J. N. Am. Benthol. Soc., 2008, 27(4):906–919
© 2008 by The North American Benthological Society
DOI: 10.1899/07-175.1
Published online: 28 October 2008

Data quality, performance, and uncertainty in taxonomic identification for biological assessments

James B. Stribling[1] AND **Kristen L. Pavlik**[2]
Tetra Tech, Inc., 400 Red Brook Blvd., Suite 200, Owings Mills, Maryland 21117-5159 USA

Susan M. Holdsworth[3]
Office of Wetlands, Oceans, and Watersheds, US Environmental Protection Agency, 1200 Pennsylvania Ave., NW, Mail Code 4503T, Washington, DC 20460 USA

Erik W. Leppo[4]
Tetra Tech, Inc., 400 Red Brook Blvd., Suite 200, Owings Mills, Maryland 21117-5159 USA

Abstract. Taxonomic identifications are central to biological assessment; thus, documenting and reporting uncertainty associated with identifications is critical. The presumption that comparable results would be obtained, regardless of which or how many taxonomists were used to identify samples, lies at the core of any assessment. As part of a national survey of streams, 741 benthic macroinvertebrate samples were collected throughout the eastern USA, subsampled in laboratories to ~500 organisms/sample, and sent to taxonomists for identification and enumeration. Primary identifications were done by 25 taxonomists in 8 laboratories. For each laboratory, ~10% of the samples were randomly selected for quality control (QC) reidentification and sent to an independent taxonomist in a separate laboratory (total $n = 74$), and the 2 sets of results were compared directly. The results of the sample-based comparisons were summarized as % taxonomic disagreement (PTD) and % difference in enumeration (PDE). Across the set of QC samples, mean values of PTD and PDE were ~21 and 2.6%, respectively. The primary and QC taxonomists interacted via detailed reconciliation conference calls after initial results were obtained, and specific corrective actions were implemented (if needed) prior to a 2nd round of comparisons. This process improved consistency (PTD = 14%). Corrective actions reduced the proportion of samples that failed the measurement quality objective for PTD from 71 to 27%. Detailed comparisons of results for individual taxa and interpretation of the potential causes for differences provided direction for addressing problematic taxa, differential expertise among multiple taxonomists, and data entry and recording errors. The taxa that proved most difficult (i.e., had high rates of errors) included many Baetidae, Odonata, Ceratopogonidae, selected groups of Chironomidae, and some Hydropsychidae. We emphasize the importance of experience and training and recommend approaches for improving taxonomic consistency, including documentation of standard procedures, taxonomic data quality standards, and routine and rigorous quality control evaluations.

Key words: bioassessment, taxonomy, uncertainty, data quality, precision, accuracy, performance, quality assurance/quality control.

Current biological monitoring and assessment programs use regionally calibrated assemblage-level biological indicators to document the status and trends of water resources. Indicators most often take the form of a multimetric index (Karr et al. 1986, Hughes et al. 1998, Barbour et al. 1999, Hill et al. 2000, 2003) or a predictive model based on the River Invertebrate Prediction and Classification System (RIVPACS; Clarke et al. 1996, 2003, Hawkins et al. 2000, Hawkins 2006). Beyond study design and field sampling protocols, the foundation of any of these indicators is a description of sample content, that is, the identification and enumeration of organisms in the sample.

[1] E-mail addresses: james.stribling@tetratech.com
[2] kristen.pavlik@tetratech.com
[3] holdsworth.susan@epa.gov
[4] erik.leppo@tetratech.com

For many laboratories that do or contribute to biological monitoring, taxonomic error rates are unknown and efforts to document them can be haphazard. This situation is a result of traditional reliance on expert opinion (Dines and Murray-Bligh 2000) and lack of clarity concerning the need for such information. However, greater attention is being given to environmental protection and improvement, protection of biodiversity, and assuring sustainability of development and other human activities, and, thus, interest in understanding the uncertainty associated with biological assessments is increasing (Clarke 2000, Dines and Murray-Bligh 2000, Moulton et al. 2000, Cao et al. 2003, Haase et al. 2006), including efforts to document the quality of taxonomic data.

Conceptually, any quality control (QC) activity beyond standard procedures has 3 parts. First, documenting error rates and sources for the activity of interest establishes a data-quality (or performance) baseline, helps to determine acceptability of the data, and allows development of necessary corrective actions. Second, results of QC analyses can be used to determine effects of error on ultimate uses of the data (Cao et al. 2003, Yuan 2007) and can be used to inform decision making on the acceptability of different error rates. Third, data quality can be monitored routinely over time to track error rates and allow performance evaluations of monitoring programs, laboratories, or individual staff. Continuous programs of taxonomic QC lead to demonstrable reductions in error rates (Haase et al. 2006) regardless of how they are interpreted as affecting uses of the data.

Data quality is defined as "the magnitude of error associated with a particular dataset" (Keith 1988, Peters 1988). Error in taxonomic identification is application of incorrect nomenclature to a specimen, and the error rate is the frequency of that occurrence (Klein 2001, Dalcin 2004, Haase et al. 2006) within a sample and within a data set. Taxonomic error can have several causes, including incorrect interpretation of technical literature; transcription or recording errors; coarse definitions of terminology, nomenclature, and standard procedures; differences in optical equipment; and sample handling and preparation techniques (Stribling et al. 2003, Dalcin 2004, Chapman 2005).

The ability to describe the uncertainty associated with the use of nonresearch taxonomists (i.e., production taxonomists [Stribling et al. 2003] or parataxonomists [New 1996, Smith et al. 2005]) to identify and enumerate specimens from large, multitaxon samples is critical when the goal is to describe the number of individuals attributed to each taxon in a sample. Two distinct approaches are used to describe such uncertainty. The 1st approach relies on confirming the identities of individual specimens, and the 2nd approach is to replicate whole-sample taxonomy, including enumeration.

Traditional taxonomic QC focuses on whether the name put on a particular specimen is correct, i.e., whether the specimen adequately matches truth (Stribling et al. 2003), or some specified gold standard. Taxonomic truth can take several forms, including: 1) a type specimen or a specimen from a type series; 2) a reference specimen that has been compared directly to types; 3) a reference specimen that has been verified by a specialist in that particular taxonomic group; 4) peer-reviewed technical literature, including accepted dichotomous identification keys describing diagnostic characteristics or the original description; or 5) DNA barcode or other genetic fingerprint. How a specialist is defined can vary for different taxa, but criteria include combinations of research experience and education, quality of peer-reviewed publications on the taxon of question, professional relationships to relevant research institutions, such as museums or universities, and respect of peers on the subject matter. Any of these approaches to attaining taxonomic truth is functionally an evaluation of *taxonomic accuracy*. Assigned names that do not match the analytical truth are considered errors.

