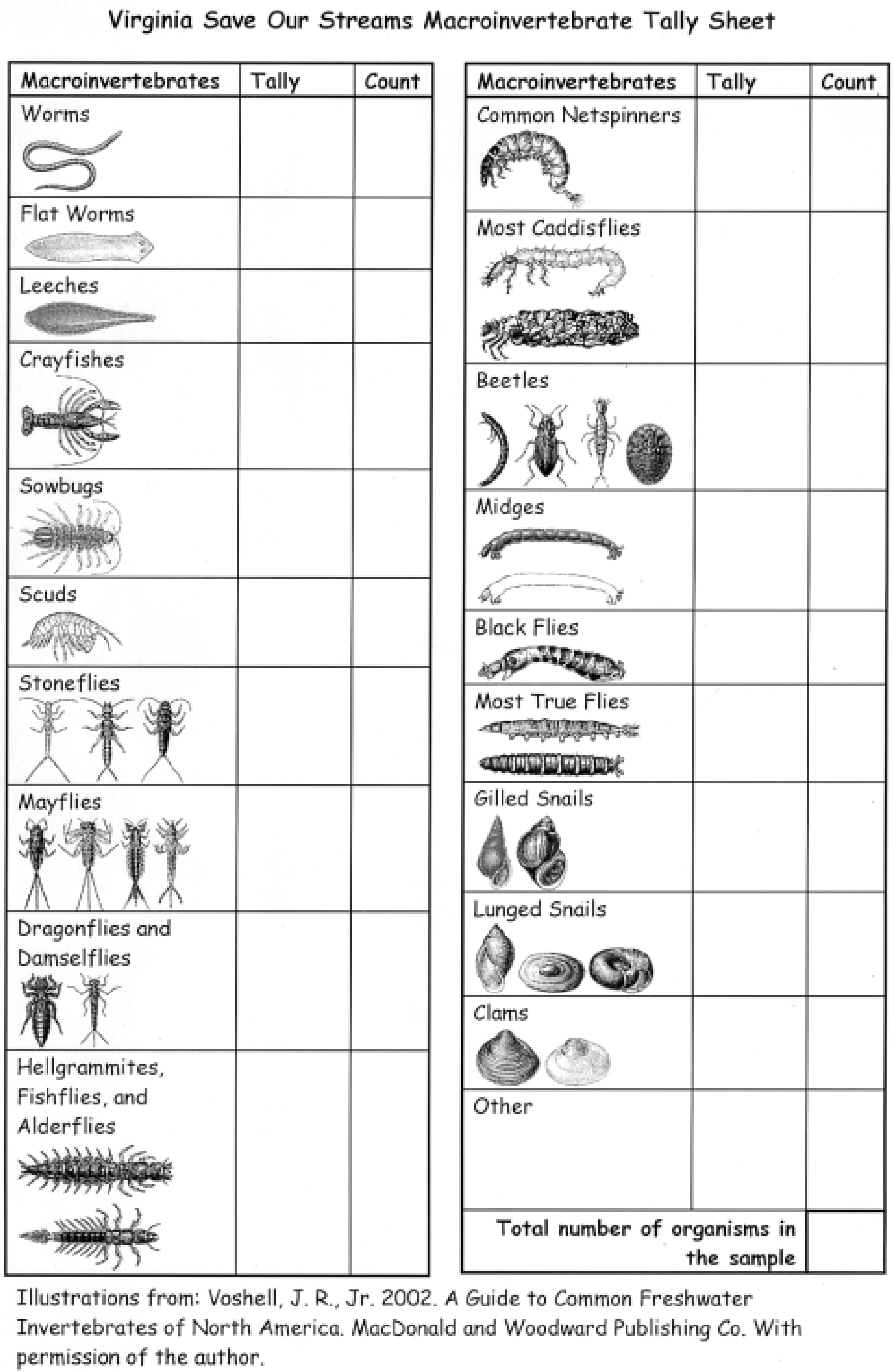
This decision meant that the only probable way to improve the accuracy of the Virginia SOS protocol was to develop a numerical score based on actual counts of the different kinds of macroinvertebrates in the samples. Although the original SOS sampling methods were not implicated in the inaccurate conclusions about ecological conditions, the sampling method had to be changed to obtain estimates of relative abundance that were both feasible and reliable. With the original SOS sampling method, volunteers did not sort and identify all of the organisms that were captured on the kick screen. They only looked at the material on the net until they reached a certain number of organisms within a taxonomic group, or until they had stopped finding any different kinds of organisms. To make accurate estimates of the relative abundance of the different kinds of macroinvertebrates, the entire contents of the sample collected by the volunteers must be sorted and identified. To only remove a predetermined number of the first organisms encountered introduces an appreciable bias toward the larger and more active macroinvertebrates. However, the kick screen used according to the original SOS protocol captured so many organisms that it would not be feasible to sort and identify all of them in a timely fashion. The concurrent samples collected by volunteers in 1998 contained an average of approximately 1500 organisms (range, 150 to >5,000 organisms). As a result, volunteers only sorted about 15% of the total organisms in the 1998 concurrent samples (range, 4-41%).

The protocols recommended by the U.S. Environmental Protection Agency for rapid assessment (Barbour et al. 1999) suggest taking standardized subsamples of large macroinvertebrate samples. Opinions are divided on the validity of using subsamples for monitoring (Barbour and Gerritsen 1996, Courtemanch 1999, Vinson and Hawkins 1999). We decided that subsampling would not be wise for volunteer monitoring because it would add another aspect of training and introduce another possible source of error. Various studies have reported that the required number of organisms in a sample to reach accurate conclusions about ecological condition ranges from 100 to >300 (Vinson and Hawkins 1996, Larsen and Herlihy 1998, Somers et al. 1998, Barbour et al. 1999). A comparative analysis of different sample sizes in Virginia streams indicated that values for most of the commonly used metrics become consistent when samples contain 200 organisms (J.R.V., unpublished data).

