J. N. Am. Benthol. Soc., 2011, 30(3):680–701

DOI: 10.1899/10-067.1
Published online: 31 May 2011

The comparability of bioassessments: a review of conceptual and methodological issues[1]

Yong Cao[2] AND Charles P. Hawkins[3]
Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, Utah State University, Logan, Utah 84322-5210 USA

Abstract. Differences in the raw data used in bioassessments and choices regarding how those data are analyzed and summarized can affect inferences regarding the status of ecological resources and, thus, the degree to which we can trust individual ecological assessments, compare assessments across different programs and regions, or share data when developing or refining new endpoint indices. Progress in addressing these issues has been hindered by lack of consensus regarding what a general definition of comparability should be in the context of bioassessments and what measures of comparability are appropriate for ecological data. In this paper, we review the state of knowledge regarding the comparability of assessments as affected by differences in raw data (composition and relative abundance of taxa), derived measures (biotic metrics and endpoint indices), and assessment levels (condition classes). We specifically address the extent to which the comparability of assessments can be compromised by systematic differences in data, discuss the factors known to affect data comparability, and consider the techniques available to evaluate and improve comparability. Rigorous assessment of data comparability should be a standard aspect of quality assurance when developing and applying biological indices.

Key words: bioassessment, biological indices, ecological assessments, data comparability, data quality, calibration, macroinvertebrates, sampling methods.

Local, state, and national environmental agencies conduct ecological assessments to determine the condition of ecological resources in both individual water bodies and aggregates of water bodies within entire regions. These agencies may share the same general goal to assess the biological condition of aquatic life (e.g., US Clean Water Act, USC §§1251–1387; European Commission 2000), but they often differ substantially in how they collect raw data, produce biological indices from those data, and use those indices to assess biological condition. The assessments conducted by different agencies are expected to be comparable, but we actually know little about their comparability and how differences in either raw data or the biological indices derived from the data affect assessments (Buffagni and Furse 2006, Clarke and Hering 2006, Davies and Jackson 2006, Hawkins 2006, Hawkins et al. 2010a).

We need to understand how comparable the data and assessments collected by different monitoring agencies and programs are for several reasons. First, if different data or indices lead to inconsistent or conflicting assessments of individual water bodies, the use of bioassessment data for regulatory purposes could be challenged for failing legal tests of data credibility. Second, inclusion of monitoring data collected by various management and research programs could greatly strengthen local and state programs and reduce duplication of effort if the data were comparable. Third, data comparability is a key requirement for any long-term monitoring program. Fourth, assessment comparability is essential for the valid aggregation of site-specific bioassessments into broader basin-, regional-, national-, or international-scale assessments (e.g., GAO 2002, Heinz Center 2002, USEPA 2003, 2006, Buffagni et al. 2007, Erba et al. 2009). Last, the increased sample sizes that result from compiling data from different sources allow researchers to more rigorously test and refine the

[1] We dedicate this paper to Richard H. Norris, our colleague and friend, whose passionate commitment to our discipline has stimulated a generation of aquatic scientists to improve the rigor on which ecological assessments are based.

[2] Present address: Illinois Natural History Survey, Institute of Natural Resource Sustainability, University of Illinois at Urbana-Champaign, 1816 S Oak Street, Champaign, Illinois 21820 USA. E-mail: yong.cao@illinois.edu

[3] To whom correspondence should be addressed. E-mail: chuck.hawkins@usu.edu

concepts and techniques on which effective bioassessment depends.

The general importance of comparability in bioassessments has been recognized for decades (e.g., Ghetti and Bonazzi 1977, ITFM 1994, 1995). Since these initial reports, this issue has been addressed from a variety of perspectives by many authors in the USA (e.g., Diamond et al. 1996, Cao et al. 2005, Astin 2006, Herbst and Silldorff 2006, Rehn et al. 2007, Blocksom et al. 2008) and elsewhere (e.g., Buffagni and Furse 2006, Clarke and Hering 2006, Haase et al. 2006, Buffagni et al. 2007). Despite these efforts, many questions remain that span a range of issues, including the comparability of overall concepts that define biological condition and the comparability of specific types of data collected with different sampling methods or derived from different analytic approaches. To compare and integrate bioassessments, we need to understand what assessments are meant to measure and what we can confidently infer from different data sets and analyses. We also need to identify critical knowledge gaps, so we can prioritize future research needs.

Comparability is a critical issue of both scientific and regulatory importance in the USA and elsewhere. We were funded by the US Environmental Protection Agency (USEPA) Office of Water to conduct a comprehensive review of the state of our knowledge regarding comparability issues as they apply to ecological assessments. In this paper, we describe the main results of that review. Our review is based on primary literature (i.e., peer-reviewed journal articles), books and book chapters, and various forms of grey literature including government reports, web-based publications, and internal reports. We pay particular attention to benthic invertebrates because this group has been most commonly used in bioassessments (e.g., Barbour et al. 1999, Wright et al. 2000) and the methods used for their sampling and analysis are highly variable (ITFM 1994, Wright et al. 2000, Carter and Resh 2001, Friberg et al. 2006, Hawkins 2006). The bulk of bioassessment research and application has been conducted in wadeable streams, but our conclusions should be applicable to all types of ecosystems.

What Exactly is Bioassessment Comparability?

Evaluating the comparability of bioassessments requires that we agree on what comparability means. Previous efforts have defined comparability in terms of either data quality (Diamond et al. 1996, MDCB 2003) or the similarity between samples in their taxonomic composition and taxa abundances (Cao et al. 2005). For example, Diamond et al. (1996, p. 715) stated "if different methods are similar with respect to the quality of data each produces, then data from those methods may be used interchangeably or together." The authors further defined data quality to include precision, accuracy, bias, method sensitivity, and the range of conditions over which a method yields satisfactory data. In another case, the Methods and Data Comparability Board (MDCB 2003, p. 1) considered that "data comparability exists when data are of known quality and can thus be validly applied by external users, even when the project objectives differ." Last, Cao et al. (2005, p. 1106) defined data comparability as "how comparable different sampling methods are in characterizing the composition and relative abundance of taxa in an assemblage."

The first 2 definitions focus on aspects of data quality, another important element in bioassessment (Diamond et al. 1996, Cao et al. 2003). Data quality generally refers to how precisely and accurately a variable of interest is measured (Sokal and Rohlf 1987, Zar 1999). These definitions are appropriate for estimating individual water-quality variables, e.g., NH_3 concentrations, but we do not think they are useful for evaluating data comparability in the context of bioassessments. Bioassessments typically involve analysis of multitaxon sample data, in which the units of analysis are either the set of taxa in samples (e.g., counts and perhaps relative abundances) or various metrics derived from the taxon counts. The general ideas of precision and bias also apply to the measurement of multivariate data quality, but their meaning and measurement are more complex than implied by definitions 1 and 2 for 2 reasons. First, it is not at all clear how characterizations of precision for multivariate data relate to comparability. For example, consider 2 sites that both contain species A and B. At site 1, the abundance of species A is estimated as 10 ± 1 (mean $\pm$ SD) and that for species B is estimated as 6 ± 5. At site 2, the mean and precision estimates are reversed: abundance of species A is 6 ± 5 and of species B is 10 ± 1. In this example, the average (multivariate) precision of the 2 samples is the same, but we do not think that expression of precision tells us much about multivariate data comparability because our confidence in estimating different components of the assemblage differs substantially among samples. Second, a more problematic issue is that we can seldom estimate the accuracy with which we estimate the response of biological assemblages to stress from field data because we almost never have predisturbance data from which we can confidently quantify how biological conditions have changed at the site. Definition 3 recognizes the need to define the comparability of sample data in terms of how similar samples are in their taxon counts, but the comparability of sample data does not guarantee that assessments will be comparable—raw data can be treated in different ways, which may lead to different assessments.

We argue that a more general definition of comparability is needed for bioassessments in which comparability is evaluated in the context of how inferences regarding biological condition are influenced by a variety of individual and combined decisions regarding data treatment and summary (Fig. 1). The elements of a bioassessment range from the conceptual framing of what is being assessed to the details of raw-data collection and analysis and include: 1) selection of the assessment endpoint (an ecological property) that is relevant to the designated use assigned by society to a water body (e.g., viability of a species, biodiversity, ecological integrity, ecosystem productivity, nutrient retention); 2) selection of an endpoint index, which can be based on raw data or derived measures, to quantify the condition of those properties (e.g., counts of an individual species, species richness, an index of similarity between observed and expected assemblages, an index of biological integrity, a measurement of primary production, an uptake coefficient describing the rate of nutrient uptake); 3) choice of the analytical procedures used to develop and apply the indices; 4) decisions regarding how and when to collect the raw sample data used to estimate the endpoint indices; and 5) decisions regarding how sample data are treated after collection (e.g., taxonomic resolution, subsampling, data transformations).

General Issues Regarding the Comparability of Biological Assessments

Biological assessments are used to determine if a resource unit (e.g., a stream reach) is meeting its biological potential as specified by its assigned designated use. The degree to which a system is supporting its designated use is measured in terms of an assessment endpoint, which is a description of the specific ecological attributes (or properties) that society wishes to protect (USEPA 1992, Suter 2007). Many types of valued ecological attributes exist, so many types of assessment endpoints are possible. Aquatic life is a very general type of designated use commonly assigned to surficial water bodies in the USA and elsewhere, and considerable research has been directed toward developing numerical endpoint indices to quantify the assessment endpoints associated with this use. Three main types of endpoint indices have been developed to characterize the condition of multispecies assemblages (a community-level attribute of value to many societies) in streams and lakes: 1) biotic indices (Armitage et al. 1983, Hilsenhoff 1987, Lenat 1993), 2) Indices of Biological Integrity (IBIs) and other multimetric indices (MMIs) (e.g., Barbour et al. 1999, Karr and Chu 1999), and 3) observed/expected taxa (O/E) indices (e.g., Hawkins et al. 2000, Wright et al. 2000). Historically, these indices typically were developed with local needs in mind and, until recently, little reason existed to consider their comparability in terms of assessing the biological condition of a water body. However, national and transnational legislation increasingly requires that the assessments used by member states be comparable, if not in assessing equivalent ecological properties, then in the scoring of ecological condition as inferred from each endpoint index in use. In this latter context, the comparability of endpoint indices is the degree to which the application of ≥ 2 indices leads to the same inference regarding the biological condition of a site or set of sites

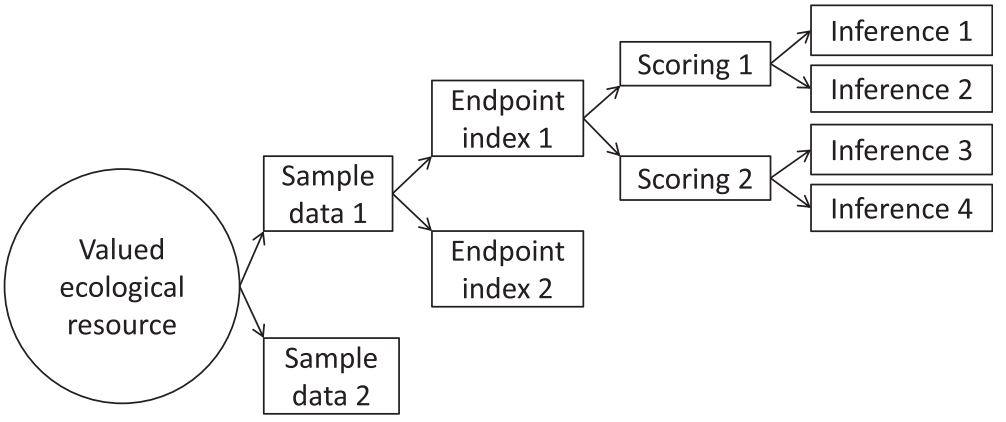

FIG. 1. A conceptual representation of how differences in data collection and treatment at different stages of a bioassessment can affect inferences regarding the biological condition of a site. In the example illustrated, only 2 choices are available for each of only 3 types of factors that can affect inferences (sampling method used to collect raw data, the endpoint index used, and how endpoint index values are scored). The different combinations of 2 sampling methods, 2 types of endpoint indices, and 2 scoring procedures could lead to 16 possible inferences, of which we show only 4. As described in the text, several other types of factors can affect inferences and >2 options usually exist for each factor.