It is important to understand how well taxonomic treatment reflects sample content because samples are the basis of biological assessments and are used to characterize ecological sites. The nearness of 2 measurements made of the same sample by independent taxonomists communicates how consistently each organism in the sample is identified; this nearness is sample-based *taxonomic precision* (Stribling et al. 2003). A key assumption in this process is that the likelihood is minimal that 2 taxonomists looking at a specimen would both be incorrect, and thus, taxonomic precision directly reflects identification error rate. Operationally, it is not critical to specify which of 2 different names might be correct. The important point is to attempt to understand what might be causing them to be different. Part of the evaluation process is to use that information to recommend potential corrective actions.

Background on the national Wadeable Streams Assessment

In 2004, the US Environmental Protection Agency (EPA) Offices of Research and Development and of Wetlands, Oceans, and Watersheds began developing the first national survey of the conditions of water resources of the contiguous US (Paulsen et al. 2008, Shapiro et al. 2008). The national Wadeable Streams Assessment (WSA) focused on wadeable, freshwater

TABLE 1. Total number of samples identified, and number of quality control (QC) samples reidentified in rounds 1 and 2, for each of 8 laboratories.

	Number of samples		
Laboratory	Total	Round 1	Round 2
A	153	18	14
B	229	20	25
D	123	12	12
E	25	3	3
F	18	4	3
H	131	10	12
I	24	3	2
J	38	4	4
Total	741	74	75

streams and used benthic macroinvertebrates as the biological indicator assemblage. WSA included a broad collaboration among the EPA, state environmental and natural resource agencies, other federal agencies, several universities and other organizations, >150 field biologists, and 25 taxonomists in 8 laboratories (USEPA 2006). The principal objective of the WSA was to produce a statistically valid answer to the question: What is the condition of US streams? (Paulsen et al. 2008). The intent of the EPA also was to have the primary taxonomic data available through the storage and retrieval (STORET; http://www.epa.gov/storet/) database for potential secondary uses, such as evaluation of geographic distributions, stressor and stressor-source diagnoses, and risk analyses.

We designed and implemented the QC process that allowed documentation of error associated with the taxonomic data of the WSA. The activities evaluated were specifically identification and counting. We used information on taxonomic identification performance and consistency to target specific corrective actions intended to reduce rates of error and to demonstrate the effects of error on documentation of presence/absence and relative abundances of taxa, the variability of metric and index values, and the consistency of final-condition narrative assessments. The purpose of our paper is to describe the results of an interlaboratory comparison and to document issues of taxonomic data quality and uncertainty. The results have implications for eventual implementation of a routine process useful for minimizing error and optimizing consistency in taxonomic data sets.

Methods

Field sampling and laboratory sorting and processing

We will not discuss field sampling and laboratory preprocessing methods (sorting and subsampling) in detail because our paper is focused strictly on the quality of taxonomic data. We review these methods briefly to provide context for consideration of the sample characteristics.

Field sampling.—The field sampling method was based on that of the EPA Environmental Monitoring and Assessment Program (EMAP) (Klemm et al. 1998, USEPA 2004b). At each site, samples were collected along 11 transects from multiple habitats with a D-frame net with 500-μm mesh openings. Transects were evenly distributed along a sampling reach length that was 40× wetted width of the channel. Organic and inorganic sample material (leaf litter, small woody twigs, silt, sand, and small gravel) was composited in containers, preserved with 95% denatured ethanol, and delivered to multiple laboratories for processing. A total of 741 samples were distributed to 9 laboratories for sorting; 1 of these laboratories did not do taxonomic identifications. Samples were primarily from streams of the eastern US that were not sampled as part of the EMAP Western Pilot Study (EMAP-West; Stoddard et al. 2005).

Sorting and subsampling.—Laboratory subsampling was done with a Caton gridded screen (Barbour et al. 1999, USEPA 2004a, Flotemersch et al. 2006) to a fixed count of 500 organisms. Samples were used in data analyses if they contained ≥300 organisms. Hereafter, all uses of *sample* refer to fixed-count subsamples.

Taxonomic identification and enumeration

Twenty-five taxonomists distributed among 8 laboratories (presented anonymously for purposes of our paper) did taxonomic identifications of the 741 samples (Table 1); laboratory capacity and taxonomic expertise dictated the number of samples assigned to each laboratory. Standard operating procedures (SOP) for the taxonomic identifications were provided to all laboratories and taxonomists (USEPA 2004a). All taxonomists were provided guidance on the kinds of biological material that should not be counted, e.g., exuviae, damaged specimens lacking head and most of thorax, oligochaete fragments without heads, mollusk shells not containing soft tissue, or taxa such as nematodes and copepods. Target taxonomic hierarchical levels were specified (primarily genus-level with a few family and genus-group targets; Table 2) based on known nomenclatural stability and general availability of technical literature, such as keys and diagnoses. In addition to these guidelines, taxonomists were instructed to use the magnification or specimen-handling technique necessary to assign target-level names with confidence and to use standardized data sheets.

Some taxonomists used morphotyping rather than

TABLE 2. Taxonomic level for Wadeable Stream Assessment (WSA) benthic macroinvertebrate identifications for which the target taxonomic level was not genus. The target taxonomic level for all other taxa was genus.

Taxon	Target
Phylum Annelida	
Class Oligochaeta	Family
Class Polychaeta	Family
Phylum Arthropoda	
Class Arachnida	
Subclass Acari	Family
Class Insecta	
Chironomidae	Genus, except certain genus groups and other complexes
Dolichopodidae	Family
Phoridae	Family
Scathophagidae	Family
Syrphidae	Family
Phylum Mollusca	
Class Gastropoda	
Hydrobiidae	Family

clearing and slide mounting all specimens of Chironomidae, but morphotyping was not specifically listed as an option in the project SOP. Morphotyping is a technique whereby similar specimens are grouped based on their appearance under dissecting microscopes, and then several specimens from each group are selected for slide mounting and identification with higher magnification from a compound microscope. If all mounted specimens from a group are identified as the same taxon with higher magnification, then the name is extrapolated to the unmounted specimens in the group. The primary taxonomists were informed that the QC taxonomist would mount any specimens they did not and, thus, would examine all chironomids on slides.

QC and documentation of data quality

The overall QC procedure was reidentification of a 10% subset of the samples identified by the 25 primary taxonomists (T1) by an independent QC taxonomist (T2), who was external to any of the primary laboratories, quantification of the magnitude and types of disagreements between the 2 sets of results, followed by a 2nd round of identification and reidentification if needed (Fig. 1). The *sample lot* is the full set of 500-organism samples originally identified by the primary laboratories and subjected to laboratory- and taxonomist-specific corrective actions. All taxonomists in both rounds followed identical procedures for target taxonomic hierarchical levels, counting rules, taxonomic comparisons, and reconciliation conference calls.