Therefore, we designed a standard sampling protocol that would enable volunteers to obtain an unbiased sample containing approximately 200 macroinvertebrates without subsampling. Volunteers select an area in a riffle that has typical rocky substrate and average current velocity for that riffle. One person holds a standard SOS kick net on the bottom, and another person thoroughly disturbs an area of approximately 0.1 m^2 with their hands in the front of the net for 20 s. The net is then brought to shore and spread on a white sheet. Volunteers sort the entire contents of the sample and keep a running tally of the total number of organisms. All organisms must be sorted, regardless of the total number. If the total number of organisms is 200 or more, the sample is complete. If there are fewer than 200 total organisms, additional samples are taken by the same technique in other places with similar features in the same riffle. All subsequent samples must be sorted in their entirety. Organisms from subsequent samples are added to the previous ones until the composite contains at least 200 organisms. The maximum number of samples is four. For subsequent samples, the times can be increased or decreased, if desired, up to a maximum of 90 s per net.

As with the original Virginia SOS protocol, organisms are identified in the field. We developed a field sheet with color illustrations to assist with identifications and forms to facilitate recording counts accurately (Fig. 3). Volunteers identify most of the arthropods to order and the other invertebrates to class. Within the insects, three kinds are identified to family. These are Hydropsychidae (net-spinning caddisflies) in the order Trichoptera, Chironomidae (non-biting midges) and Simuliidae (black flies) in the order Diptera. Within the class Gastropoda (snails and limpets), two subclasses are distinguished: Prosobranchia (gilled snails) and Pulmonata (lunged snails). The three families of aquatic insects are commonly collected by volunteers and tend to be tolerant of degraded ecological conditions, especially moderate eutrophication and organic loading. The lunged snails, commonly called left-handed snails because of the direction of their spiral shell, are much more tolerant of low dissolved oxygen concentration because they can also breathe from the atmosphere.

(A)



Virginia Save Our Streams Macroinvertebrate Tally Sheet

Macroinvertebrates	Tally	Count
Worms		
Flat Worms		
Leeches		
Crayfishes		
Sowbugs		
Scuds		
Stoneflies		
Mayflies		
Dragonflies and Damselflies		
Hellgrammites, Fishflies, and Alderflies		

Macroinvertebrates	Tally	Count
Common Netspinners		
Most Caddisflies		
Beetles		
Midges		
Black Flies		
Most True Flies		
Gilled Snails		
Lunged Snails		
Clams		
Other		
Total number of organisms in the sample		

Illustrations from: Voshell, J. R., Jr. 2002. *A Guide to Common Freshwater Invertebrates of North America*. MacDonald and Woodward Publishing Co. With permission of the author.

Fig. 3. New field sheets developed for use in the modfied Virginia Save-Our-Streams protocol. (A) Sheet for identifying macroinvertebrates and recording counts.

(B)

Individual Metrics

Metric	Number		Total number of organisms in the sample		Percent
Mayflies + Stoneflies + Most Caddisflies		Divide by		Multiply by 100	
Common Netspinners		Divide by		Multiply by 100	
Lunged Snails		Divide by		Multiply by 100	
Beetles		Divide by		Multiply by 100	

% Tolerant

Taxon	Number
Worms	
Flatworms	
Leeches	
Sowbugs	
Scuds	
Dragonflies and Damselflies	
Midges	
Black Flies	
Lunged Snails	
Clams	
Total Tolerant	
Total Tolerant divided by the total number of organisms in the sample	
Multiply by 100	

% Non-Insects

Taxon	Number
Worms	
Flatworms	
Leeches	
Crayfish	
Sowbugs	
Scuds	
Gilled Snails	
Lunged Snails	
Clams	
Total Non-Insects	
Total Non-Insects divided by the total number of organisms in the sample	
Multiply by 100	

Fig. 3. (B) Sheet for calculating individual metrics.

(C)

Save Our Streams Multimetric Index

Determine whether each metric should get a score of 2,1, or 0. Write your metric value from the previous page in the 2nd column (Your Metric Value). Put a check in the appropriate boxes for 2,1, or 0. Then calculate the subtotals and Save Our Streams Multimetric Index score and determine whether the site has acceptable or unacceptable ecological condition.

Metric	Your Metric Value	2	1	0
% Mayflies + Stoneflies + Most Caddisflies		Greater than 32.2	16.1 - 32.2	Less than 16.1
% Common Netspinners		Less than 19.7	19.7 - 34.5	Greater than 34.5
% Lunged Snails		Less than 0.3	0.3 - 1.5	Greater than 1.5
% Beetles		Greater then 6.4	3.2 - 6.4	Less than 3.2
% Tolerant		Less than 46.7	46.7 - 61.5	Greater than 61.5
% Non-Insects		Less than 5.4	5.4 - 20.8	Greater than 20.8
		Total # of 2s:	Total # of 1s:	Total # of 0s:
	Subtotals:	Multiply by 2:	Multiply by 1:	Multiply by 0:

Now add the 3 subtotals to get the Save Our Streams Multimetric Index score:________________

_____Acceptable ecological condition (7 to 12)

_____Unacceptable ecological condition (0 to 6)

Fig. 3. (C) Sheet for calculating new Virginia (USA) Save-Our-Streams multimetric index and determining the category of ecological condition.

