*Ecological Applications*, 16(4), 2006, pp. 1277–1294
© 2006 by the the Ecological Society of America

# QUANTIFYING BIOLOGICAL INTEGRITY BY TAXONOMIC COMPLETENESS: ITS UTILITY IN REGIONAL AND GLOBAL ASSESSMENTS

CHARLES P. HAWKINS[1]

*Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Aquatic, Watershed, & Earth Resources, Utah State University, Logan, Utah 84322-5210 USA*

*Abstract*. Water resources managers and conservation biologists need reliable, quantitative, and directly comparable methods for assessing the biological integrity of the world's aquatic ecosystems. Large-scale assessments are constrained by the lack of consistency in the indicators used to assess biological integrity and our current inability to translate between indicators. In theory, assessments based on estimates of taxonomic completeness, i.e., the proportion of expected taxa that were observed (observed/expected, $O/E$) are directly comparable to one another and should therefore allow regionally and globally consistent summaries of the biological integrity of freshwater ecosystems. However, we know little about the true comparability of $O/E$ assessments derived from different data sets or how well $O/E$ assessments perform relative to other indicators in use. I compared the performance (precision, bias, and sensitivity to stressors) of $O/E$ assessments based on five different data sets with the performance of the indicators previously applied to these data (three multimetric indices, a biotic index, and a hybrid method used by the state of Maine). Analyses were based on data collected from U.S. stream ecosystems in North Carolina, the Mid-Atlantic Highlands, Maine, and Ohio.

$O/E$ assessments resulted in very similar estimates of mean regional conditions compared with most other indicators once these indicators' values were standardized relative to reference-site means. However, other indicators tended to be biased estimators of $O/E$, a consequence of differences in their response to natural environmental gradients and sensitivity to stressors. These results imply that, in some cases, it may be possible to compare assessments derived from different indicators by standardizing their values (a statistical approach to data harmonization). In situations where it is difficult to standardize or otherwise harmonize two or more indicators, $O/E$ values can easily be derived from existing raw sample data. With some caveats, $O/E$ should provide more directly comparable assessments of biological integrity across regions than is possible by harmonizing values of a mix of indicators.

*Key words: biological assessment of freshwater ecosystems; biological indices; Clean Water Act; conservation; harmonization; indicators of biological integrity; modeling; monitoring; multimetrics; pollution; RIVPACS; water quality.*

## INTRODUCTION

There is a critical need to assess the biological status of the world's freshwater ecosystems and determine whether conditions are improving or declining (Revenga and Kura 2003). This need was anticipated in the United States almost 30 years ago when the modern Clean Water Act was created (1972, amended in 1977), which requires that states and tribal nations monitor and assess the biological integrity of their waters. Biological integrity was defined by Frey (1977:128) as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having a species composition, diversity, and functional organization comparable to that of the natural habitat of the region." This is the definition used by the U.S. Environmental Protection Agency (USEPA) when providing guidance to states and tribes regarding bioassessment programs (*available online*).[2] There is not consensus, however, on how it should or can be measured.

Over the last two decades there has been considerable work devoted to the development of biological indicators for use in assessing the biological integrity of freshwater ecosystems (USEPA 2002*b*), and many states in the United States and several countries have active biological monitoring and assessment programs. How-

Manuscript received 25 October 2004; revised 19 August 2005; accepted 23 August 2005. Corresponding Editor: E. H. Stanley. For reprints of this Invited Feature, see footnote 1, p. 1249.

[1] E-mail: chuck.hawkins@usu.edu

[2] ⟨http://www.epa.gov/bioindicators/html/biointeg.html⟩

PLATE 1. Indicators of taxonomic completeness require site-specific estimates of the taxa expected under specific natural environmental settings. The models used to derive these estimates are calibrated with data collected at a series of reference sites that represent the range of natural conditions within a region of interest. Sampling methods that adequately characterize the biota at a site are required to obtain accurate and precise models. The photo shows Scott Rollins, a Ph.D. student from Michigan State University, completing sampling on the Verde River, Arizona, USA, as part of an effort to derive reference conditions for streams in the western United States. Photo credit C. P. Hawkins.

ever, the independent development of assessment methods by different political jurisdictions has resulted in the use of a large mix of indicators about which we have little knowledge regarding their comparability. This issue is particularly problematic given the emerging need in the United States, Europe, and elsewhere to integrate multiple assessments conducted at small scales into regional- or national-level assessments. For example, assessments made by the states in the United States are supposed to be summarized by the U.S. Environmental Protection Agency in bi-annual reports that describe the status and trends of the Nation's water quality (e.g., USEPA 2002*a*). However, meaningful summaries have been impossible because of insufficient or incompatible data (U.S. General Accounting Office 2000, Heinz Center 2002, USEPA 2003).