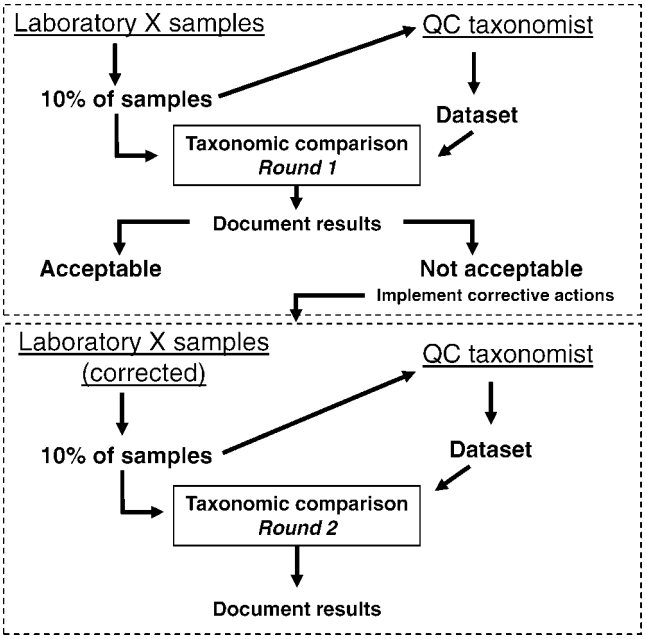

FIG. 1. Flow chart showing the process by which taxonomic results were compared among laboratories, performance results were documented, and, if necessary, corrective actions were developed and implemented. In this example, samples were 500-organism subsamples. Primary taxonomists from laboratory X identified the organisms in the samples, and 10% of these samples were randomly chosen for comparison with taxonomic results obtained by a quality control (QC) taxonomist. During the taxonomic comparison exercise, joint discussions could have led either taxonomist (primary or QC) to change identification decisions.

Round 1.—For each laboratory, 10% of the planned total number of samples already identified by T1 were randomly selected for reidentification, and all vials and slides were sent to T2 (Table 1). The same T2 taxonomist was used for all QC reidentifications for all laboratories in both rounds of identification and reidentification (see *Round 2* below). T2 was not associated with the laboratories conducting the primary identifications and used the same procedures for target taxonomic hierarchical levels and counting rules as T1. T2 also was instructed to use the magnification or specimen-handling technique necessary for confident assignment of target-level names. We compared results obtained by T1 and T2 to calculate performance measures and used interpretations of these statistics to develop recommended corrective actions that were communicated to all primary laboratories.

Reconciliation conference call.—Spreadsheet results (side-by-side sample results and calculated performance measures) from all comparisons for all laboratories were sent to T1 and T2. A reconciliation

conference call was held among the taxonomists and the taxonomic QC coordinator to review all disagreements. Greater attention was given to those taxa that were most abundant in individual samples or that were more common across the samples being compared than to uncommon or infrequent taxa. A primary goal of the reconciliation conference call was to determine possible cause(s) of the disagreements. Some of the disagreements were rectified or eliminated when a taxonomist explained her/his rationale for a name to the satisfaction of the others. Unresolved disagreements were maintained on the spreadsheets. They contributed directly to recognized and reported uncertainty in the content of that sample, were carried over to calculation of performance statistics, and described the error rate associated with the data set.

Corrective actions.—The QC coordinator reviewed all notes and results from the call and developed a list of activities for T1 to do on the entire sample lot. The QC coordinator forwarded the notes and list to EPA for review, and EPA sent written corrective actions and instructions to the taxonomists. Example corrective actions included: 1) slide mount and reidentify all Chironomidae and Oligochaeta; 2) reexamine Baetidae and Acari; 3) reexamine mollusk shells, identify and count only when soft tissue is present; 4) ensure samples sent to QC laboratory are complete, specifically include all slide-mounted material; and 5) proofread all data entries carefully.

Round 1 performance measures were used to isolate those individual laboratories and taxonomists for which additional effort was necessary. Furthermore, individual taxa exhibiting the greatest variability were parsed from the remainder of the data set for determination of whether and how improvement of identifications could be made. For some taxa, uncontrolled variability was handled by collapsing names to higher groupings, i.e., from genus to genus-group, subfamily, or family level. The amount of change in taxonomic precision and completeness between round 1 (precorrective actions) and round 2 (postcorrective actions) is characteristic of, and in part interpreted as, the effectiveness of corrective actions in improving taxonomic consistency.

Round 2.—Samples from round 1 were returned to the primary laboratories. All laboratories were given 6 to 8 wk to respond to round 1 corrective actions, after which another 10% of the total sample lot/laboratory was randomly selected (Table 1) for a 2nd round of QC identifications by T2. Samples from round 1 could potentially have been selected for round 2. The purposes of round 2 were 2-fold: 1) to document the effects of corrective actions and 2) to document the performance measures (specifically, taxonomic precision) associated with the final data set. Note that the final data set is that which existed following implementation of round 1 corrections and was used for the WSA data analysis and assessment. Round 2 was not required for Laboratory J because their round 1 error rates were low and corrective actions unnecessary. Comparisons of results obtained by T1 and T2 were used to calculate final performance measures that represented taxonomic data quality for the entire data set.

Performance measures and measurement quality objectives (MQO).—We determined the number of agreements/matches for all taxa identified by T1 and T2. We assigned errors to 3 types: 1) straight disagreements, 2) hierarchical differences, and 3) missing specimens. *Straight disagreements* occurred when it was obvious that the 2 taxonomists examined the same specimen(s) and assigned them different names. *Hierarchical differences* occurred when either T1 or T2 could not, with confidence, assign the target hierarchical-level name to the specimen(s). If both T1 and T2 assigned a nontarget hierarchical-level name to $\geq$1 specimens, it was counted as an agreement if the hierarchical levels were the same. For example, if genus level was the target for black flies, and each taxonomist identified 14 specimens as Simuliidae, that identification was scored as 14 agreements. An exception was that genus $\times$ species comparisons were called "in agreement" if the genus-level target was met. *Missing specimens* resulted from differences in actual counts that could not be attributed to differences in identifications of $\geq$1 taxa by 1 of the taxonomists.

MQOs are control points above (or below) which most observed values fall (Diamond et al. 1996, Stribling et al. 2003, 2008, Herbst and Silldorf 2006). Specific values are selected based on the distribution of values attained, particularly the minima and maxima, and should reflect performance expectations when routine techniques and personnel are used. Values that are >MQO are not automatically taken to be unacceptable data points; rather, such values are targeted for closer scrutiny to determine possible reasons for the exceedance and might indicate a need for corrective actions (Stribling et al. 2003, MDEQ 2006).

We calculated a series of performance measures for the overall WSA data set using pooled QC samples from all laboratories, and, where appropriate and necessary, we partitioned these measures by laboratory. Percentage taxonomic disagreement (PTD) for a sample (Stribling et al. 2003) is given by

$$\mathrm{PTD} = \left(1 - \left[\frac{a}{N}\right]\right) \times 100,$$

where a is the total number of agreements (matches

between T1 and T2) summed across all individuals and taxa and N is the total number of individuals identified in the larger of the 2 counts for a sample. The MQO for PTD was 15%, i.e., a sample with PTD $\geq$ 15% would be examined in more detail for the causes of disagreements.