With counts of the different kinds of macroinvertebrates in a sample, it was possible to calculate a variety of metrics and a multimetric index for volunteer monitoring, as suggested for data analysis and interpretation by the U.S. EPA (1997). The principle of multimetric indices is that individual metrics are combined to give a single score that reflects ecological condition. There are some caveats to using such an index. Some scientists are concerned that not enough is known about how individual metrics respond to impairment, what impairment they respond to, and if metrics applied to differing life stages respond differently to an impairment (Norris 1995). In addition, volunteers have to complete more calculations to arrive at a final score. However, the benefits of using a multimetric index outweigh the possible concerns. Multimetric indices, once developed, offer a cost-effective way to analyze data in a way that incorporates much ecological information. Multimetric indices also lead to an accurate final score that is easily understood by professionals and volunteers alike (Norris 1995, U.S. Environmental Protection Agency 1997, Barbour et al. 1999). Professionals can get an idea of the type, and possibly source, of impairment by interpreting the information within a multimetric index.

We generally followed the stepwise framework suggested by Barbour et al. (1995) and Barbour et al. (1996) to develop a multimetric index for use by volunteers in the Virginia SOS program. This multimetric index was developed as much as possible with data from the 145 professional samples taken in rocky-bottomed streams throughout western Virginia in 1998. Then we finalized the development of the new multimetric index and validated it with the data from the 23 professional and volunteer samples taken concurrently in 1999. The first step in the process was to evaluate all metrics that were feasible for volunteers to calculate, then choose a subset of the best ones. Feasible meant that it was possible to calculate the metric from samples that were identified mostly to the order level. We defined best metrics as those that had low variability within reference streams, but exhibited distinguishable, predictable changes in their values for streams known to be impaired. The list of metrics began with the 69 that were evaluated by Smith and Voshell (1997). We added several other metrics that we thought might be of interest. Twenty-four metrics were deemed feasible for volunteers to calculate (Table 6). Among these was the Citizen Biotic Index, which is a new, order-level index based on the Hilsenhoff biotic index. The order/select family-level tolerance values used in the new biotic index, as well as other tolerance based metrics, ranged from 1 to 10 with 1 being the most tolerant. These tolerance values for higher taxonomic levels were modified by best professional judgement from a database of genus-level tolerance values that has been developed during two decades of pollution studies in Virginia streams (J.R.V., unpublished data).

The effectiveness of these metrics was analyzed using all professional data from both the historical resampling of volunteer sites as well as the 1998 concurrent sampling events. The sites were divided into reference (acceptable) and impaired (unacceptable) ecological condition based on the MAIS scores for professional samples (≥13 acceptable, ≤12 unacceptable). The mean and coefficient of variance for all 24 metrics were calculated separately for the reference and impaired sites, and a separation statistic was calculated that compared reference and impaired sites. Our criteria for metrics being effective at distinguishing ecological condition were: coefficient of variation <50% and separation statistic >1 or < -1. In addition to these statistical criteria, we considered the ecological information contained in the metrics. Although statistical performance was given the highest priority, we tried to include metrics that provided a variety of meaningful information about the structure and function of the benthic community. This produced a total of 15 candidate metrics for possible inclusion in the Virginia SOS multimetric index (Table 6).

The various metrics that might be selected for a multimetric index have different ranges of values and even different directions of responses. Metrics cannot be combined until they have been standardized to have the same possible range of values and directions of responses. Unlike metrics were combined by standardizing the individual metrics as unitless scores of 2, 1, and 0. The highest score was assigned to indicate a close approximation of reference condition, then consecutively lower scores were assigned to progressively lower metric values indicating impaired conditions. This was accomplished by producing a boxplot of each metric for all 1998 professional data from reference sites (Fig. 4), following the methods of Barbour et al. (1996). For metrics that decrease in value when perturbation occurs (Table 6), values in the second quartile and above were assigned a score of 2 (Fig. 4). Metric values from the second quartile to the minimum possible value were divided in half, and those values in the upper half were assigned a score of 1, and those in the lower half were assigned a score of 0. For metrics that increase in value with perturbation (Table 6), the procedure was reversed. Metric values in the first three quartiles were assigned a score of 2 (Fig. 4). Metric values from the third quartile to the maximum possible value were divided in half, and those values in the lower half were assigned a score of 1 while those in the upper half were assigned a score of 0.

After transforming the values of the 15 individual candidate metrics into standardized unitless less scores, 20 selected groups of these metrics were aggregated into different multimetric indices by summing the scores of the individual metrics. Each of the 20 aggregations contained 5-10 individual metrics from the list of 15 candidate metrics. The potential aggregations were chosen so that a variety of ecological information would be contained in the metrics, with as little redundancy as possible. In addition, it was desirable for the Virginia SOS multimetric index to have some metrics that increased in value with impairment as well as some that decreased in value with impairment. This would make the multimetric index more effective for correctly assessing the ecological condition of both pristine and highly impaired waters.

At this point in the analyses, we switched to the data from concurrent sampling in 1999 because those were the only volunteer samples taken using the modified sampling method. The sites were again divided into reference (acceptable) and impaired (unacceptable) ecological condition based on the MAIS scores for professional samples (≥13 acceptable, ≤12 unacceptable), as had been done for the 1998 samples. However, in 1999 we enlisted the aid of professional biologists in the Virginia Department of Environmental Quality (DEQ) to review our decisions about reference versus impaired ecological conditions at the 23 sites. For any volunteer monitoring program to be successful, it must reach the same conclusions as the government agencies designated with regulatory authority. The DEQ biologists sample macroinvertebrates with a D-frame dip net and analyze a 100-organism subsample. The DEQ analysis incorporates MAIS scores as part of the assessment, but also includes physical and chemical measurements and information on permitted discharges and land use. Most importantly, DEQ biologists have many years of experience and have visited these streams over a long period of time. Our assessments of ecological condition agreed with those of the DEQ biologists for 19 of the 23 sites, and we gave priority to their conclusions for the other four sites.

