J. N. Am. Benthol. Soc., 2008, 27(1):58–67

DOI: 10.1899/07-037R.1
Published online: 20 November 2007

Precision of benthic macroinvertebrate indicators of stream condition in Montana

James B. Stribling[1]
Center for Ecological Sciences, Tetra Tech, Inc., 400 Red Brook Blvd., Suite 200, Owings Mills, Maryland 21117-5159 USA

Benjamin K. Jessup[2]
Tetra Tech, Inc., 15 State St., Suite 301, Montpelier, Vermont 05602 USA

David L. Feldman[3]
Montana Department of Environmental Quality, 1520 East 6th Ave., Helena, Montana 59620-0901 USA

Abstract. The Montana Department of Environmental Quality (MDEQ) uses 2 forms of benthic macroinvertebrate indicators for detection of stream impairment, a multimetric index (MMI) and a predictive model of observed to expected taxa (O/E), each of which is calibrated to streams across the state. As part of the routine monitoring program, some sample locations were subjected to repeated sampling, i.e., multiple samples were collected from stream reaches in spatial and temporal proximity. Results from repeated sampling allow calculation of precision estimates, which are important for describing a portion of the uncertainty (systematic error) associated with field sampling and site assessments. In this project, we evaluated 131 and 77 repeated-sample pairs for the MMI and O/E, respectively, using 4 different measures of precision: coefficient of variability (CV), 90% confidence intervals, relative % difference (RPD), and % difference for the final assessments. MMI and O/E had similar consistency and repeatability. Segregating the data set and calculations by region or field method yielded generally similar precision estimates for the indicators, although precision was slightly better in the mountains using the Hess field-sampling method than in other regions or with other field methods. Evaluation of RPD showed that assessments (impaired/nonimpaired) on the basis of the MMI differed between samples in 18.3% of repeated-sample pairs and assessments on the basis of O/E differed between samples for 19.5% of repeated-sample pairs. Recommended measurement quality objectives were 10 to 15% for CV and 15 to 20% for RPD for both indicators. Field-sampling precision was the focus of our paper, but we emphasize that detecting the presence of stressors or degraded conditions is the primary objective of the MDEQ stream condition indicators.

Key words: data quality, MQO, performance characteristics, multimetric index, RIVPACS, O/E, field sampling, biological assessment, repeatability.

The Montana Department of Environmental Quality (MDEQ) recently implemented 2 indicators of stream biotic conditions on the basis of benthic macroinvertebrate samples collected throughout the state (Jessup et al. 2006). The indicators model the similarity of biotic conditions in individual samples to generalized conditions observed in environmentally similar sites that are known to have minimal human impacts (i.e., reference sites). The multimetric index (MMI) and the predictive model of observed and expected taxa (O/E) are used by state water-quality monitoring programs to assess the degree of human impact on a water body. The MMI is based on a framework developed for fishes and is now applied to different biological assemblages (Karr et al. 1986, Hughes et al. 1998, Barbour et al. 1999, Hill et al. 2000, 2003). The O/E model is based on the river invertebrate prediction and

[1] E-mail addresses: james.stribling@tetratech.com
[2] benjamin.jessup@tetratech.com
[3] dfeldman@mt.gov

classification system (RIVPACS; Clarke et al. 1996, 2003, Hawkins et al. 2000, Hawkins 2006).

Benthic macroinvertebrate-based biological assessments have multiple potential sources of variability, including those arising from field sampling, laboratory sample processing (sorting, subsampling, and taxonomic identifications), data entry, calculation of indicator variables, and site assessments (Narf et al. 1984, Diamond et al. 1996, Carter and Resh 2001, Cao et al. 2003, Clarke and Hering 2006, Clarke et al. 2006, Flotemersch et al. 2006, Herbst and Silldorf 2006). Performance characteristics are quantitative or qualitative standards of acceptable data quality for a method, protocol, data set, or program. They include factors such as precision, accuracy, bias, representativeness, completeness, and sensitivity (Diamond et al. 1996, Herbst and Silldorf 2006). Other aspects of performance (sorting/subsampling, taxonomy, and site assessments) for the data set have been evaluated elsewhere (BKJ, DLF, T. Laidlaw [US EPA], D. Stagliano [Montana Natural Heritage Program], and JBS, unpublished data). Our paper focuses strictly on precision of field-sampling methods and site assessments on the basis of the Montana MMI and O/E. Therefore, our estimates of precision are based on metrics and indices calculated for sites at which repeated samples were collected.

Sample content varies on the basis of taxon presence and the number of organisms in each taxon. The Montana MMI uses several individual metrics to describe sample content and includes counts or proportions of selected taxa, proportions of individuals in specific taxa or groups of taxa, the proportion of organisms in the sample possessing some autecological characteristic, or their relative tolerance to stressor conditions (Jessup et al. 2006). Values of each individual metric are placed on a 100-point scale (i.e., each metric potentially could receive a value from 0 to 100), and the final MMI is a mean of several metrics. The MMI also is calibrated on the basis of geographic regions in Montana (mountains, low valleys, and plains) and metric and index responses to stressors. The Montana O/E is based on the presence or absence of individual taxa and is calculated as the proportion of taxa observed in a test sample (O) to those expected to occur in a sample (E), with the latter term defined by reference samples (Hawkins 2006, Jessup et al. 2006). Values of O/E range from 0 to 1.5. Calculation of O/E accounts for geographic region by incorporating environmental predictor variables in the model.