The relative difference between the total counts from 2 taxonomists for a sample (% difference in enumeration [PDE]) is calculated as

$$\text{PDE} = \frac{|n_1 - n_2|}{n_1 + n_2} \times 100,$$

where n_1 is the number of individuals counted by T1 in the sample and n_2 is the number of individuals counted by T2. The MQO for PDE was 5%, i.e., enumeration comparisons $\geq$5% would be more closely scrutinized. Some users of this performance characteristic (MDEQ 2006) prefer that the formula reflect a relative proportional difference and, thus, divide the denominator by 2 and use MQO = 10%.

A complete identification occurred when the name placed on an individual matched the target hierarchical level (Table 2). Percentage taxonomic completeness (PTC) was calculated as

$$\text{PTC} = \frac{x}{N} \times 100,$$

where x is the number of individuals in a sample for which the identification meets the target hierarchical level, and N is the total number of individuals in the sample. No MQO was specified for PTC, but general expectations were that values would be ~95%. The absolute value of the difference between these numbers for T1 and T2 was used as indication of consistency of effort. Expectations were that the absolute difference would be <10 percentage points. Samples with absolute PTC differences >10 percentage points were examined to determine the taxa responsible for the differences. In our paper, mean PTC could be calculated for samples completed by an individual taxonomist, for a laboratory with $\geq$1 primary taxonomists (T1), or by all T1 for the project. PTC should not be confused with the observed/expected index, which Hawkins (2006) defines as taxonomic completeness.

Relative % difference (RPD; Keith 1991, Berger et al. 1996, APHA 2005, Stribling et al. 2008) is the proportional difference between 2 measures. RPD is calculated for a single taxon across all samples (n = 72) as

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

where A is the number of individuals of a taxon counted by pooled T1 and B is the number of individuals counted by T2. Low RPD values indicate better consistency than do high values. However, when evaluating RPD values, 2 cautions should be borne in mind: 1) results can be misleading when numbers are very low or 0, and 2) results should be evaluated in the context of the number of samples in which individuals of a taxon were found. For example, if T1 identified 2 individuals of taxon A across all samples, and T2 identified only 1, RPD would be 67% for that taxon. One or 2 specimens of a taxon in 1 of 72 samples does not provide sufficient information to judge taxonomic consistency for that taxon. However, if T1 identified 200 individuals in 20 samples and T2 identified 100 individuals in 10 samples, then an RPD of 67% would be reason to question data for that taxon. Other than these cautions, low values of RPD indicate similarity in counts.

Statistical differences in performance measures.—We used 2-sample t-tests assuming equal variances to determine whether PTD, PDE, and PTC differed between rounds 1 and 2 across the entire data set (round 1, n = 74; round 2, n = 75) and for individual laboratories (Table 1).

Potential effects of error at different scales

Identification data.—We compiled counts by T1 and T2 of individuals for each taxon across the set of QC samples (n = 72) and calculated RPD for each taxon. We pooled T1 counts across all taxonomists, and T2 counts were from the single QC taxonomist.

Macroinvertebrate assemblage metrics and multimetric index.—The multimetric indexes developed for the 9 WSA assessment regions were based on 19 metrics (Stoddard et al. 2008). We assessed variation among replicate samples to determine the effect of taxonomic variability on metrics. We treated the value of a metric calculated from T2 data for a sample as a *taxonomy replicate* of the value of the metric calculated from T1 data for the same sample. We calculated mean absolute values (*meanABS*) of the differences between T1 and T2 values (round 1 precorrective action) for each metric across all samples for which $\geq$300 organisms were attained (n = 72) as

$$meanABS = \frac{\sum |x - y|}{n},$$

where x = value of the metric calculated with taxonomic results from T1 and y = value of the metric calculated with taxonomic results from T2.

During the WSA, *field replicates* were collected at sites (n = 60) that were randomly selected from the sample frame and were sampled <2 wk after collection of the primary samples. We evaluated the

TABLE 3. Mean (SD) taxonomic error rates (% taxonomic disagreement [PTD] and % difference in enumeration [PDE]), and % taxonomic completeness (PTC) across all samples. Identifications were done by 25 primary taxonomists (T1) and 1 quality control (QC) taxonomist (T2). The QC taxonomist was the same individual for all samples in both rounds of identifications. The differences between means of PTD, PDE, and PTC in rounds 1 and 2 had p values of 0.001, >0.05, and >0.05, respectively (2-sample t-tests).

Taxonomic QC	n	PTD	PDE	PTC
Round 1	74	21.0 (13.4)	2.6 (11.6)	89.9 (10.7)
Round 2	75	14.0 (10.3)	1.1 (1.3)	92.5 (9.1)

magnitude of the effect of taxonomic differences on resulting metric values by comparing *meanABS*s from taxonomy replicates to *meanABS*s from field replicates (i.e., where x = value of the metric calculated with taxonomic results from the primary sample and y = value of the metric calculated with taxonomic results from a field replicate). These comparisons were intended to determine whether the effects of taxonomic error on metrics could exceed effects introduced by field variability. We calculated Pearson correlation coefficients between x and y values for each metric for field and taxonomic replicates.

Condition assessment narratives.—In the WSA, multimetric index scores were converted to narrative condition classes (good, fair, and poor) based on numeric thresholds (Van Sickle and Paulsen 2008). We used the overall index scores for the 72 samples (T1 and T2, round 1, precorrective action) to quantify the number and proportion of instances where narrative assessments based on taxonomic data from T1 agreed with narrative assessments based on taxonomic data from T2. When narrative assessments disagreed, we totaled the number of instances in which the magnitude of difference was 1 or 2 assessment categories. We calculated Cohen's κ, with adjustment above what would be expected by chance (Fleiss 1981), to quantify the narrative agreement rate.

Results

Documentation of data quality and the effectiveness of corrective actions

PTD across all QC samples was 21.0% for round 1 and 14.0% for round 2 (Table 3). The proportion of samples meeting the 15% MQO for PTD increased from 27 to 71% after corrective actions were taken (i.e., from round 1 to round 2). PTDs for individual laboratories ranged from 29.7 to 8.1 in round 1 and 19.1 to 8.1 in round 2 (Table 4). PTDs decreased significantly between rounds 1 and 2 for 2 laboratories (A and H), but did not change significantly for the other laboratories (Table 4). PDE was substantial (25.9%) for 1 laboratory in round 1, but decreased to <2% in round 2. PDEs were <2% during both rounds for all other laboratories (Table 4). Mean PTC was 92.5% in round 2 (SD = 9.1, n = 75 samples; Table 3). *meanABS* between taxonomic replicates was ≥10% in only 7 of 75 PTC comparisons (9.3%). Laboratory J was not required to complete round 2 evaluations because its error rates were low (mean PTD = 8.1, mean PDE = 0.6) and its PTC was high (mean PTC = 98.4) in round 1. Round 1 results for laboratory J were simply carried over to round 2 summaries.