The essential element in the Virginia SOS multimetric index had to be a numerical threshold for acceptable versus unacceptable ecological condition. We determined this threshold the same way as Smith and Voshell (1997), which was to average the multimetric index mean for reference sites and the multimetric index mean for impaired sites. This approach is based on the classification cutoff from linear discriminant analysis. For the final selection of the suite of metrics to be combined into the new Virginia SOS multimetric index, we determined which aggregation of volunteer metrics best agreed with the conclusions about ecological condition reached by using the MAIS and the experience of professional biologists. This was done by the same correlation and classification procedures used in the original analysis of the Virginia SOS water quality rating score.

Several aggregations of the candidate metrics produced multimetric indices that correlated well with MAIS scores and as-

Table 6. Feasible metrics for use with volunteer data

Metric	Response to perturbation[a]	Reference mean	Reference CV[b]	Impacted mean	Impacted CV[b]	Separation statistic
% Amphipoda	+	0.02	398.79	0.56	357.41	0.41
% Bivalves	-	0.55	314.97	0.34	223.30	-0.15
% Chironomidae[c]	+	13.06	61.48	35.89	55.10	1.59
Citizen Biotic Index[c]	-	6.61	11.43	4.79	22.30	-2.00
% Coleoptera[c,d]	-	15.48	77.60	8.47	126.67	-0.61
Coleoptera/(Coleoptera + Hydropsychidae)[c]	-	0.52	49.62	0.34	93.31	-0.62
% Crustacea + Mollusca	+ or -	1.49	201.92	3.68	271.39	0.32
% Diptera -Chironomidae	-	7.44	101.89	7.15	141.19	-0.03
% Ephemeroptera[c]	-	32.03	47.56	14.90	116.40	-1.05
% EPT[c,e]	-	56.72	25.07	36.82	60.89	-1.09
EPT/(EPT + Chironomidae)[c,e]	-	0.81	14.63	0.50	49.12	-1.71
% EPT – Hydropsychidae[c,d,e]	-	43.52	36.51	18.66	97.69	-1.46
% Prosobranchia	-	3.31	206.93	1.03	267.40	-0.41
% Hydropsychidae[c,d]	+	13.20	73.77	18.16	96.70	0.36
Hydropsychidae/ Trichoptera[c]	+	71.98	35.76	82.65	35.50	0.39
% Intolerant[c]	-	64.44	25.30	31.10	61.66	-1.88
% Isopoda	+	0.06	456.88	3.82	326.01	0.46
% Pulmonata[c,d]	+	0.27	225.54	1.56	380.48	0.33
% Non-insects[c,d]	+	4.89	159.69	10.90	144.98	0.50
% Oligochaeta	+	0.45	260.62	3.05	208.58	0.61
% Plecoptera	-	6.39	159.61	0.78	304.55	-0.70
% Gastropoda	+ or -	3.58	195.34	2.59	249.21	-0.15
% Tolerant[c,d]	+	35.56	45.86	68.90	27.82	1.88
% 1 Dominant[c,d]	+	30.30	45.04	46.32	33.62	1.10

[a] For expected response to perturbation, - = decrease in value and + = increase in value.
[b] CV = Coefficient of Variation.
[c] 15 candidate metrics for possible aggregation into a multimetric index.
[d] Six final choices for metrics to be aggregated into a multimetric index.
[e] EPT = Ephemeroptera, Plecoptera, and Trichoptera.

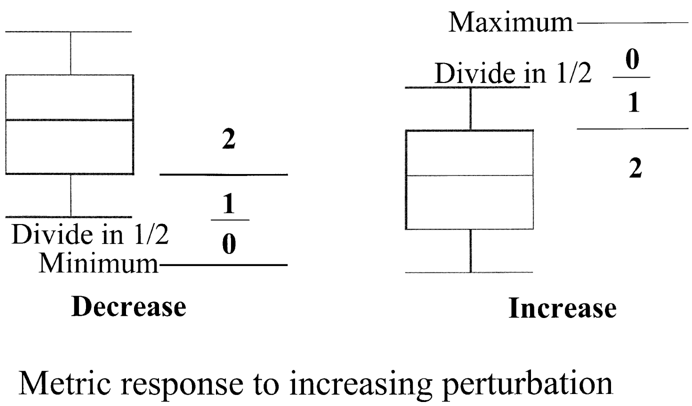


Fig. 4. Box plot method (Barbour et al. 1996) used to standardize metrics into unitless scores for aggregation into the new Virginia Save-Our-Streams multimetric index.

sessed ecological condition similarly to professional biologists (Table 7). We decided that the best multimetric index for the Virginia SOS program was the one composed of the following six metrics: % tolerant, % EPT-Hydropsychidae, % Hydropsychidae, % Pulmonata, % non-insects, and % Coleoptera. The Pearson product-moment correlation between this multimetric index and the professional MAIS resulted in an *r*-value of 0.6923 ($P = 0.0003$) (Fig. 5), which was only narrowly below our *a priori* criterion of 0.70. The classification analysis comparing this multimetric index's assessment of ecological condition to those of professional biologists indicated that the two methods came to the same conclusion 95.7% of the time (Table 8). This was well above our criterion of 86%, and McNemar's test indicated that the volunteer protocol did not significantly over or under-rate ecological condition ($P = 0.3173$).