How closely one sample resembles another sample collected from the same or adjacent reaches is a direct measure of repeatability, a form of precision (Narf et al. 1984, Diamond et al. 1996, Zar 1999, Cao et al. 2003). Narf et al. (1984) and Stark (1993) used a measure of precision (detectable difference [DD]) to evaluate the variability of the Hilsenhoff biotic index and the effect of different environmental variables on biological indicator values. Most biological monitoring programs seek to sample data with: 1) known precision and 2) precision adequate to detect real differences between samples. Precision can be used as a statement of measurement error associated with the entire data set if stream locations from which repeated samples are taken are randomly selected. The more similar sample data are when acquired from repeated samples, the better the precision and the lower the measurement error of the entire data set.

The accuracy of indicators and the validity of impairment thresholds are not affected by normal variability introduced by field sampling because indicator values and thresholds are developed during the calibration process using data for which that variability is inherent. Stated more simply, if samples used for indicator calibration meet performance standards, then direct comparison of the indicator value to the threshold or sample is valid. However, *precision* of the MMI and O/E indicator values are affected by field sampling, and the extent to which they are affected can be characterized using data from repeated samples (i.e., sample pairs) taken in spatial and temporal proximity. Moreover, each component metric of the MMI can be characterized with the same precision statistics used to describe the indicators.

Precision estimates of individual metrics, overall indices, and final site assessments are useful information for several reasons. Precision estimates: 1) help determine those metrics most responsive to stressors, 2) provide improved confidence in determining which metric(s) most influence the overall indicator score for particular samples, 3) quantify repeatability of field sampling activities in the context of the indicators, and 4) allow communication of a component of uncertainty associated with final site condition assessments. MDEQ is most concerned with the aspects of precision that improve reliability of the final assessments because the MMI and O/E are used in programmatic decisions related to impairment of water bodies. Our study was focused on the precision of the indicators, which is dependent on many factors in 2 basic categories: natural variability and measurement error. Natural variability of biotic samples is affected by the physical, chemical, hydrological, and biological characteristics of the stream reach. Measurement error is the variability introduced by field samplers, sample processors, taxonomists, data managers, and analysts.

We characterized the variability introduced by field sampling and based precision statistics on pairs of

samples collected in spatial and temporal proximity. We controlled the variability caused by natural conditions and other sampling and analysis processes as much as possible through routine quality-control activities and spatial and temporal stratification of field sampling and laboratory analyses. Nevertheless, some of this variability undoubtedly existed in our samples and had some unmeasured effect on our results.

We addressed 3 primary questions: 1) *Field-sampling precision*: If one indicator value is produced for a sample, how close is that value to expected mean values? 2) *Field-sampling precision*: What is the nearness of indicator values calculated for each sample in a repeated-sample pair gathered from the same stream reaches? 3) *Site-assessment precision*: How often do final site assessments on the basis of different samples collected from the same site differ? The purpose of our paper is to use several quantitative performance characteristics to present estimates of field-sampling and site-assessment precision for Montana stream-condition indicators. The techniques and statistics that we present are transferable to other data sets containing repeated-sample data and the concepts can help broaden our understanding of repeatability for indicators used by other biological monitoring programs.

Methods

Data reduction

We searched the MDEQ database of benthic macroinvertebrate samples collected from streams throughout Montana to identify replicate samples that were collected from the same site, on the same day, using the same sampling protocol. We then extracted the replicate assessment indicator results from the database and compiled supporting metadata, such as sampling protocol, model performance criteria, and site class, for each sample. We filtered the data set to exclude from further analysis those samples with <200 organisms, insufficient environmental data, or data indicating unusual conditions. For the O/E model, we excluded 108 samples from 54 sites because environmental characteristics of the site differed from any combination of characteristics used in model calibration (many of the sites lacked delineation of the drainage area). These filters resulted in 262 samples from 131 sites for MMI analysis, and 154 samples from 77 sites for O/E analysis. We represented each site with a single repeated-sample pair. We transferred all data to a spreadsheet for analysis.

The 131 sites used for MMI analysis spanned most geographic regions of Montana (Fig. 1), and were collected using 1 of 4 field-sampling protocols (see below; Table 1). The MMI is specific to the site class (mountains, low valleys, or plains) at the sampling location. Previous studies have indicated that differences among Montana sampling protocols have negligible effect on assessment results (BKJ, DLF, T. Laidlaw, D. Stagliano [Rhithron Associates], and JBS, unpublished data), but this information is still of interest. Therefore, we analyzed the data both as a pooled data set and within subsets defined by site class and method.

Most samples were obtained with the MDEQ traveling-kick or Hess field-sampling methods (Table 1). Other sampling methods are not as well represented in the database, and we chose not to report precision statistics generated from small sampling-method subsets separately because we considered them unreliable. Therefore, we treated data from Hess samples and Surber samples as having been collected by a single method because they covered similar substrate areas. We included data from samples collected using methods of the Science to Achieve Results (STAR; Hawkins et al. 2003) and the Environmental Monitoring and Assessment Program (EMAP; Lazorchak et al. 1998, Klemm et al. 2002) large river methods (Table 1) in our data set. Most sample sets consisted of 2 replicate samples (1 repeated-sample pair) per site. However, sample sets collected with Hess sampler in the Clark Fork project consisted of 4 replicate samples, so we randomly selected 2 of the 4 replicates to include in our data set.