(e.g., Vølstad et al. 2003, Birk and Hering 2006, Buffagni et al. 2006, 2007, Hawkins 2006, Herbst and Silldorff 2006). This definition applies to different types of indices and to situations in which a similar type of index is developed in different regions; e.g., the USA's National Wadeable Stream Assessment was based on 9 different regional MMIs that consisted of different metrics (Stoddard et al. 2008) and 3 different O/E indices (Yuan et al. 2008), all of which were separately calibrated. Within an index type, the authors of the national assessment implicitly assumed that all indices were comparable in how they measured the same assessment endpoint (i.e., biological condition).

Comparability of raw data and derived metrics

Comparability of bioassessment data is essential when combining different sets of data to develop indices (e.g., Astin 2006, Hughes and Peck 2008) or applying a previously developed index to data that were sampled differently (e.g., USEPA 2005a). A key requirement in evaluating data comparability is that comparisons be based on the same biological property (e.g., Cull et al. 1997, Griffith et al. 2006). Biological sampling typically produces a list of taxa and their counts in a sample (i.e., raw data), and the comparability of those raw data can be affected by how and when samples are collected and how samples are processed (e.g., Storey et al. 1991, Cao et al. 2005, Blocksom et al. 2008). Comparability also has been assessed in terms of derived metrics, i.e., assemblage-level attributes, such as taxon richness or the value of various other metrics derived from taxon counts within a sample (Turner and Trexler 1997), or assemblage metrics (e.g., Eaton and Lenat 1991, Lenz and Miller 1996, Houston et al. 2002, Tetra Tech 2005, Friberg et al. 2006, Rehn et al. 2007, Weilhoefer and Pan 2007, Blocksom et al. 2008, Stepenuck et al. 2008).

Factors affecting data comparability in bioassessments

Bioassessments involve many steps that can affect the comparability of inferences regarding the condition of a water body including sampling design; field sampling methods; laboratory processing; index development, analysis, and interpretation; and reporting of biological condition. Most studies on bioassessment comparability conducted to date have focused on the factors that affect comparability of raw data, and other factors that can affect the comparability of bioassessments have been examined less often.

Raw data

The comparability of raw data is influenced by how samples of aquatic biota are collected and processed. Below, we summarize how several factors can influence the comparability of raw data and discuss how comparability of raw data might be measured and improved.

Sampling effort.—The sampling effort applied in both field sampling and laboratory subsampling varies substantially among programs (Carter and Resh 2001) and can affect estimates of taxonomic richness (Gotelli and Graves 1996) and taxonomic composition and relative abundances of taxa in a sample (Cao et al. 1997). When the number of individuals collected varies among samples, investigators can adjust for these differences by taking fixed-count subsamples (e.g., 300 individuals) or by applying rarefaction techniques to the field samples (Sanders 1968, Hurlbert 1971, Gotelli and Graves 1996). For example, Cao et al. (2005) greatly improved the comparability of 2 sets of samples that differed in sampling area and sampling intensity by creating fixed-count subsamples from the original sample counts. However, use of rarefaction or fixed-count samples to estimate differences in taxon richness among sites can lead to misleading estimates of the true differences in taxon richness among sites (e.g., Baltanás 1992, Cao et al. 2002, 2007a, Cao and Hawkins 2005), especially when low fixed counts are used.

Sampling device.—The sampling devices used to collect macroinvertebrates vary markedly among monitoring agencies and programs compared with those used for stream fish and periphyton surveys (ITFM 1994). To what extent are samples collected with different sampling devices comparable when used jointly to develop an endpoint index or when samples collected with one device are applied to an index that was derived from samples collected with another device? Carter and Resh (2001) noted that 12 sampling devices are used in the USA to collect freshwater invertebrates. The number of devices creates a high potential for device-associated problems with data comparability. Many investigators have shown that specific sampling devices have inherent sampling biases (e.g., Elliott and Drake 1981, Mackey et al. 1984, Turner and Trexler 1997, Taylor et al. 2001). These biases affect data comparability in 2 main ways: 1) they capture different numbers of individuals, and 2) they collect different subsets of taxa that occur at a site. Differences in taxon richness associated with different counts potentially can be corrected as described above, but rarefaction or the use of fixed-count samples can do nothing to correct the biases of different devices for capturing different taxa.

Habitat sampled.—Different sampling protocols sometimes target different habitats (e.g., riffles, pools, and edges), which typically support somewhat different assemblages. Therefore, the choice of what

habitat to sample can affect the taxa collected and their relative abundances. For example, Parsons and Norris (1996) collected macroinvertebrate samples from 4 types of habitats and found that samples were clustered in ordination space based on habitat type rather than sites, a result that revealed the dominant effect of habitat type sampled on sample classification. Gerth and Herlihy (2006) and Roy et al. (2003) also observed that the similarity among samples of stream invertebrates was strongly associated with the type of habitats sampled. Thus, the comparability of samples collected from different habitat types can be a significant issue that constrains use of independently collected data sets when compiling data for index development (e.g., Blocksom et al. 2008).

Sampling personnel.—Large-scale and long-term monitoring programs often use many sampling crews to collect samples. If individual operators differ in how many organisms they collect or are biased in the types of microhabitats they sample, data comparability can be compromised. Early studies showed that taxon richness and composition in samples can vary significantly among individual collectors (Furse et al. 1981, Mackey et al. 1984), but more recent analyses indicated that the effect of operators accounted for only 4 to 12% of the variance in taxon richness across samples (Clarke et al. 2003). Cao et al. (2003) and Ostermiller and Hawkins (2004) found that sampling crews were similar in capturing taxa and estimating their relative abundance, results implying that the effects of crew on data comparability may be relatively small compared with effects of other factors if crews are well trained and follow data quality-control procedures.

Mesh size.—The mesh size used to collect stream invertebrates can influence the size composition of the organisms in a sample (Slack et al. 1991, Gage et al. 2000). These differences can affect sample comparability because mesh sizes used in invertebrate sampling devices typically range between 250 and 1200 μm (Carter and Resh 2001). A fine mesh usually captures more small individuals than a coarse mesh. This difference can significantly affect estimates of the detectability of taxa in fixed-count samples and, hence, the comparability of raw data. Tetra Tech (2005) tried to evaluate the effect of different mesh sizes on data comparability by collecting samples with the same type of device but different mesh sizes at several sites. They found that mesh sizes did not affect the separation of sampling sites in ordination space, but their result implied only that the effect of mesh size was smaller than the effect of environmental differences among sites.

Sampling year.—Natural interannual differences in climatic conditions can potentially affect the taxa and their abundances collected from a site (e.g., Bradley and Ormerod 2001, Medeiros and Maltchik 2001, Durance and Ormerod 2007) and, thus, the inferences drawn in any given year regarding the biological condition of a site. Such a situation can arise if biological indices are calibrated with data that were collected during a set of years that do not accurately represent the natural range of climatic conditions observed in a region. For example, if data for index calibration were collected only during years of near-average climatic conditions, any assessment based on data collected during dry or wet years would be potentially confounded by systematic differences among calibration and application data in their climatic conditions. If we ignore (or are ignorant of) such natural climatic variability, the real range of natural biological variability expected at a site probably will be underestimated and the central tendency in climatic conditions probably will be estimated with bias. Either situation probably would lead to inaccurate assessments in some years and inconsistent assessments among years (Mazor et al. 2009, Hawkins et al. 2010b). Ecologists have just begun to understand how natural interannual variability in population sizes and assemblage structure can affect estimates of the reference condition as expressed by biological indices or the inferred biological condition at an assessed site (e.g., Gilbert et al. 2008, Rose et al. 2008).

Sampling season.—Seasonal changes in the composition and abundances of stream biota associated with timing of life-cycle events of individual species (e.g., timing of reproduction, egg hatch, individual growth, and mortality) can potentially affect bioassessments. As in interannual variability, accurate inferences regarding biological condition assume that naturally occurring temporal variability in biota has been incorporated in estimates of the reference condition (i.e., the range of natural variation) or that assessments adjust or control for systematic temporal variation in taxon composition and abundances. Individual monitoring programs often are designed to minimize the effects of such phenological processes on inferences by requiring collection of samples within a specified set of dates (an index window), within which presumably little phenological variation occurs (Barbour et al. 1999). However, in practice, sampling windows typically are selected based on ease of access (e.g., summer base flows) and chronological time rather than developmental (i.e., degree days) time, which should be better correlated with the growth and mortality schedules of ectothermic organisms. If temperatures differ among sites, samples taken on the same date may not be strictly comparable because they were sampled at different physiological times. The length of sampling windows also can differ among programs

regardless of differences in regional climate conditions, a feature that can further compromise comparability. In some large-scale surveys, sampling can occur over several months (e.g., June–September in North America). Many biotic metrics vary substantially among months, e.g., the coefficient of variation (CV) of the metric % mayflies was as high as 106% (J. L. Carter, US Geological Survey, personal communication), and 1 of us (CPH) has documented changes $>10^3$ in mean densities of common stream invertebrate taxa over a 4-mo period.

Several researchers have examined how seasonal differences in sampling can affect bioassessments. For example, Reece et al. (2001) reported that assessments derived from **BE**nthic **A**ssessment **S**edimen**T** (BEAST), an assessment tool based on ordination (Reynoldson et al. 1995), were sensitive to sampling season. Hawkins et al. (2000) and Hawkins (2006) found that sampling date often was a strong predictor of the expected (reference) assemblage in River Invertebrate Prediction and Classification System (RIVPACS)-type models. However, Hose et al. (2004) observed that samples collected over a relatively short period (e.g., 4–6 wk) gave similar O/E values in Australian River Assessment System (AUSRIVAS), an Australian version of RIVPACS. The original British version of RIVPACS avoids this problem by combining samples collected from 3 seasons into 1 composite sample (Moss et al. 1987). Therefore, 3 options are available when dealing with seasonal effects: 1) standardize sampling on a short window of time (Hose et al. 2004), 2) aggregate samples across seasons (e.g., Moss et al 1987, Humphrey et al. 2000), or 3) adjust for the effect of seasonal variation on assemblage composition by modeling (e.g., Hawkins et al. 2000, Hawkins 2006).

Sorting and specimen enumeration.—Sorting of macroinvertebrates from debris can potentially affect the comparability of bioassessments through sorting bias and sorting efficiency. Two sorting methods are used commonly: hand-picking a randomly selected set of individuals from a preserved sample in the laboratory (laboratory sorting) or hand-picking live animals in the field (live sorting). The samples obtained with these 2 techniques do not appear to be comparable when estimating the values of MMIs or O/E indices (Haase et al. 2004, Nichols and Norris 2006). Live sorting tends to be biased toward relatively large, conspicuous, and active taxa relative to laboratory sorting (Nichols and Norris 2006). Laboratory sorting appears to meet the assumption that picked individuals represent a random subsample of the entire sample better than does live sorting. Cao et al. (2003) tested this assumption by comparing subsamples that were hand-picked in the laboratory with those randomly drawn from fully processed samples (i.e., all individuals in a sample were sorted, identified, and counted) by a computer program. They found no significant differences in taxon richness or composition between these 2 types of subsamples. The effects of differences in sorting efficiency on comparability is less well understood, but differences in sorting efficiency should affect estimates of taxon richness metrics because taxon richness usually increases with the number of individuals encountered, especially across the range of fixed-count subsamples frequently used in bioassessments (100–500).

Taxonomic accuracy and resolution.—A primary data-quality objective in bioassessment programs is establishment of a consistent level of taxonomic resolution that can be applied to all samples. Ideally, all specimens in a sample would be identified to the species level, and thus, these data would accurately represent the number of distinct taxa in the sample. The long-standing difficulties in identifying many freshwater organisms to species coupled with differences across programs in the level of expertise among individuals who identify the specimens creates potentially significant problems when comparing bioassessments across programs or when merging data when developing indices. Differences in the specific taxonomic resolution applied to samples occur for several reasons. First, in many programs, specimens of some taxonomic groups (e.g., chironomids, oligochaetes, and water mites) are identified only to coarse levels of taxonomic resolution (e.g., subfamily, family, order, or class) because of cost, whereas in other programs, specimens are identified to the finest level of taxonomic resolution that is practical (Carter and Resh 2001). Second, personnel at taxonomic laboratories often differ in expertise and, thus, in the taxonomic level to which they can identify organisms with confidence. A related and very common problem is that, within a single sample, some individuals of a higher-level taxon (e.g., family) can be identified at a finer taxonomic resolution than others. How such inconsistent resolution in taxonomy is handled depends on what indices are used. When MMIs are developed, richness metrics typically are derived from the number of distinct taxa within individual samples. For example, consider 2 samples. In the 1st sample, of 50 individuals in the family Heptageniidae, 20, 15, and 5 individuals were identified to genus (e.g., *Epeorus*, *Rhithrogena*, and *Cinygmula*) and 10 could be identified only to the family level. This sample would be considered to contain 3 distinct Heptageniidae taxa because the individuals identified to family could very well belong to 1 or more of the genera found in the sample. In a 2nd sample, all Heptageniidae individuals

were identified only to the family level. Thus, we know with certainty only that at least 1 distinct Heptageniidae taxon in the samples should contribute to the estimate of sample richness. The philosophy behind this approach is to retain as much biological information as possible in each sample, i.e., this approach provides an estimate of the minimal richness present in a sample and retains counts of individuals for use in other metrics (e.g., % of the sample that consists of mayfly individuals). However, use of such inconsistent taxonomic resolution among samples compromises estimates of the similarity in taxonomic composition among samples. For example, the calculation of compositional similarity between these 2 samples has to be based on a standard level of taxonomic resolution applied to the Heptageniidae individuals. One option is to consider that the 2 samples have 3 and 0 genera each. The other option is that the 2 samples both contain the same family. We must either assign all individuals to the family level or drop the individuals identified to family from the calculations. This issue is especially problematic for O/E indices, which are based on comparisons of composition, but we suspect the use of variable taxonomic resolution among samples also affects MMIs.