Incompatibility between assessments can occur for two reasons: (1) biota are sampled in different ways and (2) we use different indicators to measure biological condition (e.g., Houston et al. 2002, Davies and Jackson 2006). In the United States three main types of indicators are commonly used to measure biological condition: biotic indices, multimetric indices, and measures of taxonomic completeness. There are at least two reasons why these different types of indicators may yield different inferences regarding the biological status of a water body: (1) they are based on different ideas of what biological condition is, and (2) they differ in how expected values are derived.

Biotic indices (BI) measure the average pollution tolerance of taxa found at a site and are typically calculated as $\Sigma \mathrm{TV}_I \times n_i / N$, where $\mathrm{TV}_i =$ the tolerance value of taxon $i$, $n_i =$ abundance of taxon $i$, and $N =$ the total number of individuals in the sample. Biotic indices are based on the idea that unpolluted water bodies contain many pollution-sensitive taxa (low tolerance values), whereas polluted water bodies contain mostly pollution-tolerant taxa (e.g., Chutter 1972, Hilsenhoff 1987). A low BI value implies high biological integrity. Until recently, most tolerance values used to estimate BI values were derived by comparing how abundances of different taxa vary across gradients of known or presumed stress or water quality (e.g., Lenat 1993). Biotic indices are used as the main indicator of biological quality in several countries (see overview by Johnson et al. [1993]) and in at least one U.S. state. In the United States, biotic indices are used most often as one of the component metrics in a multimetric index.

Multimetric indices (MMI) were conceived as a way to quantify Frey's (1977) concept of biological integrity, e.g., Karr's (1981) index of biological integrity (IBI). Index values are calculated by summing the standardized values of several different types of individual metrics (e.g., richness, tolerance, composition, guild

structure) derived from a sample of organisms. The selection of metrics used in the index can be based on either a conceptual understanding of what attributes are biologically important (e.g., Karr's original IBI) or identification of that subset of the many possible metrics that best discriminates between reference and degraded water bodies (e.g., Barbour et al. 1999). Assessments are made by comparing observed MMI values to expected values that are derived from an appropriate set of reference sites (sensu Stoddard et al. 2006). MMI values that fall within the range of expected values imply high biological integrity, whereas values lower than that observed at reference sites imply biological degradation. Although the original MMI approach most closely parallels Frey's concept of biological integrity, it is not clear that all MMIs will lead to directly comparable inferences. Regional differences in the fauna or flora, which will affect the set of metrics used in MMIs (e.g., Fore et al. 1996, Klemm et al. 2002), and differences in the intensity of stress at non-reference sites used to calibrate MMIs might affect their comparability.

Measures of taxonomic completeness are based on estimates of the difference between observed ($O$) and expected ($E$) taxonomic composition. In the most widespread implementation of this idea (e.g., Moss et al. 1987, Hawkins et al. 2000, Simpson and Norris 2000, Wright 2000), the ratio, $O/E$, represents the proportion of predicted taxa that were observed in a sample. $O/E$ values near 1 imply high biological integrity and values $<1$ imply biological degradation. $O/E$ quantifies a fundamental component of ecological capital, one of the three general indicators that the National Research Council identified as critical to monitor (NRC 2000). Given that $O$ is always a subset of $E$ (the predicted taxa), it is a measure of the integrity of the *native* biota. Values near 1 are consistent with descriptions of those biological attributes characteristic of biological integrity in the highest quality tier of the biological condition gradient described by Davies and Jackson (2006). A unique property of $O/E$ assessments is that, unlike biotic indices and multimetric indices, values are not derived from or calibrated against any stressor gradient. Instead, empirical models that relate taxonomic composition to naturally occurring environmental gradients are developed from data collected at a set of reference-quality sites that differ in their natural environmental setting (see Plate 1). These models are then used to predict what the probability of capturing (PC) each taxon in the regional taxon pool would be at specific sites if those sites were in reference condition. The expected number of taxa ($E$) at a site is then estimated as $\Sigma PC_i$ for given PC threshold values (e.g., 0, 0.1, 0.5, etc.), where $i =$ each taxon in the region of interest. $O$ is that set of taxa with PC greater than the specified threshold value that were collected in a sample. The performance of $O/E$-based assessments is therefore largely dependent on how well models predict the PC for different taxa under different environmental settings (Clarke et al. 2003). Models might differ in their accuracy and precision because of differences in the predictor variables used, the methods used to select predictor variables, differences among models in the taxonomic resolution applied to the biota being modeled, methods used to sample biota, decisions regarding the PC threshold to use when calculating $E$ and $O$, or decisions regarding what subset of the biota inhabiting a water body are used in assessments (e.g., Hawkins et al. 2000, Ostermiller and Hawkins 2004).