Precision statistics

We measured precision in 2 forms: in the context of analysis of variance (ANOVA) for computational convenience and as relative % difference (RPD; Keith 1991, Berger et al. 1996, APHA 2005) to give direct comparisons of within-site proportional differences. We set up the ANOVA such that groups were defined as the sites from which repeated-sample pairs were collected. We used the mean squared error (MSE) term from ANOVA as an estimate of within-group variance (Zar 1999), and estimated standard deviation within groups as the square root of the MSE (root mean squared error [RMSE]). The RMSE is lower when measures are more precise. We used coefficient of variability (CV) to standardize variability of the data set to the mean value of the data set ($CV = 100[RMSE/mean]$) so that we could compare relative precision among indicators, metrics, and data sets (Diamond et al. 1996). We also used information about relative precision to determine which indices and metrics gave the most consistent signals and had high relative consistency in different situations (i.e., different data subsets).

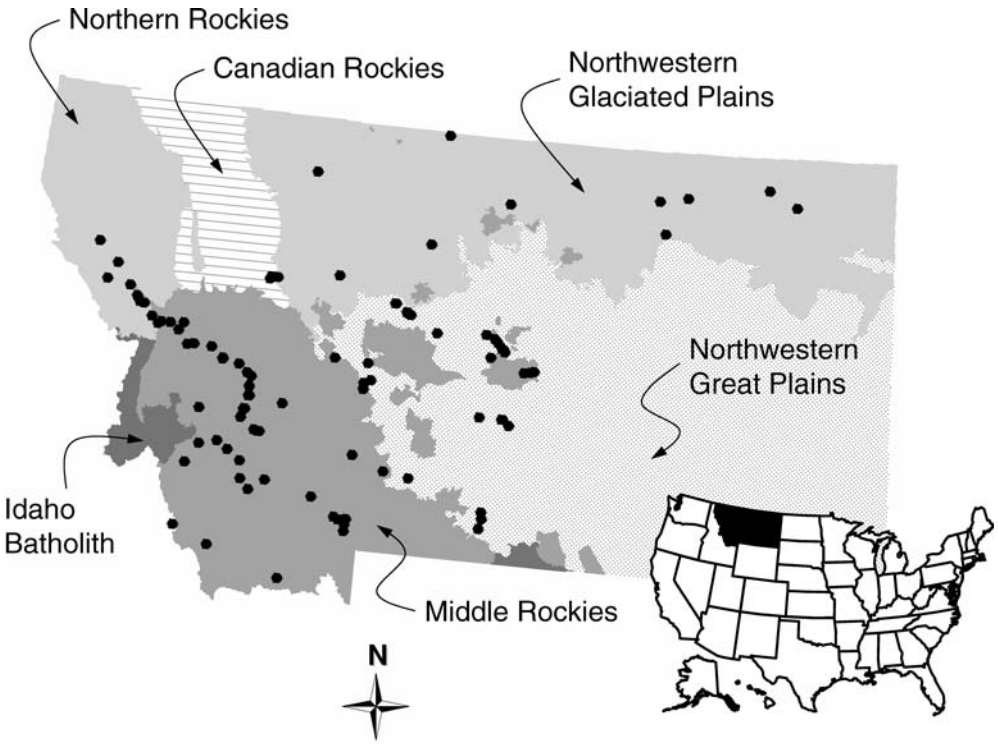

FIG. 1. Locations of 131 stream sites from which samples were analyzed for field sampling and site-assessment precision.

We used confidence intervals (CI) to indicate the magnitude of separation of 2 indicator values before the values could be considered different with statistical significance. We chose a 90% significance level for the CI (i.e., the range around the observed value within which the true mean is likely to fall 90% of the time, or a 10% probability of type I error [α]). We calculated CI90 from the RMSE using the equation:

TABLE 1. Description of lotic field-sampling methods used by Montana Department of Environmental Quality (MDEQ) for all samples used in our analyses. Sampling protocols were based on US Environmental Protection Agency (EPA) Science to Achieve Results (STAR), EPA Environmental Monitoring and Assessment Program—Large River (EMAP-LR), or MDEQ methods.

Protocol	Method	Gear	Mesh size (μm)	Subsample size (no. organisms)	Citation
Traveling kick	Traveling kick: kick for ≥1 min in a riffle, diagonally upstream and across channel; effort standardized by time	D-frame net	1200	300	MDEQ 2006a
Hess	4–8 samples in riffles, 0.61 m^2 total, composited	Hess sampler	1000	300	MDEQ 2006a
STAR	Fast-water substrate within each of eight 0.09-m^2 sampling frames (real or visualized; 2 in each of 4 different riffles) is disturbed with hands to a depth of ~10 cm, 0.72 m^2 total substrate area, composited	Surber, Surber-on-a-stick, or D-frame net	500	300	Hawkins et al. 2003
EMAP-LR	1 kick net sample (0.09 m^2) from a shallow area (<1 m deep) near the bank of the river at each of 11 transects (multiple habitats), alternating banks every 2 samples, 0.99 m^2 total substrate area, composited	D-frame net	500	300	Lazorchak et al. 1998, Klemm et al. 2002

TABLE 2. Statistics used to estimate precision of field-sampling methods and Montana indicator values (multimetric index [MMI] and ratio of observed to expected [O/E] taxa) using all available data (All) or partitioned by site class (mountains, low valleys, plains) or Montana Department of Environmental Quality sampling method (traveling kick [Kick], Hess). Statistics for the US Environmental Protection Agency Science to Achieve Results and Environmental Monitoring and Assessment Program—Large River are not shown because sample sizes were too small to give meaningful results for these subsets ($n = 2$ and 3, respectively). MSE = mean square error, RSME = square root MSE, CV = coefficient of variation, CI90 = 90% confidence interval for 1, 2, or 3 samples, n = number of repeated-sample pairs, n/a = not applicable.