Achieving comparability in taxonomic composition and assemblage structure across sets of samples requires use of a standardized taxonomic resolution that is applied to all samples. Given the variable taxonomic resolution that can exist both within and among samples, such standardization will require that some individuals be assigned to coarser levels of taxonomic resolution than originally assigned or be dropped from the analyses. Such standardization comes at a cost. If use of finer-levels of taxonomic resolution better discriminates ecological signals than does use of coarse levels (Lenat and Resh 2001, Waite et al. 2004, Feio et al. 2006, Hawkins 2006), retaining the finest level of resolution possible would be desirable. However, retaining only the finest-resolved taxa within a group often would result in dropping most individuals from a sample. A more practical goal is to strike a balance between the information gained by including individuals that are identified to finer levels of resolution and the information that would be lost if those individuals identified to a coarser level of resolution were dropped. Cuffney et al. (2007) recently compared 4 methods of assigning potentially ambiguous individuals to taxa and recommended a method similar to the process described above. Indicator species analysis (Dufrêne and Legendre 1997) could be used to determine if personnel at different taxonomic laboratories tend to identify certain taxa more frequently than do personnel at other laboratories. If associations are evident, then the treatment of those taxa should be re-examined and certain taxa might have to be aggregated prior to use in a biological assessment.

To assess the degree of taxonomic comparability among samples, Stribling et al. (2003) proposed the use of Percent Taxonomic Disagreement (PTD)

$$\mathrm{PTD} = \left(1 - \mathrm{comp}_{\mathrm{pos}}/N\right) \times 100$$

where $\mathrm{comp}_{\mathrm{pos}}$ = the number of individuals on which 2 taxonomists (or laboratories) agree in their identifications, and N = the number of individuals that are examined by both taxonomists. Several investigators have used this index to quantify the comparability of taxonomy used when identifying specimens (e.g., USEPA 2005a, Herbst and Silldorff 2006). Alternatively, one can use an index of similarity, such as the Bray–Curtis Index or the Jaccard Coefficient, to measure how consistently samples are treated by 2 taxonomists or laboratories (e.g., Kelly 1999, 2000, Cao et al. 2003).

Error propagation.—Biological surveys typically are based on a variety of sometimes similar, but seldom identical, sampling protocols (Carter and Resh 2001). These sampling protocols often differ in ≥ 1 way, e.g., in habitat types sampled, the number of field replicates taken, and the mesh size used in the sampling device (e.g., ITFM 1994, Houston et al. 2002). The effects of inconsistencies in early steps (e.g., field sampling) on data comparability will be carried through to later steps (e.g., subsampling in the laboratory), and these sequential sources of variability will result in some overall cumulative variability. Knowledge of the relative amount of variability associated with each step could help guide development of quality-control practices designed to minimize variability. However, no methods have been proposed to quantify the cumulative effects of variability in sequential sampling procedures for biological survey data.

Endpoint indices

Inferences regarding ecological condition may not be comparable even if raw sample data are comparable because several factors other than raw data can affect how sites are rated following application of an endpoint index. Some of these factors are general and others are index-specific.

Variation in reference site quality.—Reference conditions typically are used to set biological expectations for individual sites regardless of the type of index used (Norris and Hawkins 2000, Bailey et al. 2004, Stoddard et al. 2006, Hawkins et al. 2010b). In general, variation in

reference-site quality occurs as a consequence of applying different criteria to screening and selection of reference sites across regions (Stoddard et al. 2006, USEPA 2005a). Given the high natural variability among regions and types of water bodies, these criteria (e.g., PO_4 concentrations or conductivity) should be as site-specific as possible (e.g., exceeding natural background levels by <10%). In such ideal cases, all reference sites would have comparable and high quality, i.e., be minimally disturbed sensu Stoddard et al. (2006) or be classified in Tiers 1 or 2 sensu Davies and Jackson (2006). In reality, the level of human alteration of landscapes and waterways varies significantly among and within regions (Herlihy et al. 2008). As a consequence, investigators taking a reference-site approach to developing reference conditions often must adjust screening criteria downward to have a sufficiently large set of sites from which natural variability can be characterized within a region. Such differences in reference-site quality prompted Stoddard et al. (2006) to request that scientists be as explicit as possible regarding the criteria used to define reference-site quality (i.e., historical condition, minimally-disturbed, least-disturbed, or best-attainable conditions). In general, if it is important from either a regulatory, policy, or ecological context that reference benchmarks be based on the same ecological standard, then any difference in the criteria used to define the reference condition will result in incomparable biological assessments.

The USEPA (2005a) recently showed how strongly differences in reference-site quality can affect the comparability of regional assessments. In a comparison of MMIs developed in 3 neighboring states (Virginia, West Virginia, and Maryland), the % stream miles rated as degraded was much higher according to the West Virginia index than according to the Virginia or Maryland indices because the reference sites used to establish the reference condition were generally of higher quality in West Virginia than in Virginia or Maryland.

Effects of the population of calibration sites on index comparability.—The way in which the population of sites used to calibrate indices is selected also can affect the inferences derived from endpoint indices, but this potentially serious issue has received little attention to date. This issue potentially affects all types of indices, but MMIs may be especially sensitive. For example, the data used to develop MMIs typically are derived from either samples collected during probability-based surveys (sensu Stevens 1994) or from a set of sites that were targeted for sampling for ≥1 reasons. In a probabilistic design, every element in a population, such as all stream reaches or lakes in a region, has a chance to be sampled, and site selection is determined by a randomization procedure (Stevens 1994, Hughes et al. 2000, Olsen and Peck 2008). In contrast, sets of targeted sites often represent a human-caused stress gradient (e.g., the US Geological Survey [USGS] National Water Quality Assessment [NAQWA] program; Gilliom et al. 2005) or sites that were independently identified as sites of potential concern.

Sets of randomly selected and targeted sites can represent very different populations of sites and, thus, can affect the calibration of indices. For example, the cumulative frequency distributions (CFDs) of metric values derived from calibration data often are used to scale metric values in a MMI (e.g., Gerritsen et al. 2000, Klemm et al. 2003, USEPA 2006), and this scaling can affect the comparability of both individual metrics and MMIs. We illustrate this effect by showing how CFDs of the values of a metric vary among 3 sets of calibration samples that were collected from the same hypothetical population but that differ in how randomly they were drawn from that population of sites (Fig. 2). A specific percentile value of the CFD must be set prior to determining the discriminatory ability of a metric. Thus, the resulting scaled value of a metric

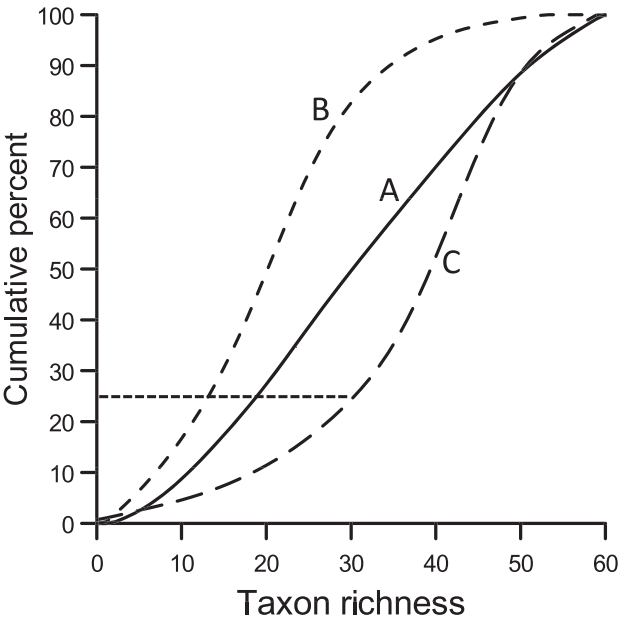

FIG. 2. Taxon richness cumulative frequency distributions (CFD) for 3 hypothetical data sets of samples collected at nonreference sites. Curve A represents samples collected following a targeted design for which sites were selected along a predetermined stress gradient. Curves B and C represent samples collected following 2 probability-based sampling designs in which different types of site stratification were used. The horizontal dashed line intersects each CFD at the 25th percentile of taxon richness values.

depends on the specific distributional properties of the calibration data. For example, if the 25th percentile of all sites in Fig. 2 is used as the lower boundary for scaling taxon richness, the boundary is equal to 20 taxa in set A, but 30 and 15 taxa in sets B and C, respectively. Similarly, the upper boundary also can be set differently. At present, we know little about how such differences in calibration data affect the distribution of metric values, their discriminatory ability, and thus, the comparability of different MMIs. RIVPACS O/E indices should be less prone to differences in the distributional properties of calibration data because these indices are calibrated on reference sites only. However, the scope of O/E indices is a function of how well the calibration data represent the true ecological heterogeneity present in the region of interest.

Effects of method for predicting the reference condition on index comparability.—The application and interpretation of biological indices require comparison of index values observed at an assessed site with the range of values expected under reference conditions (Barbour et al. 1999, Hawkins et al. 2010b). Two general approaches have been used to predict the reference condition: classification and modeling. Classification has been based on a variety of landscape and water-body attributes including ecoregion, thermal regime, channel slope, and water-body size. Such classifications have been commonly used for setting expected values for biotic metrics at potentially impaired sites based on values observed at reference sites in the same stream class (Karr and Chu 1999, Stoddard 2006). Modeling has been based on various statistical methods ranging from traditional linear models to more recently developed prediction-tree approaches such as Random Forests (Breiman 2001). These modeling methods have been used mainly to predict the expected composition and richness of native taxa (Wright et al. 2000, Hawkins 2006), but are being used increasingly to predict how the values of biotic metrics are expected to vary across natural environmental gradients (e.g., Pont et al. 2006, Cao et al. 2007b, Hawkins et al. 2010b). If the method of predicting reference condition (e.g., classification or modeling) affects reference-site index values, then application and interpretation of otherwise identical biological indices may not be directly comparable. At this time, the bioassessment community generally has not recognized how the method used to predict the reference condition can influence index performance (Hawkins et al. 2010b), although recent studies show that prediction methods can differ substantially in both their precision and accuracy and, hence, the comparability of the assessments (e.g., Cao et al. 2007b, Hawkins et al. 2010a).

Effects of spatial and environmental extent on index comparability.—In the USA and elsewhere, biological indices have been developed from data collected from sites that span vastly different scales of both geographic and environmental space (e.g., individual catchments to subcontinents). We know little about how the spatial or environmental extent of calibration data affects the performance and comparability of indices. However, the few studies that have been conducted showed that variation in the spatial extent of calibration data is another factor that can compromise index comparability. For example, Ode et al. (2008) showed that both MMI and O/E indices developed for subregions of California (CA), subregions of the western USA, and the western USA as a whole differed in their performance. These authors found that CA O/E indices were weakly correlated with the other 2 sets of O/E indices (r^2 = 0.32–0.35) and differed in the assessments of 21 to 22% of the test sites. The 3 groups of MMIs were more strongly correlated with one another (r^2 = 0.70–0.76), but the CA indices assessed 25 to 31% more sites as being in nonreference condition than did the other 2 sets of indices. Ode et al. (2008) argued that calibration with more localized data better accounted for the effects of key natural environmental gradients on assemblage composition and, hence, the accuracy in predicting reference state taxa and index values. Aroviita et al. (2009) showed that the performance of O/E indices was affected by the spatial extent of the calibration data. Indices derived from the spatially more extensive data performed less well than indices derived from less extensive data. Differences in performance were more severe for index values derived from a typology approach (classification) to estimating the reference condition than for values derived from a modeling approach.