Samples were numerically dominated by Chironomidae, Ephemeroptera, Trichoptera, Coleoptera, and Oligochaeta, which had final (round 2) rates of errors of 11.3, 16.5, 12.8, 6.9, and 22.4%, respectively (Table 5). PTC for major taxonomic groups ranged from 84 to 99% (not shown). The total number of errors for the overall data set (straight, hierarchical, and missing) dropped by ~38 percentage points from round 1 (6856) to round 2 (4233) (Table 5). The highest

TABLE 4. Mean taxonomic error rates (% taxonomic disagreement [PTD] and % difference in enumeration [PDE]), % taxonomic completeness (PTC), and number of samples for rounds 1 and 2 by laboratory. PTC is calculated for all taxonomists (T1) associated with individual laboratories. * indicates round 2 values are significantly different ($p < 0.05$) from round 1 values (laboratory A PTD: $p = 0.007$, laboratory H PTD and PTC: $p = 0.016$).

	Round 1				Round 2			
Laboratory	n	PTD	PDE	PTC	n	PTD	PDE	PTC
A	18	29.7	1.7	93.6	14	14.4*	1.4	90.8
B	20	16.9	1.1	83.8	25	13.6	0.9	89.7
D	12	16.7	1.9	95.4	12	19.1	1.6	95.3
E	3	11.8	0.5	97.9	3	9.6	0.8	99.6
F	4	16.2	25.9	96.6	3	15.9	1.8	98.1
H	10	29.6	1.0	80.8	12	12.7*	0.8	93.7*
I	3	17.6	0.2	88.6	2	9.5	0.1	84.6
J	4	8.1	0.6	98.4	4	8.1	0.6	98.4

TABLE 5. Identification errors and types of error for major taxa in rounds 1 (R1) and 2 (R2). Straight errors occurred when 2 taxonomists examined the same specimen(s) and assigned them different names. Hierarchical errors occurred when one or the other taxonomist could not assign the target-level name to the specimen(s). Missing errors occurred when differences in counts could not be attributed to ≥1 taxa identified by one of the taxonomists. Total number identified is the number of specimens identified for each taxon across all samples. Other taxa includes Bivalvia, Crustacea, Enopla, Hydrozoa, and Turbellaria.

							Error-type distribution					
	Total number identified		Number of errors		% errors		Straight		Hierarchical		Missing	
Taxon	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
All taxa	31,640	31,460	6856	4223	21.7	13.5	2965	1460	2427	1406	1464	1367
Arachnida	229	258	74	86	32.3	33.3	38	47	15	9	21	30
Chironomidae	10,579	10,832	2632	1225	24.9	11.3	1422	563	927	296	283	366
Coleoptera	2463	2829	267	195	10.8	6.9	83	60	64	49	120	86
Ephemeroptera	4837	6238	937	1029	19.4	16.5	247	190	509	632	181	207
Gastropoda	2060	1560	251	250	12.2	16.0	100	58	99	86	52	106
Hirudinea	32	21	11	4	34.4	19.0	6	0	4	2	1	2
Oligochaeta	2190	1946	494	435	22.6	22.4	232	220	75	6	187	209
Other taxa	5077	4097	1174	488	23.1	11.9	308	148	454	31	412	209
Plecoptera	796	742	144	151	18.1	20.4	21	72	77	48	46	32
Trichoptera	3377	2883	872	370	25.8	12.8	508	103	203	147	161	120

percentages of taxon-specific errors were for Chironomidae, Trichoptera, and other taxa in round 1, and these groups had the greatest reduction in errors after corrective action was taken (13.6, 13.0, and 11.2 percentage points, respectively; Table 5). Overall rates of errors associated with Chironomidae fell from 24.9 (round 1) to 11.3 (round 2) after all laboratories began mounting all chironomids on slides as a corrective action. Laboratory-specific error rates for Chironomidae ranged from 8 to 33% (mean = 20.6%, SD = 8.8) in round 1, and these rates decreased to 8 to 18% (mean = 12.3%, SD = 3.9) in round 2. A 6% increase in the percentage of errors that occurred in 1 laboratory (E) was caused by a single sample that skewed the mean proportion of errors. Laboratory-specific decreases in errors associated with chironomids ranged from ~2 to 17 percentage points. The overall change for Hirudinea was 15 percentage points, but the error rate for this group was elevated because few leeches were found (32 and 21 individuals in rounds 1 and 2, respectively).

Potential effects of error at different scales

Differences in detection of taxa in samples.—Comparison of T1 and T2 data in round 1 (precorrective action) showed large differences in nomenclature and counts for some taxa (*Thienemannimyia* genus group [Chironomidae], Hydropsychidae [Trichoptera], and Baetidae [Ephemeroptera]; Table 6). Substantial differences were particularly evident with *Meropelopia* (Chironomidae) and several hydropsychid genera (*Ceratopsyche*, *Cheumatopsyche*, and *Diplectrona*). Among Baetidae, inconsistencies in identifications for *Acerpenna*, *Baetis*, *Fallceon*, and *Plauditus* were substantial, and confusion over how to separate *Procloeon* and *Centroptilum* (a common problem) was evident in the data.

Effects of error on metric values and index.—*meanABS* of 8 of 19 metrics was greater for taxonomy than for field replicates (Table 7). *r* values for 6 different metrics and 1 of the previous 7 were lower for taxonomy than for field replicates. For field replicates, *r* values for individual metrics ranged from 0.39 (shredder richness) to 1.0 (Ephemeroptera richness); the *r* value for the multimetric index was 0.72. For taxonomy replicates, *r* values for individual metrics ranged from 0.48 (% individuals in top 5 taxa) to 0.98 (% noninsect individuals); the *r* value for the multimetric index was 0.89.

Error effects on narrative condition assessments.—Narrative condition assessments based on data from T1 and T2 were in agreement for 77.8% of the round 1 samples (Table 8). Narrative condition assessments differed by 1 category (i.e., good–fair, fair–poor) for 15 samples (20.8%) and by 2 categories (good–poor) for 1 sample (1.4%). Cohen's κ statistic adjusted the 77.8% agreement rate to account for chance agreement. For the data in Table 8, $\kappa = 0.66$ (95% confidence interval = 0.52–0.81), where κ can range from 0 to 1. This value of κ indicates fair to good agreement above what would be expected by chance (Fleiss 1981).

Discussion

Value of QC for taxonomic data

The purpose of QC is to reduce the error rate in an existing data set. In addition, QC: 1) documents data

TABLE 6. Counts of individuals by 25 primary taxonomists (T1) and 1 quality control (QC) taxonomist (T2) for selected taxa across all QC samples (round 1). The QC taxonomist was the same individual for all samples. Number of samples is the number in which either T1, T2, or both identified the taxon. RPD = relative % difference.