								CI90		
Indicator	Site class	Method	n	MSE	RMSE	Mean	CV	1	2	3
MMI	All	All	131	23.0	4.80	54.4	8.8	7.87	5.56	4.54
	Mountains	All	40	14.4	3.80	55.6	6.8	6.23	4.41	3.60
	Low valleys	All	71	26.5	5.15	56.1	9.2	8.45	5.98	4.88
	Plains	All	20	28.5	5.34	46.2	11.5	8.76	6.19	5.06
	Mountains	Kick	21	18.8	4.33	59.1	7.3	7.11	5.03	4.10
	Low valleys	Kick	19	25.9	5.09	56.6	9.0	8.34	5.90	4.82
	Plains	Kick	15	30.5	5.53	40.6	13.6	9.06	6.41	5.23
	Mountains	Hess	19	9.6	3.10	51.8	6.0	5.09	3.60	2.94
	Low valleys	Hess	52	26.8	5.17	55.9	9.3	8.48	6.00	4.90
O/E	n/a	All	77	0.0126	0.112	0.83	13.4	0.18	0.13	0.10
	n/a	Kick	40	0.0153	0.124	0.91	13.6	0.20	0.14	0.12
	n/a	Hess	32	0.0071	0.084	0.73	11.6	0.14	0.10	0.08

$$\text{CI90} = \pm([\text{RMSE}][z\alpha])$$

where $z\alpha$ is the z value for 90% confidence (i.e., $p = 0.10$) with degrees of freedom set at ∞. In this analysis $z\alpha = 1.64$ (appendix 17 in Zar 1999). As the number of sample repeats increases, CI becomes narrower; we provide CI that would be associated with 1, 2, and 3 samples per site.

RPD is the proportional difference between 2 measures. We calculated RPD as:

$$\text{RPD} = \left(\frac{|A - B|}{(A + B)/2}\right)100$$

where A is the metric or index value of the 1st sample and B is the metric or index value of the 2nd sample (Keith 1991, Berger et al. 1996, APHA 2005). Lower RPD values indicate improved precision (as repeatability) over higher values.

We characterized variability of the final site assessment provided by the MMI or O/E on the basis of whether both samples in a sample pair receive the same rating (impaired or nonimpaired). If the final scores given to each sample in a pair fall on different sides of the decision threshold (in all cases, the 10th percentile of the reference distribution) then, by definition, the final ratings will be different. We quantified site-assessment precision as the proportion of repeated-sample pairs in which assessments of impairment or nonimpairment differed between samples, i.e., % disagreement. Site-assessment precision values potentially range from 0 to 100%, with lower values indicating greater agreement. We present the results in tabular form, pooled across the entire data set and partitioned by site class. We also present the results graphically by individual repeated-sample pairs.

The specific value for a measurement quality objective (MQO) is based on the distribution of values, in particular the minima and maxima, attained in a calibration (or pilot) study, such as ours. We selected MQO subjectively on the basis of our observations of the ranges of precision estimates. The MQO is a control point above (or below) which most observed values fall. Subsequent values exceeding the MQO are not automatically taken to be *unacceptable* data points; rather, such values should receive closer scrutiny to determine reasons for the exceedence and might indicate a need for corrective actions.

Results

The CV and CI90 of the MMI calculated across all site classes and methods were 8.8% and ±7.87 index units, respectively (Table 2). Among the data subsets that were large enough to analyze, index values were most precise for samples collected in the mountains with the Hess sampler (CV, 6.0%; CI90, ±5.09 index units; Table 2). Index values also were very precise for samples collected in the mountains with the traveling-kick method (CV, 7.3; CI90, ±7.11 index units). The most variable index values were obtained from samples collected in the plains with the traveling-kick method (CV, 13.6%; CI90, ±9.06 index units).

The overall CV and CI90 of O/E calculated across all

TABLE 3. Statistics used to estimate precision of component metrics of the Montana multimetric index (MMI) using data partitioned by site class (mountains, low valleys, plains). EPT = Ephemeroptera, Plecoptera, Trichoptera taxa, MSE = mean square error, RSME = square root MSE, CV = coefficient of variation, CI90 = 90% confidence interval for 1, 2, or 3 samples, n = number of repeated samples.

Metric	MSE	RMSE	Mean	CV	CI90		
					1	2	3
Mountains (n = 40)							
Ephemeroptera taxa	0.89	0.94	5.25	17.9	1.55	1.10	0.89
Plecoptera taxa	0.81	0.90	2.42	37.3	1.48	1.05	0.85
% EPT	78.45	8.86	47.98	18.5	14.53	10.27	8.39
% Noninsects	9.02	3.00	7.30	41.1	4.93	3.49	2.85
% Predators	28.26	5.32	16.91	31.4	8.72	6.17	5.03
% Burrower taxa	15.45	3.93	28.91	13.6	6.45	4.56	3.72
HBI	0.22	0.47	4.27	10.9	0.76	0.54	0.44
Low valleys (n = 71)							
% EPT (excluding Hydropsychidae or Baetidae)	93.87	9.69	23.08	42.0	15.89	11.24	9.17
% Chironomidae	48.05	6.93	20.56	33.7	11.37	8.04	6.56
% Crustacea and Mollusca	15.57	3.95	3.20	123.2	6.47	4.57	3.74
Shredder taxa	0.38	0.62	1.37	44.9	1.01	0.71	0.58
% Predators	44.34	6.66	13.09	50.9	10.92	7.72	6.30
Plains (n = 20)							
EPT taxa	1.67	1.29	10.30	12.5	2.12	1.50	1.22
% Tanypodinae	0.42	0.64	1.73	37.3	1.06	0.75	0.61
% Orthocladiinae of Chironomidae	186.30	13.65	46.94	29.1	22.38	15.83	12.92
Predator taxa	1.87	1.37	5.44	25.1	2.24	1.58	1.29
% Filterers–collectors	55.74	7.47	78.78	9.5	12.24	8.65	7.07