Effects of other factors on index comparability.—The procedures for developing biological indices typically involve multiple steps. When developing a MMI, setting the ranges of individual metrics and standardizing the values of metrics are important steps (Barbour et al. 1999, Klemm et al. 2003). These 2 steps often are treated together and are referred to as metric scoring. Several scoring methods have been used (e.g., Blocksom 2003, Astin 2006), but reports regarding the effects of scoring method on index comparability are not consistent. For example, Maxted et al. (2000) reported that the detectability of impairment differed for indices derived from 3 different scoring methods. Similarly, Blocksom (2003) found that the choice of scoring method affected both signal-to-noise ratios (i.e., the variability among sites/variability within sites) and the number of reporting classes that could

be statistically discriminated. In contrast, Southerland et al. (2005) reported that the choice of scoring method only slightly affected the assessment of stream condition. Hawkins et al. (2010a) pointed out that any scoring method that does not use the full possible range of metric values probably will compromise the assessment of those sites that are outside of the range of conditions used in calibration.

Comparability of different types of index.—If indices differ in how precisely they can be predicted and measured or how responsive they are to different stressors, they are unlikely to be directly comparable with one another (Birk and Hering 2006, Hawkins 2006, Hawkins et al. 2010a). This issue can be especially problematic when states or countries need to integrate their assessments into a regional or continental context. Birk and Hering (2006) compared how values of 10 biotic indices used in Europe varied across the same set of sites distributed among 7 countries. These indices included Saprobic indices, Biological-Monitoring-Working-Party indices (BMWP), and MMIs. The correlations among these indices varied greatly (Pearson's $r = 0.20–0.77$ for small mountain streams), and those indices calculated based on similar lists of taxa and similar types of tolerance values were most strongly correlated. In the USA, MMI and O/E values can be moderately correlated, e.g., Spearman's $r = 0.70–0.86$ in Herbst and Silldorff (2006) and Pearson's $r = 0.65–0.71$ in Hawkins et al. (2010a). These moderate correlations are consistent with the view that different types of biological indices assess different aspects of ecological condition, i.e., they represent somewhat different ecological properties and, hence, should not be expected to be completely comparable. Much more consideration is needed regarding how different indices characterize different ecological attributes and properties of importance to society (e.g., Davies and Jackson 2006) and whether we should use standard attributes when developing indices.

Reporting class

Even if endpoint indices are comparable, the reporting classes used to summarize and categorize assessments may not be comparable because the threshold values used to distinguish classes can differ and, thus, affect comparisons (Yoder and Davis 1996, Quataert et al. 2004, Birk and Hering 2006). A variety of statistical tests can be used to determine if 2 procedures produce the same assessments across classes, e.g., the Kappa test (Roth et al. 2002, Vølstad et al. 2003, Llansó et al. 2009) for site-level assessments and McNemar's test for regional assessments (Engel and Voshell 2002, Blocksom et al. 2008).

Ideally, the threshold values used to define reporting classes should be based on an understanding of the biological or ecological properties that the endpoint index used was designed to characterize (e.g., USEPA 2005b, Davies and Jackson 2006) and not on arbitrary statistically derived percentiles. However, a statistical approach usually is used to define class boundaries. For example, Barbour et al. (1999) and Shelton and Blocksom (2004) considered a site to be impaired if its index value was $<17^{th}$ percentile of reference-site values. The USEPA (2006) used the 5^{th} percentile of reference-site values as the threshold for inferring impairment. The effects of varying thresholds on the comparability of reporting classes are largely unknown and are difficult to evaluate because thresholds are typically arbitrary and the biological implications of these threshold values have not been addressed.

Methods for Assessing and Improving the Comparability of Raw and Assessment Data

The bioassessment community is aware of the need to evaluate data comparability and has attempted to develop new methods or to adopt methods from other fields for use with biological data (e.g., Diamond et al. 1996, Cao et al. 2005, Buffagni et al. 2007). These methods have not yet been examined critically, and we will show that some methods have been misused. Below, we examine the different methods that have been applied to different types of data and discuss their strengths and limitations.

Assessing the comparability of raw data and derived metrics

When assemblage samples that were collected differently are combined to develop a new index (e.g., a regional O/E index) or an index is applied to samples collected differently from the samples used for calibration, we need to evaluate the potential effects of sampling method on comparability of the raw data.

Direct evaluation of assemblage similarity.—The comparability of the raw data on which bioassessments are based probably is measured most directly by the similarity of assemblage composition and structure between samples. Two main arguments support this approach. First, assemblage similarity estimates comparability in terms of the entire assemblage, and if both taxonomic composition and taxon relative abundances are comparable between 2 sampling methods, any metrics derived from the raw data must be comparable as well. If the raw data are not comparable, many metrics may not be. Second, calculation of assemblage similarities is the starting

point for many multivariate analyses (Green 1980). Therefore, this approach is highly relevant to assessing how sampling method can affect indices based on taxonomic composition (e.g., RIVPACS O/E values).

Application of similarity measures to evaluate sample comparability is conceptually appealing and potentially useful (Storey et al. 1991, Herbst and Silldorff 2006), but their use has a number of potential shortcomings. First, the taxon diversity and composition that occurs at specific sites can influence similarity estimates (Cao et al. 2005). Second, similarity increases with sampling effort (Cao et al. 2005), so direct assessments of comparability among different methods will be difficult. Third, true data comparability always will be underestimated because replicate samples collected with a single method, which should be considered fully comparable, are rarely identical (Wolda 1981, Plotkin and Muller-Landau 2002). Therefore, interpretations of data comparability must be couched in terms of within-method variability. Last, the choice of similarity index used can substantially affect similarity values. Some similarity indices, such as Horn's Index and Morisita's Index, strongly weight abundant taxa, whereas other indices (Canberra Metric and Jaccard Coefficient) weight all taxa equally (see Cao et al. 1997). Use of the former indices would typically yield much higher between-method similarity values than the latter indices because abundant taxa probably will be captured by any sampling method.

Ordinations and cluster analyses.—Many researchers have used a variety of less direct similarity-based approaches to evaluate the comparability of raw data. These approaches include ordination, cluster analysis, and multivariate permutation tests (Furse et al. 1981, Storey et al. 1991, Somerfield and Clarke 1996, Turner and Trexler 1997, Blocksom and Flotemersch 2005, Tetra Tech 2005). Samples typically are collected with multiple methods at several different sites, and these combined sample sets are analyzed. If samples collected with different methods at the same site cluster by site rather than sampling method, the methods are considered comparable in collecting the biological data important in discriminating among sites.

These methods can be used to scan quickly for major effects of sampling method on data comparability, but they do not quantify data comparability, and interpretation is not always straightforward. First, the outcome of multivariate analyses can depend on the specific sites selected for comparisons. If sites are very different in the biota they contain, samples typically will cluster by site. In contrast, if sites are not very different biologically, samples probably will cluster by sampling method. Second, the choice of multivariate technique used and the options selected during the analyses can affect the outcome of analyses. For example, differences among samples collected by different methods may be apparent in a 3-dimensional ordination but not in a 2-dimensional ordination. Last, the comparability that is visually apparent in ordination space cannot be quantified meaningfully because sample scores are all calculated on a relative basis.

Sampling-method comparability.—To accurately quantify the comparability of samples of biota collected with different sampling methods, one must account for the effects of within-site sampling variability and differences in sampling effort. Cao et al. (2005) modified Van Sickle's (1997) method of quantifying classification strength (Mean Similarity Analysis) to meet these requirements. They defined sampling-method comparability (SMC) as

$$\mathrm{SMS} = 100(2S_{\mathrm{between}}/[S1_{\mathrm{within}} + S2_{\mathrm{within}}])$$

where S_{between} is the mean similarity between 2 sets of replicate samples collected with 2 sampling methods, and $S1_{\mathrm{within}}$ and $S2_{\mathrm{within}}$ are the mean similarities between 2 replicate samples collected with methods 1 and 2, respectively. To quantify SMC accurately, several replicates are required for each method. If true replicates are not available, a randomization procedure for resampling can be used to generate a large number of pseudoreplicates. SMC can then be estimated based on the averages of these replicates. When applied to presence–absence data, SMC can be measured with either the Jaccard Coefficient or the Sørensen Index. When applied to abundance data, SMC can be measured with the Bray–Curtis Index. An SMC value of 100 indicates that the 2 methods are entirely comparable, whereas a value of 0 means they are completely noncomparable. SMC measures the overall similarity in assemblage composition or structure, so a high SMC may not imply comparability of a specific biotic metric (Blocksom et al. 2008). In addition, researchers must decide what level of SMC is acceptable based on the specific goals of their studies. Moreover, SMC does not measure the quality of samples, although the estimate of within-method similarity used in the analysis does (Cao et al. 2002).

Direct comparison of metric values.—Most previous assessments of the comparability of biological data have involved the comparison of assemblage attributes (e.g., taxon richness and evenness) and biotic metrics (e.g., % EPT individuals and the number of tolerant taxa) as estimated from different types of sampling method (e.g., Eaton and Lenat 1991, Blocksom and

Flotemersch 2005, Friberg et al. 2006, Blocksom et al. 2008). However, as an assessment of raw data comparability, this approach has at least 2 serious limitations. First, sampling methods often affect the comparability of some metrics more than others, and therefore, the comparability of the overall sample data is uncertain (Cao et al. 2005). Second, some indices, e.g., O/E (Moss et al. 1987, Hawkins 2006) and tolerance-based indices (see Birk and Hering 2006, Clews and Ormerod 2009) require data on the composition of specific taxa observed at each site. Metrics do not provide the information needed for comparing these indices.

Assessing the comparability of endpoint indices

Comparing index values directly.—For indices that share the same range of values, the differences in values of each index observed at the same sites indicate their comparability. Several studies have directly examined the comparability of endpoint indices this way. For example, Herbst and Silldorff (2006) compared 3 MMIs and 3 O/E indices based on samples collected from 40 streams in the Sierra Nevada Mountains. They found that the 3 MMIs yielded very similar values at the same sites and rated similar proportions of the test sites as being in nonreference condition, a pattern that also was observed for the 3 O/E indices they examined. Correlations have been used to assess the comparability of indices that do not share the same range of values (Birk and Hering 2006, Hawkins et al. 2010a).

Other studies have compared the performances of endpoint indices developed in one area and used in another (Houston et al. 2002, USEPA 2005a). For example, the USEPA (2005a) found values produced by applying the West Virginia (WV) MMI and the Virginia (VA) MMI to streams in Maryland were highly correlated ($r = 0.96$; after adjusting for differences in calibration data, sampling method, and taxonomic resolution). The WV and VA MMIs assessed similar proportions (38 and 43%) of the Maryland sites as being in nonreference condition. The Maryland MMI was less strongly correlated with the WV MMI ($r = 0.86$) and rated 53% of the sites as being in nonreference condition.

Use of simulations in evaluating index performance and comparability.—Comparisons based on simulated data of known properties may be the only method available for evaluating the behavior and comparability of different indices. Simulation studies have been used to address general ecological questions for >30 y (e.g., Faith et al. 1987, Minchin 1987), but they have been used in a bioassessment context only recently (e.g., Trebitz et al. 2003, Cao and Hawkins 2005, Mazor et al. 2006, Lamb et al. 2009, Hawkins et al. 2010a). Cao and Hawkins (2005) developed a simple model to simulate biological impairment of macroinvertebrate assemblages caused by stress. Changes in taxon abundances in response to increasing stress were related to differences among taxa in their tolerance to stress as

$$Y_i = X_i(1 - C[1 - TV_i])$$

where X_i = the initial number of individuals in taxon i, TV_i = the tolerance value of taxon i, C = a coefficient that controls the levels of stress, and Y_i is the number of individuals in taxon i after a stress occurs. The basic idea is that assemblage composition and structure can be altered in a realistic way to allow evaluation of how different indices track known alterations in assemblages.