Discussion

In addition to good statistical agreement between the new SOS volunteer multimetric index and professional approaches, the individual metrics that comprise the volunteer multimetric index contain meaningful ecological information about the structure and function of the benthic community. The percentage tolerant metric is based on numerical values that characterize the general ability of the various taxa to withstand pollution or other environmental stress. In the data analysis technique that we developed, these numerical values range from 1 to 10, with 1 being the most tolerant. We consider values ≤5 to reflect taxa that can withstand a great deal of stress and probably not be eliminated from the benthic community. Thus, if the proportion of organisms belonging to taxa with tolerance values ≤5 increases appreciably, this is a reliable sign that the stream may be affected by pollution.

EPT is an abbreviation for Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies), which are three orders of aquatic insects that are common in streams. Almost all species in these orders are sensitive to pollution. Hydropsychidae is a common family of Trichoptera that is considerably different by being somewhat sensitive, or facultative, to pollution. Therefore, the percentage of organisms belonging to these three orders, minus the facultative Hydropsychidae, will be high in undisturbed streams but will become lower if pollution is introduced.

The percentage non-insects metric responds in the opposite direction. Most of the invertebrates other than insects that are commonly collected in streams by volunteers are Turbellaria (flatworms), Oligochaetes (aquatic earthworms), Hirudinea (leeches), Isopoda (aquatic sow bugs), Amphipoda (scuds), and Decapoda (crayfishes). Almost all of the species in these groups range from being tolerant to facultative to pollution, so high percentages of these organisms are indicative of impaired ecological conditions.

The percentage Pulmonata metric provides information about the level of dissolved oxygen, which usually decreases for certain types of pollution, such as organic wastes or nutrients. The class Gastropoda (snails and limpets) is divided into two subclasses according to how they breathe, Prosobranchia (gilled snails) and Pulmonata (lunged or pouch snails). Pulmonata can breathe from a body cavity that they refill with air, so they do not depend on dissolved oxygen being present in the water. In polluted waters with little dissolved oxygen, Pulmonata often become very abundant because there is a great deal of organic matter for food and little competition from other invertebrates. Volunteers can easily recognize the most common kinds of Pulmonata because the opening of the spiral shell is on the left side when the narrow end is held up (commonly referred to as left-handed snails).

The remaining two metrics, percentage Hydropsychidae and per-

Table 7. Summary of statistical analysis of possible multimetric indices for volunteer samples that correlated most closely with Macroinvertebrate Aggregated Index for Streams (MAIS) values for professional samples and agreed most closely with assessments of ecological condition made by professional biologists

Aggregation of metrics	Acceptable scores	Unacceptable scores	Correlation analysis *r*-value	Classification analysis % agreement
6 metrics: % Tolerant, % EPT-Hydropsychidae, % Hydropsychidae, % Pulmonata, % non-insects, and % Coleoptera	7 - 12	0 - 6	0.6923	96%
8 metrics: % Tolerant, % EPT-Hydropsychidae, % Hydropsychidae, % Pulmonata, % Coleoptera/ (Coleoptera+Hydropsychidae), Hydropsychidae/ Trichoptera, % non-insects, and % Coleoptera	9-16	0-8	0.6933	91%
7 metrics: % Tolerant, % EPT-Hydropsychidae, % Hydropsychidae, % Pulmonata, % Coleoptera/ (Coleoptera+Hydropsychidae), % non-insects, and % Coleoptera	8-14	0-7	0.6910	91%
5 metrics: % Tolerant, % Hydropsychidae, Coleoptera/ (Coleoptera+Hydropsychidae), % non-insects, and % Coleoptera	6-10	0-5	0.6914	87%

All data were from 1999 concurrent samples.

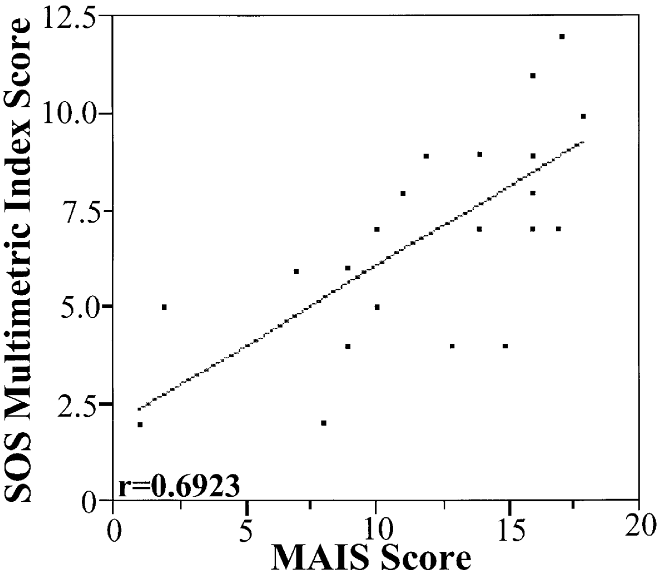


Fig. 5 Reults of Pearson product-moment correlation analysis comparing new Virginia (USA) Save-Our-Streams (SOS) multimetric index scores with professional macroinvertebrate aggregated index for streams (MAIS) scores. Data were from concurrent volunteer and professional samples taken in 1999.

centage Coleoptera are based on taxonomic composition, but they also provide information on trophic dynamics and movement habits. Hence, these metrics provide insight into the ecological function of the community, as well as its structure. Hydropsychidae (common net spinner caddisflies) feed by constructing fine mesh nets out of silk, which they position in current to filter particles of detritus and algae from the water. When streams are polluted with moderate amounts of organic wastes or nutrients (various ions of phosphorus and nitrogen), the amount of organic matter suspended in the water increases. This provides additional food for the filter-feeding Hydropsychidae, and their populations increase accordingly. Volunteers can easily identify Hydropsychidae, and a high percentage of these organisms is a reliable indicator of organic loading.