sampling methods were 13.4% and ±0.18 index units, respectively (Table 2). Data subsets were differentiated on the basis only of sampling method because site class was taken into account during calculation of the indicator. Fewer sample sets were used in analysis of O/E precision because necessary environmental information was not available for all sites. Among methods, indicator values for samples collected with the Hess sampler were more precise (CV, 11.6%; CI90, ±0.14 index units) than for samples collected with the traveling-kick method (CV, 13.6%; CI90, ±0.20 index units).

The CVs for all component metrics of MMIs were higher than the CVs of the respective MMIs for all site class data subsets (Table 3). The only exception was a single metric (% filterers–collectors) in the plains data subset (Table 3). The most precise metrics (on the basis of CV) were the HBI (10.9%; mountains), % Chironomidae (33.7%; low valleys), and % filterers–collectors (9.5%; plains). The least precise metrics (greatest variability) among site classes were for samples in the low valleys, where CV for all metrics were >30%. Metrics with the largest CVs were % noninsects (41.1%; mountains), % Crustacea and Mollusca (123.2%; low valleys), and % Tanypodinae (37.3%; plains).

For the MMI, values for RPD on individual sample pairs ranged from 0.08 to 49.8% (the mean across all sampling methods = 10.5% [Table 4]). RPDs typically were lowest (better repeatability) for samples collected in the mountains (<~10%), intermediate for samples collected in the low valleys (~10–11%), and highest for samples collected in the plains (14–15%). RPDs of O/Es ranged from 0 to 50 (mean across all sampling methods = 15.3%; Table 4). O/Es for samples collected using Hess field-sampling methods were more repeatable (mean RPD = 13.9%) than O/Es for samples collected using traveling-kick methods (mean RPD = 15.9%).

Percent disagreement for assessments on the basis of the MMI calculated across all site classes and methods was 18.3% (Table 5, Fig. 2). That is, assessments of both samples in a repeated-sample pair agreed in 81.7% of the repeated-sample pairs. The highest % disagreement for assessments on the basis of the MMI for a site class was 22.5% (mountains). Percent disagreement for assessments on the basis of the O/E calculated across all methods was 19.5%, only slightly worse than the overall MMI (Table 5, Fig. 2).

Discussion

We calculated 3 precision estimates (CV, CI90, RPD) to describe the variability of the Montana biological

TABLE 4. Mean relative % difference (RPD) statistic used to estimate precision (repeatability) of Montana indicator values (multimetric index [MMI] and ratio of observed to expected [O/E] taxa) using all available data (All) or partitioned by site class (mountains, low valleys, plains) or Montana Department of Environmental Quality sampling method (traveling kick [Kick], Hess). Statistics for the US Environmental Protection Agency Science to Achieve Results and Environmental Monitoring and Assessment Program—Large River methods are not shown because sample sizes were too small to give meaningful results for these subsets (n = 2 and 3, respectively). n = number of repeated-sample pairs, n/a = not applicable.

Indicator	Site class	Method	n	RPD
MMI	All	All	131	10.5
	Mountains	All	40	8.0
	Low valleys	All	71	10.9
	Plains	All	20	14.0
	Mountains	Kick	21	10.0
	Low valleys	Kick	19	10.5
	Plains	Kick	15	14.8
	Mountains	Hess	19	5.8
	Low valleys	Hess	52	11
O/E	n/a	All	77	15.3
	n/a	Kick	40	15.9
	n/a	Hess	32	13.9

indicators (MMI and O/E). These precision estimates can be used to: 1) evaluate the differences in indicator values between and among multiple samples, 2) document consistency of field sampling and assessments, 3) establish MQOs, and 4) evaluate whether a field team, data set, or the overall program continues to meet the MQOs in future sampling and analysis.

The most precise indicator evaluated was the MMI (Table 2), yet the O/E seemed to be more stable among the different field-sampling protocols. The types of precision estimates we calculated have been used to evaluate other data sets. For example, Narf et al. (1984) calculated DD for the HBI in 42 Wisconsin stream locations and used it in the same way we used CI90 for the Montana data. RMSE and DD were equivalent to 3.9 and 8.4%, respectively, of the range of HBI values calculated for the Wisconsin data set (Narf et al. 1984). RMSE and CI90 were equivalent to 4.7 and 7.6%, respectively, of the range of HBI values calculated for the Montana data set. We took the same approach but extended it to all component metrics and the overall MMI.

Precision estimates can and should be used to describe the uncertainty associated with field sampling. However, laboratory processing of benthic macroinvertebrate samples also affects the quality of the data obtained from field sampling. The very nature of sampling and analysis for biological assessments means that the quality of data obtained from field sampling cannot be evaluated until sample processing has been completed. Thus, investigators must recognize the implicit assumption that other factors that contribute to data quality, such as sorting/subsampling and taxonomy, are acceptable when attributing variability in indices to field-sampling sources of error (i.e., that type I and II errors are not related to laboratory processing error).