Taxon	Number of samples	Number of individuals counted		RPD
		T1	T2	
Chironomidae:Tanypodinae: *Thienemannimyia* genus group				
Conchapelopia	23	119	25	130.6
Conchapelopia genus group	20	2	95	191.8
Hayesomyia	4	14	4	111.1
Helopelopia	8	15	25	50.0
Meropelopia	16	5	72	174.0
Meropelopia genus group	2	0	6	200.0
Rheopelopia	3	2	5	85.7
Telopelopia	2	4	0	200.0
Thienemannimyia	8	60	26	79.1
Thienemannimyia genus group	39	122	88	32.4
Hydropsychidae				
Ceratopsyche	26	569	697	20.2
Ceratopsyche/Hydropsyche	1	0	1	200.0
Cheumatopsyche	42	681	541	22.9
Diplectrona	8	27	60	75.9
Hydropsyche	29	440	415	5.8
Hydropsychidae	29	193	220	13.1
Macrostemum	1	1	1	0.0
Potamyia	3	14	8	54.5
Baetidae				
Acentrella	14	153	162	5.7
Acerpenna	16	28	51	58.2
Baetidae	46	317	307	3.2
Baetis	41	762	575	28.0
Callibaetis	5	2	8	120.0
Camelobaetidius	3	0	5	200.0
Cloeon	1	1	0	200.0
Diphetor hageni	2	0	29	200.0
Fallceon	7	16	38	81.5
Paracloeodes	6	5	28	139.4
Plauditus	12	64	122	62.4
Centroptilum	9	34	0	200.0
Procloeon	6	7	7	0.0
Procloeon/Centroptilum	22	0	78	200.0
Pseudocloeon	5	2	5	85.7

quality, 2) provides direction for improvement through implementation of corrective actions, and 3) improves quality of data sets over time by monitoring performance standards. Had QC not been implemented for the WSA, no objective guidelines would have been available for determining whether data quality was acceptable and whether corrective actions were warranted. In short, the data would have been of unknown quality.

Taxonomic data have multiple uses, of which 1 is to conduct site assessments that can be aggregated for condition assessments at broader spatial scales, such as that of the WSA. Potential secondary uses of the WSA data set include identification and evaluation of stressors, determination of relationships between taxa and stressor gradients, biogeographical uses (documentation of spatial ranges and patterns of emergence), development of biological criteria and standards, conservation planning, and contributions to phylogenetic studies. The data set is more likely to be assessed for these and other potential applications when the amount of information about the quality of the data is high. Another critical reason to document data quality is that knowledge of data quality enhances our ability to defend data sets against potential misuse. Our evaluation process allowed specification of individual operators (laboratories, taxonomists) and taxa for direct corrective actions and provided a valid and straightforward statement of taxonomic data quality for the WSA data set. The

TABLE 7. Absolute differences in metrics used to create multimetric indexes for the Wadeable Stream Assessment (WSA) (Stoddard et al. 2008) when metrics were calculated from taxonomic identifications based on replicate samples. Taxonomy replicates consisted of identifications by 25 primary taxonomists (T1) and by 1 quality control taxonomist (T2) ($n = 72$, round 1, precorrective action). Field replicates consisted of identifications by T2 from the set of primary samples collected during the WSA and a set of samples collected from WSA sampling sites ~2 wk after primary sampling ($n = 60$, round 1, postcorrective action). Correlation coefficients (r) were calculated for paired (T2 primary vs T2 field replicate, T1 vs T2) values of indexes and metrics. Numbers in bold are the larger of the 2 absolute differences or the smaller of the 2 correlation coefficients. Min = minimum, max = maximum, EPT = Ephemeroptera, Plecoptera, Trichoptera, PTV = pollution tolerance value.

	Field					Taxonomy				
	Absolute difference					Absolute difference				
Index and metrics	Mean	SD	Min	Max	r	Mean	SD	Min	Max	r
Multimetric index	**9.4**	9.8	0	48.1	**0.72**	6.2	5.4	0.01	30.6	0.89
% EPT taxa	**5.8**	5.4	0.3	30.8	**0.74**	4.1	4.8	0	34.3	0.82
% EPT individuals	**10.9**	10.6	0.3	50.9	**0.74**	3.7	10.8	0	82.6	0.75
% noninsect individuals	**10.3**	9.3	0.1	38.2	**0.68**	2.6	3.3	0	19.0	0.98
% Ephemeroptera taxa	**3.7**	3.1	0.1	19.2	0.67	3.3	4.1	0	25.7	**0.62**
% Chironomidae taxa	**8.6**	7.4	0.2	44.6	**0.48**	5.4	7.2	0	48.8	0.68
Shannon diversity	**0.3**	0.3	0.0	1.2	0.63	0.2	0.4	0	2.8	**0.58**
% individuals in top 5 taxa	**9.3**	8.6	0.5	41.0	0.55	6.3	11.4	0	75.4	**0.48**
Scraper richness	**1.6**	1.4	0	7	0.71	1.4	1.3	0	6.0	**0.65**
Shredder richness	1.9	1.6	0	6	0.39	**18.0**	30.7	0	5.0	0.92
% burrower taxa	**6.0**	5.1	0	24.7	0.49	4.9	4.6	0.2	19.5	0.54
% clinger taxa	**6.2**	5.6	0	23.6	0.77	5.1	7.6	0	51.4	**0.64**
Clinger richness	**3.2**	2.9	0	12	0.84	2.2	8.5	0	18.0	**0.83**
Ephemeroptera richness	1.5	1.0	0	5	1.00	**3.8**	3.2	0	13.0	**0.69**
EPT richness	2.7	2.4	0	11	**0.82**	**7.8**	6.7	0	33.0	0.87
Intolerant richness	2.3	2.3	0	10	**0.83**	**6.3**	5.2	0	18.0	0.85
% tolerant taxa	5.6	5.3	0.1	29.2	**0.69**	**13.1**	11.6	0	68.8	0.83
PTV 0–5.9 richness	4.6	4.1	0	18.0	**0.81**	**22.1**	32.7	0	24.0	0.94
% PTV 0–5.9 taxa	7.7	5.8	0	22.6	**0.70**	**26.6**	33.9	0	68.6	0.99
% PTV 8–10 taxa	4.0	2.9	0	11.7	**0.75**	**20.9**	34.9	0	18.4	0.79

known mean rate of error associated with benthic macroinvertebrate taxonomy in the data set is 14% (SD = 10.3; Table 3).

The QC process provided taxon-specific data on the consistency of identifications and allowed communication of problems with data quality to individual laboratories and taxonomists. The results are of value to taxonomists (as data producers) because they provide knowledge of the taxa in need of increased scrutiny and of the relative capacities of different taxonomists to provide consistent identifications. The results are of value to data analysts and decision makers (as data users) because they document and communicate uncertainty associated with identifications of individual taxa, laboratories/taxonomists, and the data set.

TABLE 8. Narrative assessments based on multimetric index scores calculated from identifications by 25 primary taxonomists (T1) and 1 quality control (QC) taxonomist (T2) (round 1, $n = 72$ samples). Numbers in bold indicate samples for which assessments between T1 and T2 were in agreement.