In an extension of the work by Cao and Hawkins (2005), Hawkins et al. (2010a) developed 3 RIVPACS-type O/E indices and 5 MMIs from the same set of data collected from the Interior Columbia River Basin. They evaluated the performance of these indices by simulating stress at 13 different reference sites. They found that O/E values decreased almost linearly with increasing taxon loss or changes in assemblage structure across the entire range of impairment, whereas the MMIs responded similarly up to intermediate levels of impairment but were less responsive to further impairment. These results imply that the comparability between O/E and MMIs may depend on the specific range and magnitude of biological impairment occurring. Hawkins et al. (2010a) argued that the damped response of the MMIs at the highest levels of simulated impairment might be expected if those levels exceeded the levels of impairment in the field samples that were used to calibrate the indices.

Simulations should help researchers assess the performance and comparability of different indices, but their application must be tempered by the recognition that our best simulations are far from perfect representations of the dynamics of real assemblages (e.g., Kenkel 2006, Hurst et al. 2008). For example, the true tolerance of a taxon to a stressor at a specific location probably cannot be known because various natural environmental gradients and species interactions will probably affect its response. Furthermore, the simulations that have been used to date did not incorporate species additions in response to stress, a response that does happen in nature (e.g., Riley et al. 2008, Walters et al. 2008). Accurate simulation of the addition of taxa as environmental conditions are altered will require detailed knowledge regarding the specific

taxa in the regional pool, their dispersal abilities, and the abiotic and biotic factors that affect their establishment.

Assessing and improving the comparability of reporting classes

Performance-based method systems.—A performance-based method system (PBMS) is defined as a system that permits the use of any appropriate sampling and analysis method that demonstrates the ability to meet established data criteria and complies with specific data-quality requirements or data-quality objectives (Diamond et al. 1996). This method was originally developed for chemical analysis, later accepted for toxicity tests (USEPA 1990, ASTM 1995), and subsequently recommended for bioassessment by the Intergovernmental Task Force on Monitoring Water Quality (ITFM 1995) and the Methods and Data Comparability Board (MDCB 2001). Detailed descriptions are available in Diamond et al. (1996) and Barbour et al. (1999). Diamond et al. (1996) outlined a 6-step procedure for applying the PBMS approach to MMIs: 1) Have trained personnel sample replicate reaches or subreaches within a site. 2) Sample ≥ 5 reference sites in the same site class (habitat type, stream sizes, and ecoregions). 3) Compute metrics for each site. 4) Compute precision (variability) for each metric among sites. 5) Repeat step 3 and 4 for ≥ 3 test sites in each site class examined in step 2. Test sites should have different types and apparent levels of impairment. 6) Compare data precision, bias, and method sensitivity for each site class.

Proponents of this approach state that the key to applying this method is to define realistic data-quality objectives in terms of precision, accuracy, sensitivity, and the range of conditions to which the method can be applied. Diamond et al. (1996) and Barbour et al. (1999) defined these performance characteristics at different steps of bioassessment. For example, for evaluating the comparability of sampling devices, they defined precision as the repeatability of an assemblage attribute estimated in a given habitat, bias as exclusion of certain taxa, and performance range as different sampling efficiencies in different habitats. Since its introduction, several studies have used the PBMS approach to evaluate comparability of raw data and reporting classes (e.g., Diamond et al. 1998, Houston et al. 2002, Herbst and Silldorff 2006). Houston et al. (2002) compared 5 indices at 2 Alabama streams following the procedure of Diamond et al. (1996) and found that precision and sensitivity of those indices were similar but that site assessments did not always agree. The latter result implied that the indices were biased relative to one another at some sites. Herbst and Silldorff (2006) showed that different indices were similar in their precision, sensitivity, and relative accuracy as well as in individual site assessments. However, the indices they compared were based on almost identical raw data (SMC values were close to 100% based on our recalculation of their published data) and similar sets of biotic metrics (MMIs) or analysis procedures (RIVPACS-type models). Therefore, the high comparabilities of the indices and site assessments were expected.

In our view, a robust validation of the PBMS approach for bioassessments has not yet been conducted. A key reason is that accuracy, probably the most important performance characteristic in the PBMS, is impossible to evaluate with bioassessment data. This problem arises because the true biological impairment that exists at the physically or chemically altered sites used to calibrate MMIs normally cannot be known with any accuracy or precision prior to calibration (Cao and Hawkins 2005). That is, field ecologists do not have access to the equivalent of stock solutions of known properties with which to calibrate indices or to evaluate their performance. The physical and chemical alterations that are apparent to humans may or may not be of significance to the biota at those sites. Moreover, if these habitat alterations have affected biota, the degree of biological alteration produced by a unit change in habitat conditions will almost certainly vary with the specific ecosystem examined, i.e., the general environmental setting and the suite of specific taxa native to each site (e.g., Leibold et al. 1997, Nydick et al. 2003). Assuming that a certain amount of physical or chemical alteration at a site is equivalent to a certain amount of biological alteration creates a vexing logical dilemma in bioassessment. Bioassessments were developed to provide a direct and independent assessment of biological condition to avoid the problems associated with inferring biological condition from physicochemical assessments (e.g., Hawkes 1979, USEPA 1990, Karr and Chu 1999). Tying the performance of biological assessments to the degree to which they mimic physicochemical alterations severely compromises the unique value and independence of bioassessments.

In an attempt to circumvent this problem, Diamond et al. (1996) substituted the term ''method accuracy'' for accuracy and defined it as the minimum detectable difference between known or assumed impaired sites and reference sites with respect to endpoint indices. In our view, this definition implies sensitivity or discriminative ability rather than accuracy and does not eliminate the fundamental problem of independence. The term *relative accuracy*, defined as the % of sites assumed a priori to be impaired that are

assessed as impaired by an index also has been used as a measure of accuracy (Diamond et al. 1996). However, this concept does not remove the issue of lack of independence discussed above and also suffers from additional issues. For example, consider 2 indices (A and B) that both assess 10 of 20 test sites as impaired, achieving identical relative accuracy (50%). However, if these 2 indices do not agree on the specific sites they assess as impaired, they are not comparable in their site-specific accuracy. When PBMS is applied to indices developed for different regions, this method can be even more problematic because the criteria used to define both reference conditions and impaired conditions probably differ among regions.

In our view, the PBMS approach as applied to bioassessment has additional critical flaws. For example, the precision, sensitivity, and relative accuracy of different indices will almost certainly differ in their statistical properties (e.g., range, frequency distributions). Indices based on data derived from different sampling methods can be similar in some performance characteristics, but not in others. Therefore, estimating overall data comparability would be difficult. In addition, we do not think the PBMS approach is useful for estimating the comparability of reporting classes among different states, the greatest challenge for integrating data into national or regional bioassessments. First, if each of the 50 states in the USA used only 1 index (they often use several), we would potentially need to evaluate the comparability of as many as 1225 pairs of indices ($50!/[50 - 2! \times 2!]$). Second, even if such an exercise were practical, a PBMS could not be applied because each index is developed from specific sets of reference sites and test sites. The criteria for selecting reference sites vary from state to state, as do the type and magnitude of stressors affecting degraded sites. As a result, no performance characteristic is likely to be comparable.

Harmonizing different indices based on a benchmark data set.—Buffagni et al. (2006, 2007) proposed a procedure to harmonize the macroinvertebrate indices used by different countries in Europe. Erba et al. (2009) subsequently improved this procedure. The revised procedure consists of several steps: 1) Use a national standard protocol to sample a set of streams of a given type, e.g., small mountain streams (as defined in the Water Framework Directive) in a country and assign each of those sites to 1 of 5 quality classes (high, good, moderate, poor, bad) based on the national biotic index. 2) Compile a large set of samples collected from streams of the same type across Europe based on the European Union (EU) protocol, and select a subset as reference sites based on analyses of chemical and physical stressors. 3) Calculate the values of 6 widely applicable biotic metrics for each of the national and panEuropean sites and then standardize a metric by dividing each value by the median of the metric at the reference sites. 4) Use the percentiles of the panEuropean MMI (i.e., STAndardisation of River classifications—Intercalibration Common Metrics index or STAR_ICMi) at the reference sites to set boundaries for the 5 quality classes described above (e.g., the 25^{th} percentile as the boundary of high/good). 5) Adjust the threshold of the national index for each type of stream against the STAR_ICMi-based assessment.

This method would be useful for improving the comparability of reporting classes used in different countries if the classification of streams can effectively partition the natural variability of biological assemblages across a large region, such as Europe. If not, the performance of the STAR_ICMi in a specific country can be compromised and so will the whole procedure of harmonization.

Mapping disparate indices to a common biological condition gradient.—Davies and Jackson (2006) proposed a conceptual model, the Biological Condition Gradient (BCG), that used 10 general ecological attributes to describe how biological conditions change across a range of human-caused alterations (pristine to severely degraded). Six of the ecological attributes described aspects of taxonomic composition or taxon relative abundances. One was based on organism condition, 1 on ecosystem function, and 2 on the spatial distribution of ecosystem elements. The potential value of this conceptual model is that it is independent of any specific bioassessment method and geographic region and, therefore, offers a way to "facilitate communication of the current biological condition of a water body compared to natural conditions" (USEPA 2005b; p. 31). In other words, it was designed to standardize the interpretation of biological conditions as measured by different indices and, thus, to allow creation of comparable reporting classes of biological condition. Similar approaches have been used elsewhere, including by the State of Maine (2003) and the State of Ohio (2003). To evaluate the potential utility of this approach, Davies and Jackson (2006) described how consistently ecologists assigned 54 stream macroinvertebrate samples and 58 stream fish samples to 6 different condition tiers defined by differences in taxonomic composition and taxon relative abundances. Participants agreed on 82% of benthic samples and 73% of fish samples.

The information presented by Davies and Jackson (2006) supported the view that such a conceptual model might be able to serve as a general translation

tool. However, application of the BCG to specific regions or types of systems is just beginning to be evaluated. Following guidance presented in USEPA (2005b), Gerritsen and Leppo (2005) developed a BCG model for streams in New Jersey. They concluded that separate BGCs should be developed for 2 broad groups of streams: high- and low-gradient channels. They also made several modifications to the general BCG model to fine-tune how biota in these 2 types of streams were expected to respond to stress. In general, they found that of 6 possible condition tiers, Tier 1 (pristine) sites did not exist in the state and Tiers 3 and 4 were difficult to distinguish from one another. Both of these observations are important in developing translation schemes across regions in which the range of biotic conditions present and the distinctiveness of condition classes can vary.

Theoretically, we can assign all index values to each condition tier by considering how each index relates to the narrative descriptions of each tier. The challenge is how to identify the threshold values for each individual index. If 2 indices both clearly differentiate the 6 tiers based on samples occurring in each tier, a direct transfer is straightforward. If an index poorly differentiates the BCG tiers from one another, a revision of the index may be required (USEPA 2005b). Davies and Jackson (2006) described how Discriminant Function Analysis could be used to select a group of biotic metrics that maximize the separations of BCG tiers. These selected metrics would then be used to develop a new MMI.

In our view, the BCG provides a useful heuristic framework to aid our thinking about how region-specific bioassessments can be integrated or compared across large, heterogeneous regions, but it is not yet clear if the features of the BGC can be quantified in such a way as to allow meaningful and repeatable comparisons of the many biological assessments that are currently based on a varied mix of sampling methods and index types. Some of the nontrivial challenges in implementing an operational BCG, as applied to streams, include: 1) Development of regional BCG models will require consistency in sampling method, habitat sampled, sampling season, sampling effort, and how water bodies are classified (Gerritsen and Leppo 2005, USEPA 2005b). All of these procedures often differ among states and programs (Carter and Resh 2001), and how these inconsistencies will affect the comparability of regional BCG models themselves is not clear. 2) Language in USEPA (2005b) and Davies and Jackson (2006) often describes the taxa expected in a tier as predicted. However, it does not specify how these predictions are made. Gerritsen and Leppo (2005) established criteria for each attribute for each condition tier through use of an expert panel and an iterative process. The panel examined samples from a subset of sites that were impacted to varying degrees and assigned each of these sites to a tier based on each attribute. They then established quantitative or semi-quantitative criteria of each attribute for each tier based on their professional judgment. The degree of consistency among experts described by Davies and Jackson (2006) and Gerritsen and Leppo (2005) is encouraging, but how consistently such criteria can be developed or applied across other regions is unknown. 3) Assignment of a sample to a condition tier in the BCG is based on the values of 10 different attributes. However, it is likely that the values for some attributes will imply assignment to one tier, whereas the values of other attributes will imply assignment to other tiers. Furthermore, different experts may weight these attributes differently in their assignments. For example, Gerritsen and Leppo (2005) reported that a panel of experts in New Jersey weighted some attributes more heavily when differentiating certain tiers. If theoretical justification of the weighting of each attribute is not possible, some form of guidance will be needed to ensure consistency in the weightings.