Assessments on the basis of samples separated by space or time would be considered truly different from each other if the difference between their indicator values exceeded the CI90. The CI90s for the data set considered in our study are similar to those arrived at earlier (Jessup et al. 2006; Table 6). These rates of error (roughly 1 in 5) should be of little concern as long as indicator values for samples in repeated-sample pairs do not fall near the 10^{th} percentile decision threshold (of the reference distribution) (Fig. 2). However, if indicator values for repeated samples are close to the threshold *and* fall on either side of the threshold, then other factors should be used to communicate the

TABLE 5. Site-assessment precision matrices showing the number of cases in which assessments (impaired, nonimpaired) based on the Montana indicator values (multimetric index [MMI] and ratio of observed to expected [O/E] taxa) of individual samples in repeated-sample pairs agreed (bold font) or disagreed. MMI data were analyzed using all (All) available data or data partitioned by site class (low valleys, mountains, plains). n = number of repeated-sample pairs.

Indicator	Site class	n	Assessment	Nonimpaired	Impaired	% disagreement
MMI	All	131	Nonimpaired	**70**	—	
			Impaired	24	**37**	18.3
	Low valleys	71	Nonimpaired	**48**	—	
			Impaired	12	**11**	16.9
	Mountains	40	Nonimpaired	**11**	—	
			Impaired	9	**20**	22.5
	Plains	20	Nonimpaired	**11**	—	
			Impaired	3	**6**	15.0
O/E	All	77	Nonimpaired	**38**	—	
			Impaired	15	**24**	19.5

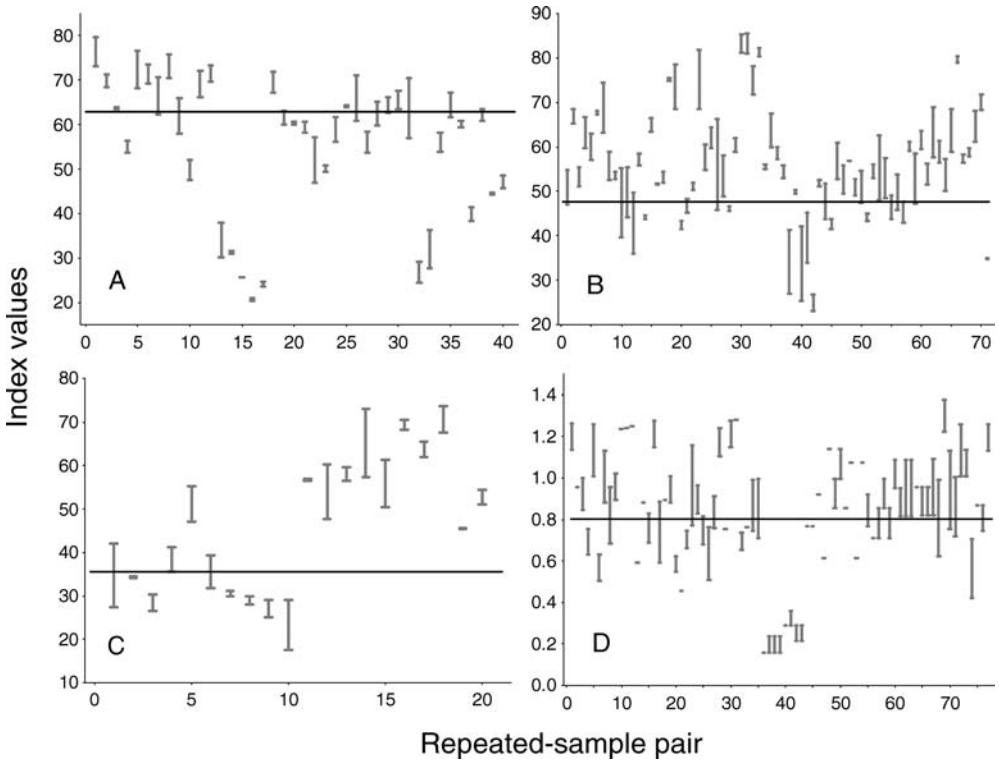

FIG. 2. Montana multimetric index (MMI) and ratio of observed to expected (O/E) taxa index values calculated for individual samples within repeated-sample pairs. The end bars of each vertical line indicate the sample scores for each pair. The length of the vertical bar indicates the degree of dissimilarity between each pair of scores. If no vertical line is visible, the index values were identical for both samples in the pair. The horizontal line represents the impaired/nonimpaired decision threshold (10th percentile of reference values). Samples connected by vertical bars that cross the threshold have different impairment assessments. A.—MMI values for samples collected from sites in the mountains site class. B.—MMI values for samples collected from sites in the low valleys site class. C.—MMI values for samples collected from sites in the plains site class. D.—O/E values for samples collected at all sites.

condition of the site following the MDEQ sufficient credible data/beneficial use determination approach and macroinvertebrate collection standard operating procedures (Norton et al. 2002, Suter et al. 2002, MDEQ 2006a, b).

The purpose of a CI is to enhance comparison of observed indicator values to other values or decision thresholds. CIs should never be used to justify assignment of a value other than the specific observed value, nor should CIs be used to question the position of a value relative to a threshold or another point. Rather, the appropriate use of the CI is expression of uncertainty associated with an observation. If the CI contains a threshold of concern, then individual streams in the data set might be targeted for continued or more intensive assessments using additional, ancillary information.