		Primary (T1)		
Assessment narratives		Good	Fair	Poor
QC (T2)	Good	**17**	7	0
	Fair	5	**12**	2
	Poor	1	1	**27**

The QC process included implementation of corrective actions in a manner that enabled us to relate subsequent results to the MQO (PTD = 15%). Corrective actions improved the consistency of taxonomic data; the percentage of samples that met the MQO increased from 27 to 71% from round 1 to round 2. Our experience has been that PTD = 10% typically is difficult to reach, and PTD = 20% usually is attained with minimal difficulty. The MQO used in this project simply splits that difference. Chessman et al. (2007) reported very low error rates (mean PTD = 4.2, mean PDE = 0.05) but did not specify an MQO or other acceptability criteria.

Problematic taxa and corrective actions

Problems.—The taxa with the highest error rates, Chironomidae, Ephemeroptera, and Trichoptera (Table 5), are often cited by production taxonomists as the groups for which consistency is difficult. More detailed examination of RPD showed that the consistency of identifications for individual genera within these difficult groups is not uniform (Table 6). Thus, different decisions can be made to address the errors. For example, some laboratories used morphotyping for chironomid identifications, whereas others did not. One corrective action was to ask those laboratories that used morphotyping to slide mount all chironomids in the sample lot (not just those in QC samples) and reidentify the chironomids from the slide mounts. The error rate for the group fell ~14 percentage points from 24.9 to 11.3% from round 1 to round 2 (Table 5). This improvement could be attributed to 2 factors, of which 1 might be slide mounting itself. The other is that complete slide mounting explicitly forces the taxonomist to look at every specimen and to forgo subsampling, which is a component of morphotyping. The issue with consistency in chironomid identifications is not specifically whether morphotyping was used. Rather, it is the training and experience of the person doing the work. The more experience a taxonomist has with chironomids, the less important slide mounting is for good identifications. We observed this pattern with several taxonomists and in different laboratories.

A combination of factors can lead to increased variability in taxonomic data. One of these factors is poor sample condition (New 1996, Stribling et al. 2003, Ball et al. 2005, Schander and Willassen 2005, Cuffney et al. 2007). Changes in sample preservation, shipping, and handling are potential approaches to controlling specimen damage as a cause of error. However, more experienced taxonomists will have an easier time attaining positive identifications with poor or damaged specimens than will inexperienced taxonomists, a situation conceptually similar to the slide-mounting issue described above.

Operational taxonomic units (OTUs).—Based on consistency of identifications in this data set, recommendations were provided to WSA data analysts for OTUs that should be used for analysis (metric and index calculation). Most chironomids were identified consistently enough that analysts could be reasonably confident of data labeled as certain taxa. Several groups were problematic. These groups included genera often thought of as making up the *Thienemannimyia* genus group (*Conchapelopia, Hayesomyia, Helopelopia, Meropelopia, Rheopelopia, Telopelopia,* and *Thienemannimyia,* and, occasionally, *Conchapelopia* genus group and *Meropelopia* genus group). Our recommendation was to collapse these genera and genus groups to *Thienemannimyia* genus group for analysis. Other complexes of chironomid genera that were reported inconsistently were *Cricotopus/Orthocladius* and *Eukiefferiella/Tvetenia* (Table 6). Our recommendation was to use *Cricotopus, Orthocladius, Cricotopus/Orthocladius,* and *Orthocladius/Cricotopus* as *Cricotopus/Orthocladius* for analysis; the same recommendations were given for *Eukiefferiella/Tvetenia*.

Fourteen genera of Ceratopogonidae were reported in the QC data set. Only 2 of these genera (*Culicoides* [14 samples, RPD = 16.7%] and *Probezzia* [20 samples, RPD = 10.8%]) were identified consistently in enough samples to maintain them for analysis. Our recommendation was to collapse all other genera to family level. Extremely inconsistent results for 13 reported genera of Baetidae (Ephemeroptera) (Table 6) suggested that they would be only minimally useful for analysis *at that hierarchical level*, and our recommendation was that they be collapsed to family level. Six genera of Hydropsychidae (Trichoptera), including *Ceratopsyche, Cheumatopsyche, Diplectrona, Hydropsyche, Macrostemum,* and *Potamyia* were reported. Substantial problems related to consistency in recognition of *Diplectrona* and segregation of specimens of *Hydropsyche, Ceratopsyche,* and *Cheumatopsyche* were observed; however, our recommendation was to maintain hydropsychid caddisfly data at genus level. Extreme inconsistency for genera in Gomphidae and Calopterygidae (Odonata) led us to recommend that all data be collapsed to family level for analysis.

Applications

Cao et al. (2003) and Yuan (2007) correctly point out that some types and magnitudes of error have little effect on ultimate biological assessments and, furthermore, that attention to issues of data quality should be based on the ultimate uses. Our analyses show that the magnitude of sample-based taxonomic error varies among taxa, laboratories, and taxonomists and that the variability can affect interpretations of taxonomic diversity and can cause differences in metrics and indexes. The intent of the entire QC exercise was *not* to show that the WSA data set is of poor quality. On the contrary, the intent was to show the data set is of *known* quality and that different levels of confidence are warranted depending on how the data are to be used. Haase et al. (2006) evaluated the occurrence of taxonomic identification error associated with European biological monitoring programs as part of the European Union Water Directive Framework. Their initial evaluations showed that the occurrence of errors

was considerable but substantial reductions in error rates were associated with subsequent QC testing (Haase et al. 2006). This result is predicted by basic QC theory—continuous improvement results from oversight of any process (a concept originating with Walter A. Shewhart and W. Edwards Deming), particularly through use of statistical QC (Shewhart 1986).

The quality of the data set did not change as a result of the 2nd round of QC reidentifications (i.e., no further corrective actions were taken). Leaders of the WSA could have chosen to move into the data-analysis phase of the assessment (metric and index calibration) with data that had been corrected but without round 2 QC comparisons. The quality of the final data set would have been communicated as having an *assumed* error rate less than that of round 1 (i.e., improved); WSA leaders did not regard resting on this assumption as a viable option. Round 2 of the QC process enabled us to accomplish our primary goals; the quality of the final data set was characterized and changes after corrective actions were documented.

Our approach to taxonomic QC is applicable for smaller scale biological monitoring programs at state, regional, county, and project-specific scales. Several state programs have made a commitment to use this form of QC with sample sizes ranging from 55 (10% of a 500+ sample lot) to 3 (100% of a 3-sample lot). The process can be used as a routine process for documenting performance of individual laboratories or taxonomists by randomly testing 5 to 10% of archived/identified samples every 3 to 12 mo. For purposes of many programs, a 2nd round of reidentification might not be critical, and the results of 1 round of external QC could be used to identify errors and corrective actions, make corrections in the primary data, and, perhaps, refine hierarchical target levels. This procedure is acceptable if the ultimate users of the data agree. In addition, the process can help identify the need for additional staff training or technical literature for certain taxonomic groups.

So, the question could be asked, "Would the WSA results be a better assessment if the quality of taxonomic data were unknown?" Most would agree that it is better to base assessments on data of known quality. Error is not a bad thing. It is *not* knowing about its existence or extent that impedes efforts at error management and control, and risks loss of credibility, defensibility, and utility of biological assessments and the data sets on which they are based.