A Brief Summary Guide to Checking Data Comparability

As discussed above, data comparability can be affected in several ways. Here, we present 2 flow diagrams that should aid practitioners who are faced with combining data from multiple sources when developing or applying MMI and O/E indices (Fig. 3). The 2 diagrams differ slightly because the 2 types of index differ somewhat in their data requirements and in how they are calibrated, which affects the factors that can influence data comparability.

The initial step in determining data comparability requires an assessment of the criteria used to define the site conditions used to calibrate the indices—i.e., the reference condition for both MMIs and O/E indices and the degraded conditions for MMIs. If the criteria used to select sites in the different data sets are not similar, the criteria will have to be adjusted (standardized) so they apply to all sites. Otherwise the candidate sites will have to be screened to remove sites that do not meet the desired criteria.

For MMI development, next confirm that the data sets that will be used to calibrate the index were drawn from the same population of sites in the same way (i.e., random selection or targeted). If they were not, the sets of samples will have to be reconciled as described by Overton et al. (1993) and Astin (2007). For O/E indices, this step involves checking that the

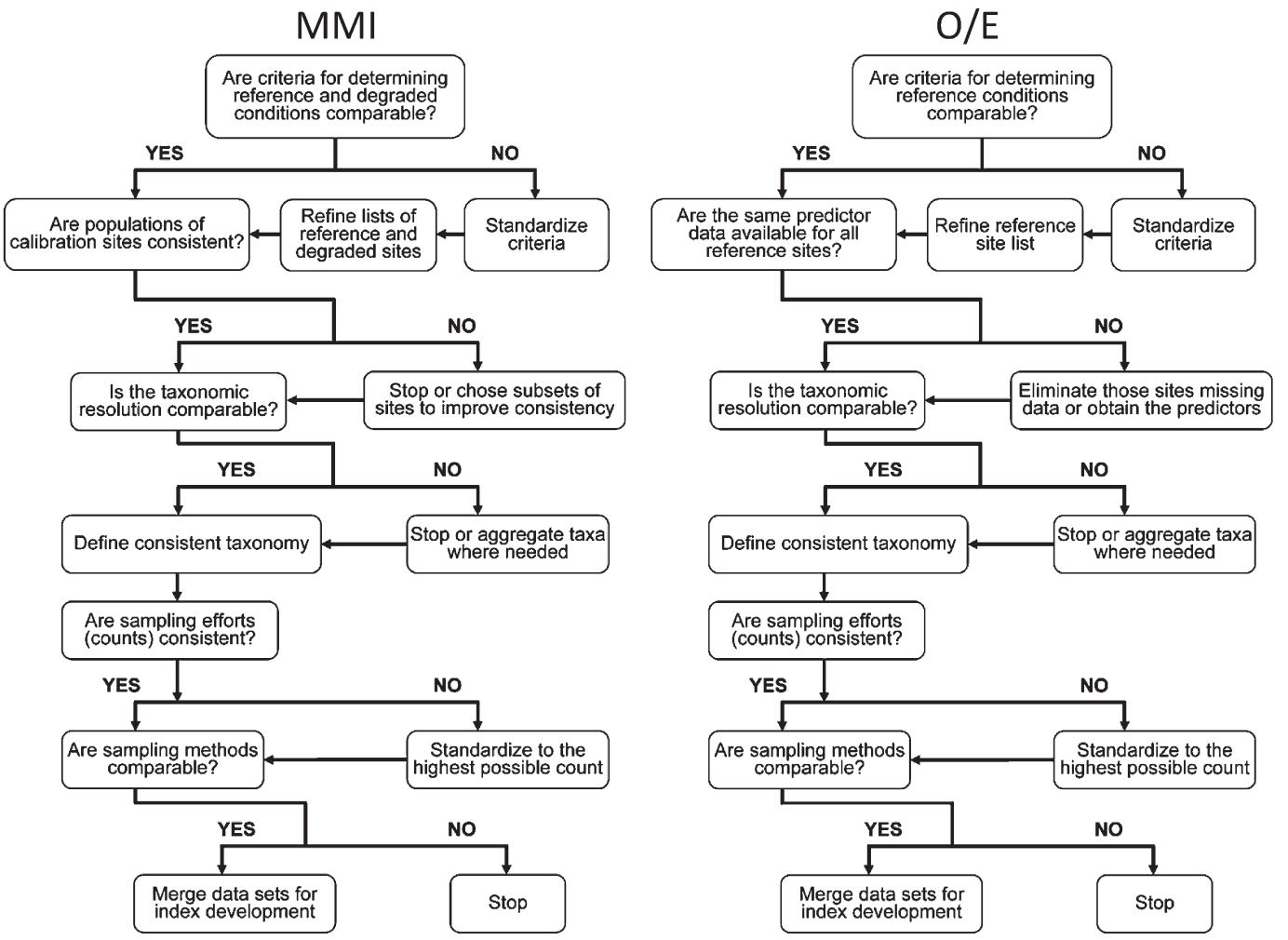

FIG. 3. Flow diagrams showing the major steps at which decisions regarding data comparability should be made when merging raw data for use in developing and applying multimetric indices (MMIs) and observed/expected (O/E) indices.

sites in the collective data set represent the natural gradients of interest and that the same predictor variables are available for all sites. If predictor variables are lacking for some sites, 3 options exist: drop the sites from index development, drop the use of ≥ 1 predictors and use a reduced set of predictors for modeling, or obtain the predictor information for those sites lacking predictors.

The remaining samples will then have to be screened for taxonomic consistency. Depending on the outcome of this screening, samples may have to be dropped or the taxonomic resolution used in the lowest-resolution data set will have to be applied to all samples. For example, chironomid midges are identified only to family or subfamily in many data sets. To integrate data from these samples with data in which midges were more highly resolved will require aggregation of midge genera to the higher-level taxon.

Integrating data also will require that data were collected with similar sampling and subsampling efforts because differences in effort will affect many metrics and O/E values. If differences in effort are largely associated with sample counts, computer resampling can be used to standardize counts. The cost associated with such resampling is that the counts used will be the lowest counts observed among all samples. If samples differ in the area sampled, resampling of counts will not completely resolve differences in taxon richness metrics associated with area sampled (e.g., Carlisle and Hawkins 2008).

Last, if different methods were used to collect samples, the comparability of the methods must be assessed. If different sampling methods target different habitats or are differentially effective at capturing different taxa from the same habitat type, it is unlikely that the samples can be suitably harmonized for the

purpose of index development and application. O/E indices may be less sensitive to differences in sampling method than MMIs. O/E indices are based on presence–absence data, which may not vary as much between habitats or sampling methods as taxon counts, which are used in constructing some metrics. If the outcome of the evaluation at any of these steps is not satisfactory, the data combination effort should be terminated.

Portions of these flow diagrams also are applicable when applying an index developed from one set of data to samples collected in different ways. In these cases, some steps in the flow charts can be skipped. For example, when an index developed from samples collected with method A is applied to a set of samples collected with method B, the steps evaluating effects of reference-site criteria are not relevant.

Concluding Remarks

Data comparability remains a vexing problem for those entities interested in maximizing the use of previously collected data when comparing assessments across either ecological or jurisdictional regions, developing region-specific biological indices (e.g., O/E and MMIs), or integrating previously conducted site assessments into regional statements of the status of freshwater biota. In this review, we have summarized the existing scientific literature regarding data comparability and identified those issues we think are in most need of attention. The bioassessment community is aware of many of the general problems that affect our ability to integrate data sets, but significant conceptual issues remain to be resolved. Once we agree on definitions regarding data comparability, we will be able to address in a more meaningful way some of the practical questions that currently constrain our general ability to mix and match data sets for the purpose of generating statistically and ecologically more robust and geographically more extensive bioassessments.

Acknowledgements

This project was funded by contract MP5045 from the USEPA to CPH and YC. At the request of the USEPA Office of Water, 14 individuals provided critical review of draft versions of this document (Mike Southerland, Sam Stribling, John Van Sickle, Leska Fore, David Herbst, Erick Silldorff, Karen Blocksom, Greg Pond, LeAnne Astin, Gil Dichter, Ellen Dickey, Bob Hughes, David P. Larsen, and Mark Vinson). Their comments and criticisms greatly improved the manuscript. Laura Gabanski of the USEPA Office of Water provided overall guidance and many useful suggestions. We thank Bruce Chessman and 2 anonymous referees for additional constructive suggestions.

Literature Cited

ARMITAGE, P. D., D. MOSS, J. F. WRIGHT, AND M. T. FURSE. 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research 17: 333–347.

AROVIITA, J., H. MYKRÄ, T. MUOTKA, AND H. HÄMÄLÄINEN. 2009. Influence of geographical extent on typology- and model-based assessments of taxonomic completeness of river macroinvertebrates. Freshwater Biology 54: 1774–1787.

ASTIN, L. 2006. Data synthesis and bioindicator development for nontidal streams in the interstate Potomac River Basin, USA. Ecological Indicators 6:664–685.

ASTIN, L. 2007. Developing biological indicators from diverse data: the Potomac Basin-wide Index of Benthic Integrity (B-IBI). Ecological Indicators 7:895–908.

ASTM (AMERICAN SOCIETY OF TESTING AND MATERIALS). 1995. Biological effects and environmental fate. Volume 11.05. Annual book of standards. American Society of Testing and Materials, Philadelphia, Pennsylvania.

BAILEY, R., R. H. NORRIS, AND T. B. REYNOLDSON. 2004. Bioassessment of freshwater ecosystems: using the reference condition approach. Springer, New York.

BALTANÁS, A. 1992. On the use of some methods for the estimation of species richness. Oikos 65:484–492.

BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Rapid bioassessment protocols for use in wadeable streams and rivers: periphyton, benthic macroinvertebrates, and fish. EPA 841-B-99-002. US Environmental Protection Agency, Washington, DC.

BIRK, S., AND D. HERING. 2006. Direct comparison of assessment methods using benthic macroinvertebrates: a contribution to the EU Water Framework Directive intercalibration exercise. Hydrobiologia 566:401–415.

BLOCKSOM, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands stream. Environmental Management 31:670–682.

BLOCKSOM, K. A., B. C. AUTREY, M. PASSMORE, AND L. REYNOLDS. 2008. A comparison of single and multiple habitat protocols for collecting macroinvertebrates in wadeable streams. Journal of the American Water Resources Association 44:1–17.

BLOCKSOM, K. A., AND J. E. FLOTEMERSCH. 2005. Comparison of macroinvertebrate sampling methods for nonwadeable streams. Environmental Monitoring and Assessment 102:243–262.

BRADLEY, D. C., AND S. J. ORMEROD. 2001. Community persistence among stream invertebrates tracks the North Atlantic Oscillation. Journal of Animal Ecology 70:987–996.

BREIMAN, L. 2001. Random forests. Machine Learning 45: 5–32.

BUFFAGNI, A., S. ERBA, M. CAZZOLA, J. MURRAY-BLIGH, H. SOSZKA, AND P. GENONI. 2006. The STAR common metrics approach to the WFD intercalibration process: full application for small, lowland rivers in three European countries. Hydrobiologia 566:379–399.

BUFFAGNI, A., S. ERBA, AND M. T. FURSE. 2007. A simple procedure to harmonize class boundaries of assessment systems at the pan-European scale. Environmental Science and Policy 10:709–924.

BUFFAGNI, A., AND M. FURSE. 2006. Intercalibration and comparison—major results and conclusions from the STAR project. Hydrobiologia 566:357–364.

CAO, Y., AND C. P. HAWKINS. 2005. Simulating biological impairment for evaluating ecological indicators. Journal of Applied Ecology 42:954–965.

CAO, Y., C. P. HAWKINS, D. P. LARSEN, AND J. VAN SICKLE. 2007a. Effects of sample standardization on mean species detectabilities and estimates of relative differences in species richness among assemblages. American Naturalist 170:381–395.

CAO, Y., C. P. HAWKINS, J. OLSEN, AND M. NELSON. 2007b. Modelling natural environmental gradients improves the accuracy and precision of diatom-based indicators for Idaho streams. Journal of the North American Benthological Society 26:566–585.

CAO, Y., C. P. HAWKINS, AND A. W. STOREY. 2005. A method for measuring the comparability of different sampling methods used in biological surveys: implications for data integration and synthesis. Freshwater Biology 50: 1105–1115.