Resource managers might conclude from analyses such as ours that precision estimates (error rates) are unacceptably high. The width of the CI associated with a site could be decreased by collecting additional replicate samples (Table 3). For example, doubling or tripling the number of replicate samples can result in substantial increases in certainty that an observed value for a site is the true mean for the indicator. However, improving precision increases costs because it requires increased field-sampling effort (more samples or samples that cover more surface area) or more time in the laboratory (e.g., larger fixed-count subsamples [>300]).

The variability of the Montana data reported here and in previous studies (Jessup et al. 2006) should be

TABLE 6. Comparison of 90% confidence intervals (CI90) for values of the Montana multimetric index (MMI) obtained in our study and by Jessup et al. (2006) for data pooled by site class.

Site class	Our study	Jessup et al. (2006)
Mountains	5.7	6.9
Low valleys	7.5	8.4
Plains	8.5	9.6

TABLE 7. Recommended measurement quality objectives for field-sampling precision on the basis of coefficient of variation (CV) and relative % difference (RPD) statistics on the basis of the Montana indicator values (multimetric index [MMI] and ratio of observed to expected [O/E] taxa) of individual samples in repeated-sample pairs. MMI data were analyzed using all (All) available data or data pooled by site class (mountains, low valleys, plains).

Model	Site class	CV	RPD
MMI	All	10	15
MMI	Mountains	10	15
MMI	Low valleys	10	15
MMI	Plains	15	20
O/E	n/a	15	20

taken as the precision performance of the MMI and O/E for field sampling and site assessment. We used our analyses of variation around the mean (CV) and of repeatability of field sampling (RPD) to develop MQOs for the MDEQ (Table 7). Estimates of precision that exceed these MQOs signal potential problems with data quality and the need for corrective actions. Examples of potential corrective actions include additional training for personnel engaged in sampling and laboratory activities, field and laboratory audits, or increased scrutiny of field and laboratory quality-control data. However, the primary goal associated with use of both Montana stream-condition indicators is to detect stressors or stressed conditions. Although consistency and repeatability are critical indicator characteristics, sustained use of the indicators by MDEQ is unlikely if they do not contribute to resource-management decisions.

Acknowledgements

This project was completed under MDEQ contract to Tetra Tech, Inc. (Owings Mills, Maryland) (task no. 206014, term contract reference no. SPB05-894P-BB). Under a separate contract, C. P. Hawkins (Utah State University, Logan, Utah), Tina Laidlaw (US Environmental Protection Agency/Region 8, Helena, Montana), and Mike Suplee, Mark Bostrom, and Bob Bukantis (MDEQ, Helena, Montana) collaborated and provided input for calibration and application of the MMI and O/E models. We thank Jerry Diamond and Jeroen Gerritsen of Tetra Tech, Mark Bostrom, and Bob Bukantis for review and comment on earlier drafts of this manuscript. Dan McGuire did most of the sample collections for the Clark Fork project. Bruce Chessman, Pamela Silver, and 2 anonymous referees provided comments that helped improve the manuscript.

Literature Cited

APHA (AMERICAN PUBLIC HEALTH ASSOCIATION). 2005. Standard methods for the examination of water and wastewater. 21st edition. American Public Health Association, American Water Works Association, and Water Environment Federation, Washington, DC.

BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Revision to the rapid bioassessment protocols for streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. EPA/841-D-97-002. Office of Water, US Environmental Protection Agency, Washington, DC.

BERGER, W., H. MCCARTY, AND R. K. SMITH. 1996. Environmental laboratory data evaluation. Genium, Amsterdam, New York.

CAO, Y., C. P. HAWKINS, AND M. R. VINSON. 2003. Measuring and controlling data quality in biological assemblage surveys with special reference to stream benthic macroinvertebrates. Freshwater Biology 48:1898–1911.

CARTER, J. L., AND V. H. RESH. 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. Journal of the North American Benthological Society 20:658–682.

CLARKE, R. T., M. T. FURSE, J. F. WRIGHT, AND D. MOSS. 1996. Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. Journal of Applied Statistics 23:311–332.

CLARKE, R. T., AND D. HERING. 2006. Errors and uncertainty in bioassessment methods—major results and conclusions from the STAR project and their application using STARBUGS. Hydrobiologia 566:433–439.

CLARKE, R. T., A. LORENZ, L. SANDIN, A. SCHMIDT-KLOIBER, J. STRACKBEIN, N. T. KNEEBONE, AND P. HAASE. 2006. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. Hydrobiologia 566:441–459.

CLARKE, R. T., J. F. WRIGHT, AND M. T. FURSE. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modeling 160:219–233.

DIAMOND, J. M., M. T. BARBOUR, AND J. B. STRIBLING. 1996. Characterizing and comparing bioassessment approaches and their results: a perspective. Journal of the North American Benthological Society 15:713–727.

FLOTEMERSCH, J. E., J. B. STRIBLING, AND M. J. PAUL. 2006. Concepts and approaches for the bioassessment of nonwadeable streams and rivers. EPA/600/R-06/127. Office of Research and Development, US Environmental Protection Agency, Cincinnati, Ohio.

HAWKINS, C. P. 2006. Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional- and global-scale assessments. Ecological Applications 16:1277–1294.

HAWKINS, C. P., R. H. NORRIS, J. N. HOGUE, AND J. W. FEMINELLA. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456–1477.