The aquatic Coleoptera (beetles) that are most often collected by volunteer stream monitors are members of two families, Psephenidae (water pennies) and Elmidae (riffle beetles). Members of both families feed by scraping algae that is tightly attached to rocks and other firm substrates in shallow areas of flowing water. Under natural conditions, there is a very thin layer of healthy, nutritious algae cells growing on the substrate, so these scrapers are able to cling to the substrate without being swept away by the swift current and obtain a nourishing diet. When streams become polluted, the layer of algae often becomes thick and contains cells that are dead or dying. The thick layer of soft, slimy algae prevents scrapers from being able to hold on to the substrate in swift current, and the algae that is present is not nearly as nutritious. The end result is that the percentage of scrapers decreases when streams become polluted because of changes in the algae growing on the solid substrate. The percentage Coleoptera metric provides volunteers a reliable way to track that effect.

Table 8. Classification analysis comparing the conclusions about ecological condition based on Macroinvertebrate Aggregated Index for Streams (MAIS) scores for professional samples to those based on the new Virginia (USA) Save-Our-Streams (SOS) multimetric index for volunteer samples

	Virginia SOS multimetric index	
	Acceptable	Unacceptable
Professional MAIS		
Acceptable	14	1
Unacceptable	0	8
% agreement	95.7%	
McNemar test P-value	0.3173	

All data were from 1999 concurrent samples. For the McNemar test, $\propto = 0.05$

Using biological assessments performed by volunteers for the authorized purposes of regulatory and natural resource agencies is a matter that should not be taken lightly. There can be important outcomes of these activities, some of which would not be desirable. If volunteers conclude that a stream is impaired, when in truth it is not, regulatory actions can be triggered that will waste the time and meager budget of professional biologists, as well as cause significant negative socioeconomic impacts. If volunteers conclude that a stream has acceptable ecological conditions, when in truth it is impaired from human activities, the problem will likely worsen and cause significant damage to the environment that could have been avoided by accurate early detection. Thus, the consequences of inaccurate volunteer biological monitoring may be worse than not making any official use of volunteer data. However, our study has shown that this does not have to be the case.

The original Virginia Save-Our-Streams program that was the subject of this study consistently overrated the ecological condition of streams. In a statistically significant number of instances ($\propto =$ 0.05), professional measures of ecological condition revealed that streams classified as being acceptable by the SOS protocol were actually impaired. Conversely, volunteers using the SOS protocol never classified a stream as impaired that was not impaired.

Although we only analyzed one volunteer biological monitoring program (the one in Virginia), our results are probably broadly applicable because many programs use almost the same protocol. Most common protocols are based on presence or absence of kinds of benthic macroinvertebrates that are identified only to higher taxonomic levels (classes, orders, a few select families) and divided into three pollution tolerance categories (sensitive, somewhat sensitive or facultative, and tolerant). Every volunteer monitoring program that is based on a similar protocol is very likely to overrate the ecological condition of streams and fail to differentiate impaired and healthy streams. Well-meaning, dedicated volunteers do not necessarily produce valid scientific results by carefully adhering to a monitoring protocol that has been promulgated for their use. Volunteer monitoring protocols must be analyzed in detail and compared with appropriate statistical techniques to confirm that they reach the same conclusions as the professional protocols being used by government agencies in an area. Without such rigorous validation studies, professional biologists will always be skeptical, and justifiably so, about the results of volunteer biological monitoring programs.

This study should serve as an example of how volunteer biological monitoring programs can be modified and validated to provide reliable data that are consistent with the results of professional biologists and suitable for making the basic assessment of whether a stream is impaired or not. We found that the essential modification was to calculate an assortment of ecologically meaningful metrics based on numbers of organisms belonging to each kind of macroinvertebrate, rather than just presence or absence of the kinds. To do this, it was necessary to make the sampling protocol more quantitative so that unbiased counts of the numbers of individual organisms in each group were obtained. It was not necessary to identify the organisms to lower taxonomic levels. In fact, our recommended protocol involves fewer family-level identifications than the original SOS protocol. Also, it was not necessary to use a net with a finer mesh. We found that a multimetric index, which aggregates a group of individual metrics into a single score, correlated well with a professional multimetric index and the conclusions about ecological condition agreed very closely with those made by professional aquatic biologists.

Finally, we demonstrated that the modified Virginia SOS protocol was feasible for volunteers currently participating in the program. However, it is only the process of this study that should be used by other volunteer monitoring programs. The specific protocol that emerged from this study should not be automatically adopted as a standard method for volunteer biological monitoring everywhere. The modified Virginia SOS protocol is probably valid in shallow rocky-bottom streams in the mid-Atlantic region of the eastern United States, but volunteer programs in other areas need to do thorough validation studies, such as conducted here. Certain elements of our results may be useful in other areas. In Table 7 there are three other aggregations of metrics that performed almost as well as the one we chose. Some of the other 24 feasible metrics listed in Table 6 may be useful, especially the 15 candidate metrics that exhibited favorable statistical properties in our study.