Acknowledgements

This work was completed under US EPA contract numbers EP-C-06-033 and 68-C-04-006 to the Great Lakes Environmental Center and Tetra Tech, Inc. We thank all taxonomists and their laboratories for efforts in sample processing, identification, and cooperation in this project. We are grateful to colleagues and coworkers, including Chuck Hawkins, John Van Sickle, Trefor Reynoldson, Evan Hornig, Jim Haney, Phil Larsen, Mike Barbour, Vince Resh, Jerry Diamond, Joe Flotemersch, Esther Peters, John O'Donnell, and Ellen Tarquinio, for reviewing various versions of our manuscript and for their ensuing comments and discussions. Ben Jessup assisted with some of the statistical treatments, and Ellen Tarquinio and Otto Gutenson (USEPA) participated in the reconciliation conference calls and development of corrective actions. Pamela Silver and John Van Sickle and 2 anonymous referees provided valuable 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.

BALL, S. L., P. D. N. HEBERT, S. K. BURIAN, AND J. M. WEBB. 2005. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. Journal of the North American Benthological Society 24:508–524.

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 Publishing Corp., 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.

CHAPMAN, A. D. 2005. Principles of data quality. Version 1.0. Report for the Global Biodiversity Information Facility (GBIF). Global Biodiversity Information Facility, Copenhagen, Denmark. (Available from: http://www2.gbif.org/DataQuality.pdf)

CHESSMAN, B., S. WILLIAMS, AND C. BESLEY. 2007. Bioassessment of streams with macroinvertebrates: effect of sampled habitat and taxonomic resolution. Journal of the North American Benthological Society 26:546–565.

CLARKE, R. T. 2000. Uncertainty in estimates of biological quality based on RIVPACS. Pages 39–54 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

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., 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.

CUFFNEY, T. F., M. D. BILGER, AND A. M. HAIGLER. 2007. Ambiguous taxa: effects on the characterization and interpretation of invertebrate assemblages. Journal of the North American Benthological Society 26:286–307.

DALCIN, E. C. 2004. Data quality concepts and techniques applied to taxonomic databases. PhD Thesis, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton, Southampton, UK. (Available from: http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf)

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.

DINES, R. A., AND J. A. D. MURRAY-BLIGH. 2000. Quality assurance and RIVPACS. Pages 71–78 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

FLEISS, J. 1981. Statistical methods for rates and proportions. 2nd edition. John Wiley and Sons, New York.

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

HAASE, P., J. MURRAY-BLIGH, S. LOHSE, S. PAULS, A. SUNDERMANN, R. GUNN, AND R. CLARKE. 2006. Assessing the impact of errors in sorting and identifying macroinvertebrate samples. Hydrobiologia 566:505–521.

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.

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.

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. (EDITOR). 1988. Principles of environmental sampling. ACS Professional Reference Book. American Chemical Society, Columbus, Ohio.

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

KLEIN, B. D. 2001. Detecting errors in data: clarification of the impact of base rate expectations and incentives. Omega 29:391–404.

KLEMM, D. J., J. M. LAZORCHAK, AND P. A. LEWIS. 1998. Benthic macroinvertebrates. Pages 147–182 *in* J. M. Lazorchak, D. J. Klemm, and D. V. Peck (editors). Environmental monitoring and assessment program—surface waters: field operations and methods for measuring the ecological condition of wadeable streams. EPA/620/R-94/004F. US Environmental Protection Agency, Washington, DC.

MDEQ (MONTANA DEPARTMENT OF ENVIRONMENTAL QUALITY). 2006. 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)

MOULTON, S. R., J. L. CARTER, S. A. GROTHEER, T. F. CUFFNEY, AND T. M. SHORT. 2000. Methods of analysis by the U.S. Geological Survey National Water Quality Laboratory—processing, taxonomy, and quality control of benthic macroinvertebrate samples. U.S. Geological Survey Open-File Report 00–212. National Water Quality Laboratory, US Geological Survey, Denver, Colorado.

NEW, T. R. 1996. Taxonomic focus and quality control in insect surveys for biodiversity conservation. Australian Journal of Entomology 35:97–106.

PAULSEN, S. G., A. MAYIO, D. V. PECK, J. L. STODDARD, E. TARQUINO, S. M. HOLDSWORTH, J. VAN SICKLE, L. L. YUAN, C. P. HAWKINS, A. T. HERLIHY, P. R. KAUFMANN, M. T. BARBOUR, D. P. LARSEN, AND A. R. OLSEN. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. Journal of the North American Benthological Society 27:812–821.

PETERS, J. A. 1988. Quality control infusion into stationary source sampling. Pages 317–333 *in* L. H. Keith (editor). Principles of environmental sampling. ACS Professional Reference Book. American Chemical Society, Columbus, Ohio.

SCHANDER, C., AND E. WILLASSEN. 2005. What can biological

barcoding do for marine biology? Marine Biology Research 1:79–83.

SHAPIRO, M. H., S. M. HOLDSWORTH, AND S. G. PAULSEN. 2008. The need to assess the condition of aquatic resources in the US. Journal of the North American Benthological Society 27:808–811.

SHEWHART, W. A. 1986. Statistical method from the viewpoint of quality control. Dover Publications, New York.

SMITH, M. A., B. L. FISHER, AND P. D. N. HEBERT. 2005. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 360:1825–1834.

STODDARD, J. L., A. T. HERLIHY, D. V. PECK, R. M. HUGHES, T. R. WHITTIER, AND E. TARQUINIO. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27:878–891.

STODDARD, J. L., D. V. PECK, S. G. PAULSEN, J. VAN SICKLE, C. P. HAWKINS, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. P. LARSEN, G. LOMNICKY, A. R. OLSEN, S. A. PETERSON, P. L. RINGOLD, AND T. R. WHITTIER. 2005. An ecological assessment of western streams and rivers. EPA 620/R-05/005. US Environmental Protection Agency, Washington, DC.

STRIBLING, J. B., B. K. JESSUP, AND D. L. FELDMAN. 2008. Precision of benthic macroinvertebrate indicators of stream condition in Montana. Journal of the North American Benthological Society 27:58–67.

STRIBLING, J. B., S. R. MOULTON, AND G. L. LESTER. 2003. Determining the quality of taxonomic data. Journal of the North American Benthological Society 22:621–631.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2004a. Wadeable Stream Assessment: benthic laboratory methods. EPA 841-B-04007. Office of Water and Office of Research and Development, US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2004b. Wadeable Stream Assessment: field operations manual. EPA 841-B-04-004. Office of Water and Office of Research and Development, US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2006. Wadeable Streams Assessment. A collaborative survey of the nation's streams. EPA 841-B-06-002. Office of Research and Development and Office of Water, US Environmental Protection Agency Washington, DC.

VAN SICKLE, J., AND S. G. PAULSEN. 2008. Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. Journal of the North American Benthological Society 27:920–931.

YUAN, L. L. 2007. Effects of measurement error on inferences of environmental conditions. Journal of the North American Benthological Society 26:152–163.

Received: 27 November 2007
Accepted: 14 August 2008