HAWKINS, C. P., J. OSTERMILLER, M. VINSON, R. J. STEVENSON, AND J. OLSON. 2003. Stream algae, invertebrate, and environmental sampling associated with biological water quality assessments: field protocols. Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, Utah State University, Logan, Utah. (Available from: http://129.123.10.240/WMCPortal/downloads/USU_field_protocols_9Jun2003.pdf).

HERBST, D. B., AND E. L. SILLDORF. 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. Journal of the North American Benthological Society 25:513–530.

HILL, B. H., A. T. HERLIHY, P. R. KAUFMANN, S. J. DECELLES, AND M. A. VANDER BORGH. 2003. Assessment of streams of the eastern United States using a periphyton index of biotic integrity. Ecological Indicators 2:325–338.

HILL, B. H., A. T. HERLIHY, P. R. KAUFMANN, R. J. STEVENSON, F. H. MCCORMICK, AND C. B. JOHNSON. 2000. Use of periphyton assemblage data as an index of biotic integrity. Journal of the North American Benthological Society 19:50–67.

HUGHES, R. M., P. R. KAUFMANN, A. T. HERLIHY, T. M. KINCAID, L. REYNOLDS, AND D. P. LARSEN. 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences 55:1618–1631.

JESSUP, B. K., C. HAWKINS, AND J. B. STRIBLING. 2006. Biological indicators of stream condition in Montana using benthic macroinvertebrates. Prepared by Tetra Tech, Inc., Owings Mills, Maryland and Utah State University, Logan, Utah, for the Department of Environmental Quality, Helena, Montana. (Available from: http://www.deq.state.mt.us/wqinfo/Standards/Montana%20Indicators%20Report%20(FINALcomb_061004).pdf)

KARR, J. R., K. D. FAUSCH, P. L. ANGERMEIER, P. R. YANT, AND I. J. SCHLOSSER. 1986. Assessing biological integrity in running waters: a method and its rationale. Special publication 5. Illinois Natural History Survey, Champaign, Illinois.

KEITH, L. H. 1991. Environmental sampling and analysis. A practical guide. Lewis Publishers, Chelsea, Michigan.

KLEMM, D. J., J. M. LAZORCHAK, AND P. A. LEWIS. 2002 (*Unpublished draft*). Benthic macroinvertebrates (Revision 2 [April 2002]). Section 9 *in* D. V. Peck, D. K. Averill, A. T. Herlihy, B. H. Hill, R. M. Hughes, P. R. Kaufmann, D. J. Klemm, J. M. Lazorchak, F. H. McCormick, S. A. Peterson, P. L. Ringold, M. R. Cappaert, T. Magee, and P. A. Monaco. Environmental monitoring and assessment program—surface waters: western pilot study field operations manual for non-wadeable rivers and streams. Office of Research and Development, US Environmental Protection Agency, Washington, DC. (Available from: US EPA National Health and Environmental Effects, Research Lab/ORD Western Ecology Division, 200 S.W. 35th Street, Corvallis, Oregon 97333-4902 USA.)

LAZORCHAK, J. M., D. J. KLEMM, AND D. V. PECK (EDITORS). 1998. Environmental monitoring and assessment program—surface waters: field operations and methods for measuring the ecological condition of wadeable streams. EPA/620/R-94/004F. Office of Research and Development, US Environmental Protection Agency, Washington, DC.

MDEQ (MONTANA DEPARTMENT OF ENVIRONMENTAL QUALITY). 2006a. Sample collection, sorting, and taxonomic identification of benthic macroinvertebrates. Standard operation procedure WQPBWQM-009. Revision no. 2. Water Quality Planning Bureau, Montana Department of Environmental Quality, Helena, Montana. (Available from: http://www.deq.mt.gov/wqinfo/QAProgram/WQPBWQM-009rev2_final_web.pdf)

MDEQ (MONTANA DEPARTMENT OF ENVIRONMENTAL QUALITY). 2006b. Water quality assessment process and methods. Appendix A to 303(d) 2000–2004. Standard operation procedure WQPBWQM-001. Revision no. 2. Water Quality Planning Bureau, Montana Department of Environmental Quality, Helena, Montana. (Available from: http://www.deq.mt.gov/wqinfo/QAProgram/SOP%20WQPBWQM-001.pdf)

NARF, R. P., E. L. LANGE, AND R. C. WILDMAN. 1984. Statistical procedures for applying Hilsenhoff's Biotic Index. Journal of Freshwater Ecology 2:441–448.

NORTON, S. B., S. M. CORMIER, G. W. SUTER, B. SUBRAMANIAN, E. LIN, D. ALTFATHER, AND B. COUNTS. 2002. Determining probable cause of ecological impairment in the Little Scioto River, Ohio, USA. Part 1. Listing candidate causes and analyzing evidence. Environmental Toxicology and Chemistry 21:1112–1124.

STARK, J. D. 1993. Performance of the macroinvertebrate community index: effects of sampling method, sample replication, water depth, current velocity, and substratum on index values. New Zealand Journal of Marine and Freshwater Research 27:463–478.

SUTER, G. W., S. B. NORTON, AND S. M. CORMIER. 2002. A methodology for inferring the causes of observed impairments in aquatic ecosystems. Environmental Toxicology and Chemistry 21:1101–1111.

ZAR, J. H. 1999. Biostatistical analysis. 4th edition. Prentice/Hall, Upper Saddle River, New Jersey.

Received: 9 April 2007
Accepted: 8 October 2007