Regardless of how effective a volunteer biological monitoring protocol proves to be in a validation study, there must be an adequate quality assurance/quality control plan to guarantee that the protocol is consistently adhered to by all participating volunteer groups. Volunteers should be certified by training and testing before conducting bioassessments, and periodic recertification should be required, perhaps every 2 or 3 yr. All participants should be required to preserve at least 10% of their samples each year to have their identifications checked by a professional biologist, or at least a very experienced volunteer.

If volunteer biological monitoring programs are carefully analyzed, modified where necessary, validated, and then strictly adhered to, professional biologists and others in regulatory and natural resource agencies should accept the results, be confident about using them, and be grateful for the assistance. The work of volunteer monitoring programs is not intended to take the place of professional biologists. The advanced training and career experience of professional biologists are necessary for many aspects of such monitoring, such as documenting the cause of impairment, quantifying effects, and developing plans to solve problems. In addition, there are many instances when purely numerical results do not produce straightforward interpretations because streams are such complex ecosystems. In these instances, issues have to be resolved by relying to some degree, on best professional judgement, which is most reliably obtained from experienced professional biologists.

We have shown, however, that volunteers can reliably assess whether the ecological condition of a stream is impaired or not, if a sound protocol is developed according to scientific principles. This proven ability of volunteers should be used to the fullest extent to assess the ecological condition of the vast reaches of streams that need attention. This would provide professional biologists more time to accomplish the scientific activities that only they are qualified for. A mutualistic arrangement between volunteers and professionals would seem to be the only way that streams are going to receive appropriate environmental stewardship.

Acknowledgements

Sara Grove, Dana Roberts, and Kim Frazer provided field and laboratory technical assistance. Jay Gilliam and the Virginia Save Our Stream volunteers provided data and allowed us to sample with them in the field. A number of individual volunteers were extremely helpful at recruiting other volunteers and organizing field experiences: Llyn Sharp, Jerry Elkins, John Hermsmeier, Wayne Teel, Dianne Hoffman, and Deanna Crumbling. Nick Stone and Rick Fell reviewed the manuscript. The Virginia Department of Environmental Quality provided funding for this study.

References Cited

Barbour, M. T. 1997. The re-invention of biological assessment in the U.S. Human Ecol. Risk Assess. 3: 933-940.

Barbour, M. T., J. B. Stribling, and J. R. Karr. 1995. Multimetric approach for establishing biocriteria and measuring biological condition, pp. 63-77. *In* W. S. Davis and T. P. Simon (eds.), Biological assessment and criteria. Lewis, Boca Raton, FL.

Barbour, M. T., and J. M. Gerritsen. 1996. Subsampling of benthic samples: a defense of the fixed count method. J. North Am. Benthol. Soc. 15: 386-391.

Barbour, M. T., J. M. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White, and M. L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. J. North Am. Benthol. Soc. 15: 185-211.

Barbour, M. T., J. Gerritsen, and B. Snyder. 1999. Rapid bioassessment protocols for use in wadeable streams and rivers: periphyton, benthic macroinvertebrates, and fish, 2nd ed. EPA 841-B-99-002. Office of Water, Washington, DC.

Cooke, K. 1999. Kentucky Water Watch. Frankfort, Kentucky (http://water.nr.state.ky.us/ww/wwhomepg.htm).

Courtemanch, D. L. 1996. Commentary on the subsampling procedures used for rapid bioassessments. J. North Am. Benthol. Soc. 15: 381-385.

DeWalt, R. E. 1999. Congruence of Riverwatch and CTAP stream biomonitoring data: results for 1998. Center Biodiv. Tech. Rep. 15.

Dilley, M. A. 1991. A comparison of the results of a volunteer stream quality monitoring program and the Ohio EPA's biological indices. Undergraduate Honors Research, School of Natural Resources, The Ohio State University, Columbus.

Ely, E. 1997. Parallel testing – volunteers vs. professionals. The Volunteer Monitor 9 (http://www.epa.gov/volunteer/spring97/mettec16a.html, http://www.epa.gov/volunteer/spring97/mettec16.html, http://www.epa.gov/volunteer/spring97/mettec17.html, http://www.epa.gov/volunteer/spring97/mettec18.html).

Ely, E. 2000. Macroinvertebrate data: volunteers versus professionals. The Volunteer Monitor 12: 1 and 6-7.

Gibson, G. R., M. T. Barbour, J. B. Stribling, J. Gerritsen, and J. R. Karr. 1996. Biological criteria: technical guidance for streams and rivers. EPA/822-B-94-001. Office of Science and Technology, Washington, DC.

Godfrey, P. J. 1994. Shared waters and common goals: Enhancing partnerships in water quality monitoring and decision-making. Univ. Mass. Water Resour. Res. Center Rep. 1994.

Hilsenhoff, W. L. 1987. An improved biotic index of organic stream pollution. Gt. Lakes Entomol. 20: 31-39.

(IDNR) Illinois Department of Natural Resources. 1998. Illinois RiverWatch stream monitoring manual, 3rd ed. Division of Energy and Environmental Assessment, Springfield, IL.

Independent Sector. 1999. Giving and volunteering in the United States: executive summary. Washington, DC (http://www.independentsector.org/GandV/default.htm).

Kellogg, L. 1994. Monitor's guide to aquatic macroinvertebrates. The Izaak Walton League of America, Gaithersburg, MD.

Larsen, D. P., and A. T. Herlihy. 1998. The dilemma of sampling streams for macroinvertebrate richness. J. North Am. Benthol. Soc. 17: 359-366.

Lathrop, J. E., and A. Markowitz. 1995. Monitoring water resource quality using volunteers, pp. 303-314. *In* W. S. Davis and T. P. Simon (eds.), Biological assessment and criteria. Lewis, Boca Raton, FL.