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.

CAO, Y., W. P. WILLIAMS, AND A. W. BARK. 1997. Effects of sample size (number of replicates) on similarity measures in river Aufwuchs community analysis. Water Environment Research 69:107–114.

CAO, Y., D. D. WILLIAMS, AND D. P. LARSEN. 2002. Comparisons of ecological communities: the problem of sample representativeness. Ecological Monographs 72:41–56.

CARLISLE, D. M., AND C. P. HAWKINS. 2008. Land use and the structure of western US stream invertebrate assemblages: predictive models and ecological traits. Journal of the North American Benthological Society 27:986–999.

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, R. J. M. GUNN, J. M. WINDER, AND J. F. WRIGHT. 2003. Sampling variation in macroinvertebrate data and implications for river quality indices. Freshwater Biology 47:1735–1751.

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

CLEWS, E., AND S. J. ORMEROD. 2009. Appraising riparian management effects on benthic macroinvertebrates in the Wye River system. Aquatic Conservation: Marine and Freshwater Ecosystems 20:73–81.

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

CULL, C. A., S. E. MANLEY, I. M. STRATTON, H. A. W. NEIL, I. S. ROSS, R. R. HOLMAN, R. C. TURNER, AND D. R. MATTHEWS. 1997. Approach to maintaining comparability of biochemical data during long-term clinical trials. Clinical Chemistry 43:1913–1918.

DAVIES, S., AND S. JACKSON. 2006. The biological condition gradient: a descriptive model for interpreting changes in aquatic ecosystems. Ecological Applications 16:1251–1266.

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

DIAMOND, J., J. STRIBLING, AND C. O. YODER. 1998. Determining comparability of bioassessment methods and their results. Proceedings of the NWQMC National Monitoring Conference, July 7–9, Reno, Nevada. (Available from: http://acwi.gov/monitoring/conference/98proceedings/Papers/27-diam.htm)

DUFRÊNE, M., AND P. LEGENDRE. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

DURANCE, I., AND S. J. ORMEROD. 2007. Climate change effects on upland stream macroinvertebrates over a 25-year period. Global Change Biology 13:942–957.

EATON, L. E., AND D. R. LENAT. 1991. Comparison of a rapid bioassessment method with North Carolina's qualitative macroinvertebrate collection method. Journal of the North American Benthological Society 10:335–338.

ELLIOTT, J. M., AND C. M. DRAKE. 1981. A comparative study of 7 grabs used for sampling benthic macroinvertebrates in rivers. Freshwater Biology 11:99–120.

ENGEL, S. R., AND J. R. VOSHELL. 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? American Entomologists 48:164–177.

ERBA, S., M. T. FURSE, R. BALESTRINI, A. CHRISTODOULIDES, T. OFENBOCK, W. VAN DE BUND, J. G. WASSON, AND A. BUFFAGNI. 2009. The validation of common European class boundaries for river benthic macroinvertebrates to facilitate the intercalibration process of the Water Framework Directive. Hydrobiologia 633:17–31.

EUROPEAN COMMISSION. 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327, 22.12.2000, 1–72.

FAITH, D. O., P. R. MINCHIN, AND L. BELBIN. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68.

FEIO, M. J., T. B. REYNOLDSON, AND M. A. S. GRAÇA. 2006. The influence of taxonomic level on the performance of a

predictive model for water quality assessment. Canadian Journal of Fisheries and Aquatic Sciences 63: 367–376.

FRIBERG, N., L. SANDIN, M. T. FURSE, S. E. LARSEN, R. T. CLARKE, AND P. HAASE. 2006. Comparison of macroinvertebrate sampling methods in Europe. Hydrobiologia 566:365–378.

FURSE, M. T., J. F. WRIGHT, P. D. ARMITAGE, AND D. MOSS. 1981. An appraisal of pond-net samples for biological monitoring of lotic macroinvertebrates. Water Research 15:679–689.

GAGE, J. D., D. J. HUGHES, AND J. L. GONZALEZ VECINO. 2000. Sieve size influence in estimating biomass, abundance and diversity in samples of deep-sea macrobenthos. Marine Ecology Progress Series 225:97–107.

GAO (GENERAL ACCOUNTING OFFICE). 2002. Water quality: inconsistent state approaches complicate Nation's efforts to identify its most polluted waters. GAO-02-186. US General Accounting Office, Washington, DC.

GERRITSEN, J., J. BURTON, AND M. BARBOUR. 2000. A stream condition index for West Virginia wadeable streams. Technical Report. Region 3, Environmental Services Division, Office of Science and Technology, Office of Water, US Environmental Protection Agency, Washington, DC. (Available from: http://www.dep.wv.gov/WWE/watershed/bio_fish/Documents/WVSCI.pdf)

GERRITSEN, J., AND E. W. LEPPO. 2005. Biological condition gradient for tiered aquatic life use in New Jersey. Region 2, Office of Science and Technology, US Environmental Protection Agency, and New Jersey Department of Environmental Protection, Trenton, New Jersey. (Available from: http://www.state.nj.us/dep/wms/bwqsa/FINAL TALU NJ RPT_2.pdf)

GERTH, W. J., AND A. T. HERLIHY. 2006. Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. Journal of the North American Benthological Society 25:501–512.

GHETTI, P. F., AND G. BONAZZI. 1977. A comparison between various criteria for the interpretation of biological data in the analysis of the quality of running waters. Water Research 11:819–831.

GILBERT, B., P. J. DILLON, K. M. SOMERS, R. REID, AND L. SCOTT. 2008. Response of benthic macroinvertebrate communities to El Niño related drought events in six upland streams in south-central Ontario. Canadian Journal of Fisheries and Aquatic Sciences 65:890–905.

GILLIOM, R. J., W. M. ALLEY, AND M. E. GURTZ. 2005. Design of the National Water-Quality Assessment Program: occurrence and distribution of water-quality conditions. US Geological Survey Circular 1112. US Geological Survey, National Water-Quality Assessment (NAWQA) Program, Reston, Virginia. (Available from: http://pubs.usgs.gov/circ/circ1112/)

GOTELLI, N. J., AND G. R. GRAVES. 1996. Null models in ecology. Smithsonian Institution Press, Washington, DC.

GREEN, R. H. 1980. Multivariate approaches in ecology: the assessment of ecological similarity. Annual Review of Ecology and Systematics 11:1–14.

GRIFFITH, J. F., L. A. AUMAND, I. M. LEE, C. D. MCGEE, L. L. OTHMAN, K. L. RITTER, K. O. WALKER, AND S. B. WEISBERG. 2006. Comparison and verification of bacterial water quality indicator measurement methods using ambient coastal water samples. Environmental Monitoring and Assessment 116:335–344.

HAASE, P., S. LOHSE, S. PAULS, K. SCHINDEHUTTE, A. SUNDERMANN, P. ROLAUFFS, AND D. HERING. 2004. Assessing streams in Germany with benthic invertebrates: development of a practical standardized protocol for macroinvertebrate sampling and sorting. Limnologica 34:349–365.

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

HAWKES, H. A. 1979. Invertebrates as indicators of river water quality. Pages 2:1–2:45 *in* A. James and L. Evison (editors). Biological indicators of water quality. John Wiley and Sons, Chichester, UK.

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., Y. CAO, AND R. ROPER. 2010a. Method of predicting reference conditions affects the performance and interpretation of ecological indices. Freshwater Biology 55:1066–1085.

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 Application 10:1456–1477.

HAWKINS, C. P., J. R. OLSEN, AND R. A. HILL. 2010b. The reference condition: predicting benchmarks for ecological and water-quality assessment. Journal of the North American Benthological Society 29:312–343.

HEINZ CENTER. 2002. The state of the Nation's ecosystems: measuring the lands, waters, and living resources of the United States. Cambridge University Press, Washington, DC.

HERBST, D. B., AND E. L. SILLDORFF. 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.

HERLIHY, A. T., S. G. PAULSEN, J. VAN SICKLE, J. L. STODDARD, C. P. HAWKINS, AND L. YUAN. 2008. Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale. Journal of the North American Benthological Society 27:860–877.

HILSENHOFF, W. L. 1987. An improved biotic index of organic stream pollution. Great Lakes Entomologist 20:31–39.

HOSE, G., E. TURAK, AND N. WADDELL. 2004. Reproducibility of AUSRIVAS rapid bioassessments using macroinvertebrates. Journal of the North American Benthological Society 23:126–139.

HOUSTON, L., M. T. BARBOUR, D. LENAT, AND D. PENROSE. 2002. Multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. Ecological Indicators 1:279–292.

HUGHES, R. M., S. G. PAULSEN, AND J. L. STODDARD. 2000. EMAP-surface water: a multi-assemblage, probability

survey of ecological integrity in the U.S.A. Hydrobiologia 422/423:429–443.

HUGHES, R. M., AND D. V. PECK. 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. Journal of the North American Benthological Society 27:837–859.

HUMPHREY, C. L., A. W. STOREY, AND L. THURTELL. 2000. AUSRIVAS: operator sample processing errors and temporal variability—implications for model sensitivity. Pages 143–165 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of fresh waters. Freshwater Biological Association, Ambleside, UK.

HURLBERT, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577–586.

HURST, C. P., C. P. CATTERALL, AND J. CHASELING. 2008. A comparison of two methods for generating artificial multi-assemblage ecological datasets. Ecological Informatics 3:286–294.

ITFM (INTERGOVERNMENTAL TASK FORCE ON MONITORING WATER QUALITY). 1994. Report of the Interagency Biological Methods Workshop. U.S. Geological Survey Open-file Report 94-490. US Geological Survey, Reston, Virginia.

ITFM (INTERGOVERNMENTAL TASK FORCE ON MONITORING WATER QUALITY). 1995. Strategy for improving water-quality monitoring in the United States. Final Report of the Intergovernmental Task Force on Monitoring Water Quality. U.S. Geological Survey Open-file Report 95-742. US Geological Survey, Reston, Virginia.

KARR, J. R., AND E. W. CHU. 1999. Restoring life in running waters: better biological monitoring. Island Press, Washington, DC.

KELLY, M. G. 1999. Progress towards quality assurance of benthic diatom and phytoplankton analyses in the UK. Pages 208–215 *in* J. Prygiel, B. A. Whitton, and J. Bukowska (editors). Use of algae for monitoring rivers III. Agence de l'Eau Artois-Picardie, Douai, France, (Available from: Agence de l'Eau Artois-Picardie, 200, Rue Marceline, Centre Tertiaire de l'Arsenal BP818, F-59508 DOUAI CEDEX, France.)

KELLY, M. G. 2000. Use of similarity measures for quality control of benthic diatom samples. Water Research 35: 2784–2788.

KENKEL, N. C. 2006. On selecting an appropriate multivariate analysis. Canadian Journal of Plant Science 86:663–676.

KLEMM, D. L., K. A. BLOCKSOM, F. A. FULK, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. V. PECK, J. L. STODDARD, W. T. THOENY, M. B. GRIFFITH, AND W. S. DAVIS. 2003. Development and evaluation of a Macroinvertebrate Biotic Integrity Index (MBII) for regionally assessing Mid-Atlantic Highlands streams. Environmental Management 31:656–669.

LAMB, E. G., E. BAYNE, G. HOLLOWAY, J. SCHIECK, S. BOUTIN, J. HERBERS, AND D. L. HAUGHLAND. 2009. Indices for monitoring biodiversity change: are some more effective than others? Ecological Indicators 9:432–444.

LEIBOLD, M. A., J. M. CHASE, J. B. SHURIN, AND A. L. DOWNING. 1997. Species turnover and regulation of trophic structure. Annual Review of Ecology and Systematics 28:467–494.

LENAT, D. R. 1993. A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assigning water-quality rating. Journal of the North American Benthological Society 12:279–290.

LENAT, D. R., AND V. H. RESH. 2001. Taxonomy and stream ecology: the benefits of genus- and species-level identifications. Journal of the North American Benthological Society 20:287–298.

LENZ, B. N., AND M. MILLER. 1996. Comparison of aquatic macroinvertebrate samples collected using different field methods. Fact Sheet FS-216-96. National Water-Quality Assessment Program, US Geological Survey, Madison, Wisconsin. (Available from: http://wi.water.usgs.gov/pubs/FS-216-96/fs-216-96.pdf)

LLANSÓ, R. J., J. H. VØLSTAD, D. M. DAUER, AND J. R. DEW. 2009. Assessing benthic community condition in Chesapeake Bay: does the use of different benthic indices matter? Environmental Monitoring and Assessment 150: 119–127.