Lee, V. 1988. Rhode Island volunteers monitor the health of salt ponds. Oceanus 31: 44-48.

Lenat, D. R., and M. T. Barbour. 1993. Using benthic macroinvertebrate community structure for rapid, cost-effective, water quality monitoring: rapid bioassessment, pp. 187-215. *In* S. L. Loeb and A. Spacie (eds.), Biological monitoring of aquatic systems. Lewis, Boca Raton, FL.

Levy, S. 1998. Using bugs to bust polluters. Bioscience 48: 342-346.

Livermore, B. 1993. Riding the toxic surf. Sea Frontiers 39: 28-35.

Maas, R. P., D. J. Kucken, and P. F. Gregutt. 1991. Developing a rigorous water quality database through a volunteer monitoring network. Lakes Reserv. Manag. 7: 123-126.

Markusic, J. S. 1991. Cost-effective monitoring using lay volunteers on Shipshewana Lake. Lakes Reserv. Manag. 7: 126-131.

Matthews, R. A., A. L. Buikema, Jr., J. Cairns, Jr., and J. H. Rogers, Jr. 1982. Biological Monitoring Part IIA-Receiving system functional methods, relationships, and indices. Water Res. 16: 129-139.

Mattson, M. D., M. F. Walk, P. A. Kerr, A. M. Slepski, O. T. Zajicek, and P. J. Godfrey. 1994. Quality assurance testing for a large scale volunteer monitoring program: The acid rain monitoring project. Lakes Reserv. Manag. 9: 10-13.

Miller, D. L., et al. 1988. Regional applications of an index of biotic integrity for use in water resource management. Fisheries 13: 12-20.

Moore, D. S., and G. P. McCabe. 1993. Introduction to the practice of statistics, 2nd ed. Freeman, New York.

Norris, R. H. 1995. Biological monitoring: the dilemma of data analysis. J. North Am. Benthol. Soc. 35: 440-450.

Penrose, D., and S. M. Call. 1995. Volunteer monitoring of benthic macroinvertebrates: regulatory biologists' perspectives. J. North Am. Benthol. Soc. 14: 203-209.

Plafkin, J. L., M. T. Barbour, K. D. Porter, S. K. Gross, and R. M. Hughes. 1989. Rapid bioassessment protocols for use in streams and rivers: benthic macroinvertebrates and fish. EPA 440/4-89/001. Office of Water, Washington, DC.

Ramsey, F. L., and D. W. Schafer. 1997. The statistical sleuth: a course in methods of data analysis. Duxbury Press, Belmont, CA.

Resh, V. H., and J. K. Jackson. 1993. Rapid assessment approaches to biomonitoring using benthic macroinvertebrates, pp. 195-233. *In* D. M. Rosenberg and V. H. Resh (eds.), Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York.

Rosenberg, D. M., and V. H. Resh. 1993. Introduction to freshwater biomonitoring and benthic macroinvertebrates, pp. 1-9. *In* D. M. Rosenberg and V. H. Resh (es.), Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York.

Sall, J., and A. Lehman. 1996. JMP start statistics: a guide to statistical and data analysis using JMP and JMP IN software. Duxbury Press, Belmont, CA.

Sheehan, L. 1998. Strategies for creating a comprehensive, integrated coastal water quality monitoring network. Taking a look at California's ocean resources: an agenda for the future 2: 1442.

Smith, E. P., and J. R. Voshell. 1997. Studies of benthic macroinvertebrates and fish in streams within EPA Region 3 for development of biological indicators of ecological condition. Part 1, benthic macroinvertebrates. Report to U.S. Environmental Protection Agency. Cooperative Agreement CF821462010. EPA, Washington, DC.

Sokal, R. R., and F. J. Rohlf. 1969. Biometry. Freeman, San Francisco, CA.

Somers, K. M., R. A. Reid, and S. M. David. 1998. Rapid biological assessments: how many animals are enough? J. North Am. Benthol. Soc.17: 348-358.

Stokes, M. E., C. S. Davis, and G. G. Koch. 1995. Categorical data analysis using the SAS system. SAS Institute, Cary, NC.

Thomson, C. H. 1987. The acid drops project: pollution monitoring by young people. Biol. J. Linn. Soc. 32: 127-135.

(U.S. EPA) United States Environmental Protection Agency. 1997. Volunteer stream monitoring: a methods manual. EPA 841-B-97-003. Office of Water, Washington, DC.

(U.S. EPA) United States Environmental Protection Agency. 1999. Draft guidance for water quality-based decisions: the TMDL process, 2nd ed. EPA 841-D-99-001. Office of Water, Washington, DC.

Vinson, M. R., and C. P. Hawkins. 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among streams J. North Am. Benthol. Soc.15: 392-399.

Voshell, J. R., E. P. Smith, S. K. Evans, and M. Hudy. 1997. Effective and scientifically sound bioassessments: Opinions and corroboration from academe. Hum. Ecol. Risk Assess. 3: 941-954.

Sarah Engel received her bachelor's degree from The Ohio State University in environmental science. She went on to obtain her Master's degree fromVirginia Tech studying biomonitoring. Sarah is currently working for Environmental Services & Consulting, an environmental consulting firm based in Blacksburg, VA. **Reese Voshell** is a professor in the Department of Entomology, Virginia Tech, Blacksburg, VA 24060-0319, where he teaches courses in aquatic entomology and freshwater biomonitoring. He has received numerous grants to study the effects of pollution and environmental stress on freshwater invertebrates (e-mail: rvoshell@vt.edu).