MACKEY, A. P., D. A. COOLING, AND A. D. BERRIE. 1984. An evaluation of sampling strategies for qualitative surveys of macro-invertebrates in rivers, using pond nets. Journal of Applied Ecology 21:515–534.

MAXTED, J. R., M. T. BARBOUR, J. GERRITSEN, V. PORETTI, N. PRIMROSE, A. SILVIA, D. PENROSE, AND R. RENFROW. 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19:128–144.

MAZOR, R. D., A. H. PURCELL, AND V. H. RESH. 2009. Long-term variability in bioassessments: a twenty-year study from two northern California streams. Environmental Management 43:1269–1286.

MAZOR, R. D., T. B. REYNOLDSON, D. M. ROSENBERG, AND V. H. RESH. 2006. Effects of biotic assemblage, classification, and assessment method on bioassessment performance. Canadian Journal of Fisheries and Aquatic Sciences 63: 394–411.

MDCB (METHODS AND DATA COMPARABILITY BOARD). 2001. Towards a definition of performance-based laboratory methods: a position paper developed by the Methods and Data Comparability Board (MDCB). Technical Report 01-02. National Water Quality Monitoring Council. (Available from: http://acwi.gov/methods/pubs/perf_pubs/nwqmc.0102.pdf)

MDCB (METHODS AND DATA COMPARABILITY BOARD). 2003. The value of data comparability. Fact Sheet. National Water Quality Monitoring Council. (Available from: http://wi.water.usgs.gov/methods/about/publications/valcomp_fs.pdf)

MEDEIROS, E. S. F., AND L. MALTCHIK. 2001. Fish assemblage stability in an intermittently flowing stream from the Brazilian semiarid region. Austral Ecology 26:156–164.

MINCHIN, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69: 89–107.

MOSS, D., M. T. FURSE, J. F. WRIGHT, AND P. D. ARMITAGE. 1987. The prediction of the macroinvertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology 17:41–52.

NICHOLS, S. J., AND R. H. NORRIS. 2006. River condition assessment may depend on the sub-sampling method: field live-sort versus laboratory sub-sampling of invertebrates for bioassessment. Hydrobiologia 572:195–213.

NORRIS, R. H., AND C. P. HAWKINS. 2000. Monitoring river health. Hydrobiologia 435:5–17.

NYDICK, K. R., B. M. LAFRANCOIS, J. S. BARON, AND B. M. JOHNSON. 2003. Lake specific responses to elevated atmospheric nitrogen deposition in the Colorado Rocky Mountains, U.S.A. Hydrobiologia 510:103–114.

ODE, P. R., C. P. HAWKINS, AND R. D. MAZOR. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. Journal of the North American Benthological Society 27:967–985.

OLSEN, A. R., AND D. V. PECK. 2008. Survey design and extent estimates for the Wadeable Streams Assessment. Journal of the North American Benthological Society 27:822–836.

OSTERMILLER, J. D., AND C. P. HAWKINS. 2004. Effect of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. Journal of the North American Benthological Society 23:363–382.

OVERTON, J. M., T. C. YOUNG, AND W. S. OVERTON. 1993. Using 'found' data to augment a probability sample: procedure and case study. Environmental Monitoring and Assessment 26:65–83.

PARSONS, M., AND R. H. NORRIS. 1996. The effect of habitat-specific sampling on biological assessment of water quality using a predictive model. Freshwater Biology 36: 419–434.

PLOTKIN, J. B., AND H. C. MULLER-LANDAU. 2002. Sampling the species composition of a landscape. Ecology 83: 3344–3356.

PONT, D., B. HUGUENY, U. BEIER, D. GOFFAUX, A. MELCHER, R. NOBLE, C. ROGERS, N. ROSET, AND S. SCHMUTZ. 2006. Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. Journal of Applied Ecology 43:70–80.

QUATAERT, P., J. BREINE, AND I. SIMOENS. 2004. Comparison of the European Fish Index with the standardised European model, the spatially based models (eco-regional and European), and existing methods. Institute for Forestry and Game Management, Groenendaal-Hoeilaart, Belgium.

REECE, P. F., T. B. REYNOLDSON, J. S. RICHARDSON, AND D. M. ROSENBERG. 2001. Implications of seasonal variation for biomonitoring with predictive models in the Fraser River catchment, British Columbia. Canadian Journal of Fisheries and Aquatic Sciences 58:1411–1418.

REHN, A. C., P. R. ODE, AND C. P. HAWKINS. 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. Journal of the North American Benthological Society 26:332–348.

REYNOLDSON, T. B., R. C. BAILEY, K. E. DAY, AND R. H. NORRIS. 1995. Biological guidelines for freshwater sediment based on **BE**nthic **A**ssessment **S**ediemen**T** (the BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20:198–219.

RILEY, L. A., M. F. DYBDAHL, AND R. O. HALL. 2008. Invasive species impact: asymmetric interactions between invasive and endemic freshwater snails. Journal of the North American Benthological Society 27:509–520.

ROSE, P., L. METZELING, AND S. CATZIKRIS. 2008. Can macroinvertebrate rapid bioassessment methods be used to assess river health during drought in south eastern Australian streams? Freshwater Biology 53:2626–2638.

ROTH, N., J. VØLSTAD, G. MERCURIO, AND M. SOUTHERLAND. 2002. Biological indicator variability and stream monitoring integration: a Maryland case study. EPA/903-02/008. Office of Environmental Information and Mid-Atlantic Integrated Assessment Team, US Environmental Protection Agency, Fort Meade, Maryland.

ROY, T. V., C. T. ROBINSON, AND G. W. MINSHALL. 2003. Development of macroinvertebrate-based index for bioassessment of Idaho streams. Environmental Management 27:627–636.

SANDERS, H. L. 1968. Marine benthic diversity: a comparative study. American Naturalist 102:243–282.

SHELTON, A. D., AND K. A. BLOCKSOM. 2004. A review of biological assessment tools and biocriteria for rivers and streams in New England states. EPA/600/R-04/168. US Environmental Protection Agency, Cincinnati, Ohio.

SLACK, K. V., L. J. TILLEY, AND S. S. KENNELLY. 1991. Mesh-size effects on drift sample composition as determined with a triple net sampler. Hydrobiologia 209:215–226.

SOKAL, R. R., AND F. J. ROHLF. 1987. Biometry: the principles and practice of statistics in biological research. W. H. Freeman and Company, San Francisco, California.

SOMERFIELD, P. G., AND K. R. CLARKE. 1996. A comparison of some methods commonly used for collection of sublittoral sediments and their associated fauna. Marine Environmental Research 43:145–156.

SOUTHERLAND, M., G. ROGERS, M. KLINE, R. MORGAN, D. BOWARD, P. KAZYAK, R. KLAUDA, AND S. STRANKO. 2005. New biological indicators to better assess the condition of Maryland streams. Monitoring and Non-Tidal Assessment Division, Maryland Department of Natural Resources, Annapolis, Maryland. (Available from: http://www.dnr.state.md.us/streams/pdfs/ea-05-13_new_ibi.pdf)

STATE OF MAINE. 2003. Code of Maine Rules 06-096. Chapter 579: Classification attainment evaluation using biological criteria for rivers and streams. Department of Environmental Protection, Augusta, Maine. (Available from: http://www.maine.gov/dep/blwq/docmonitoring/biomonitoring/material.htm)

STATE OF OHIO. 2003. Ohio Administrative Code at OAC Chapter 3745-1. State of Ohio Water Quality Standards. State of Ohio, Columbus, Ohio. (Available from: http://www.epa.state.oh.us/portals/35/rules/01-00_eff031510.pdf)

STEPENUCK, K. F., R. L. CRUNKILTON, M. A. BOZEK, AND L. WANG. 2008. Comparison of macroinvertebrate-derived stream quality metrics between snag and riffle habitats. Journal of the American Water Resources Association 44:670–678.

STEVENS, D. 1994. Implementation of a national monitoring program. Journal of Environmental Management 42: 1–29.

STODDARD, J. 2006. Use of ecological regions in aquatic assessments of ecological condition. Environmental Management 34:61–70.

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. P. LARSEN, C. P. HAWKINS, R. K. JOHNSON, AND R. H. NORRIS. 2006. Setting expectations for the ecological condition of running waters: the concept of reference conditions. Ecological Applications 16:1267–1276.

STOREY, A. W., D. H. D. EDWARD, AND P. GAZEY. 1991. Surber and kick sampling: a comparison for the assessment of macroinvertebrate community structure in streams of south-western Australia. Hydrobiologia 211:111–121.

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

SUTER, G. W. 2007. Ecological risk assessment. 2nd edition. CRC Press, Boca Raton, Florida.

TAYLOR, B. W., A. R. MCINTOSH, AND B. L. PECKARSKY. 2001. Sampling stream invertebrates using electroshocking techniques: implications for basic and applied research. Canadian Journal of Fisheries and Aquatic Sciences 58: 437–445.

TETRA TECH. 2005. Comparability analysis of benthic macroinvertebrate methods in Montana. A report to Montana Department of Environmental Quality, Helena, Montana, prepared by Tetra Tech, Inc., Owings Mills, Maryland. (Available from: http://www.epa.gov/region8/water/monitoring/MTCompReport20050711A.pdf)

TREBITZ, A. S., B. H. HILL, AND F. H. MCCORMICK. 2003. Sensitivity of indices of biotic integrity to simulated fish assemblage changes. Environmental Management 32: 499–515.

TURNER, A. M., AND J. C. TREXLER. 1997. Sampling aquatic invertebrates from marshes: evaluating the options. Journal of the North American Benthological Society 16:694–709.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 1990. Methods for measuring the acute toxicity of effluents and receiving waters to aquatic organisms. 4th edition. EPA/600-4-90-0027. Office of Research and Development, US Environmental Protection Agency, Cincinnati, Ohio.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 1992. Framework for ecological risk assessment. EPA/630/R-92/001. Risk Assessment Forum, US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2003. Draft report on the environment 2003. EPA 260-R-02-006. US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2005a. Proof of concept for integrating bioassessment results from three states probabilistic monitoring programs. EPA-903-R-05-003. Office of Environmental Information, US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2005b. Use of biological information to better define designated aquatic life uses in state and tribal water quality standards: tiered aquatic life uses. Draft. EPA-822-R-05-001. Health and Ecological Criteria Division (4304T), Office of Science and Technology, Office of Water, 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, Office of Water, US Environmental Protection Agency, Washington, DC.

VAN SICKLE, J. 1997. Using mean similarity dendrograms to evaluate classification. Journal of Agricultural, Biological, and Environmental Statistics 2:370–388.

VØLSTAD, J., N. ROTH, M. SOUTHERLAND, AND G. MERCURIO. 2003. Pilot study for Montgomery County and Maryland DNR data integration: comparison of benthic macroinvertebrate sampling protocols for freshwater streams. EPA/903/R-03/005. Office of Environmental Information and Mid-Atlantic Integrated Assessment Team, US Environmental Protection Agency, Fort Meade, Maryland.

WAITE, I. R., A. T. HERLIHY, D. P. LARSEN, N. S. URQUHART, AND D. J. KLEMM. 2004. The effects of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, U.S.A. Freshwater Biology 49:474–489.

WALTERS, D. M., M. J. BLUM, B. RASHLEIGH, B. J. FREEMAN, B. A. PORTER, AND N. M. BURKHEAD. 2008. Red shiner invasion and hybridization with blacktail shiner in the upper Coosa River, USA. Biological Invasions 8:1229–1242.

WEILHOEFER, C. L., AND Y. PAN. 2007. A comparison of diatom assemblages generated by two sampling protocols. Journal of the North American Benthological Society 26:308–318.

WOLDA, H. 1981. Similarity indices, sample size and diversity. Oecologia (Berlin) 50:296–302.

WRIGHT, J. F., D. W. SUTCLIFFE, AND M. T. FURSE. 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

YODER, C. P., AND G. D. DAVIS. 1996. The Ohio EPA bioassessment comparability project: a preliminary analysis. Ohio EPA Technical Bulletin MAS/1996-12-4. Ohio Environmental Protection Agency, Columbus, Ohio. (Available from: http://www.epa.ohio.gov/portals/35/documents/biocomp.pdf)

YUAN, L. L., C. P. HAWKINS, AND J. VAN SICKLE. 2008. Effects of regionalization decisions on an O/E index for the US national assessment. Journal of the North American Benthological Society 27:892–905.

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

Received: 14 May 2010
Accepted: 5 April 2011