A hybrid assessment method has also been developed by the Maine Department of Environmental Protection (DEP) that uses discriminant-function models to predict the a priori, legally defined water-quality classes to which samples belong from the values of 1–9 biological metrics and indices measured in samples (Davies et al. 1995). Although this method shares some of the predictive machinery used by $O/E$ models (discriminant-function models), it is similar to BI and MMI methods in that the assessments are calibrated by training models to discriminate between samples collected from reference-quality and a-priori-defined degraded sites.

Given the marked differences between many programs in assessment methods and indicators, there are two possible approaches to the synthesis of existing data for the purpose of creating larger, regional assessments. In one approach, a system might be developed for translating among different types of indicators. Davies and Jackson (2006) provide a conceptual framework, the biological condition gradient, that provides qualitative guidance regarding the biological attributes that should be considered when making such translations. An alternative approach is to use a single indicator that is general enough to measure what the other indicators measure, can be easily applied to all data sets, and thus avoid the need to develop translation functions. Because $O/E$ is based on the raw compositional data from which other indicators are derived, it might serve as such a universal indicator if project specific effects on estimates of $O/E$ do not compromise its inter-project comparability.

In this paper I examine the potential use of $O/E$ as a universal indicator of biological integrity. To do so, I compare the performance of $O/E$ assessments with that of three other types of indicators: MMI, BI, and the Maine DEP methods of assessment. I examine performance of both $O/E$ and the other indicators in terms of indicator bias and precision and sensitivity to stressors. To further evaluate the robustness and comparability of $O/E$-based assessments, I also examine how variable reference-site $O/E$ values were across years, how taxonomic resolution used in models affected values, and if the type of sampling method used to collect samples of biota affected $O/E$ assessments. I conclude by discussing both the potential advantages of $O/E$ as a means of providing standardized assessments across regions as well as the pitfalls associated with its use.

TABLE 1. Comparison, by taxon, of different assessment measures with $O/E$, the ratio of observed to expected taxonomic composition.

| Taxon, U.S. source, and habitat† | Assessment measure‡ | Data-set samples (mean ± SD) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Calibration ($C$) | Validation ($V$) | Test | $C$–$T$ | 10th% C§ | %T < 10th% C§ |
| Invertebrates | | | | | | | |
| North Carolina, MH | | (208) | (202) | (984) | | | |
| | raw NCBI | 3.86 ± 0.97 | 4.14 ± 0.90 | 6.03 | | 5.16 | 77 |
| | SNCBI | 1.00 ± 0.16 | 0.96 ± 0.15 | 0.65 | 0.35 | 0.79 | 77 |
| | ASNCBI | 1.00 ± 0.09 | 1.01 ± 0.07 | 0.72 | 0.28 | 0.90 | 80 |
| | $O/E_{sp,0}$ | 0.99 ± 0.16 | 1.03 ± 0.14 | 0.70 | 0.29 | 0.78 | 61 |
| | $O/E_{sp,0.5}$ | 1.01 ± 0.14 | 0.98 ± 0.13 | 0.62 | 0.39 | 0.83 | 78 |
| | $O/E_{g,0}$ | 0.99 ± 0.15 | 1.03 ± 0.14 | 0.72 | 0.27 | 0.82 | 64 |
| | $O/E_{g,0.5}$ | 1.01 ± 0.13 | 0.98 ± 0.11 | 0.65 | 0.36 | 0.83 | 77 |
| | $O/E_{f,0}$ | 1.00 ± 0.13 | 1.03 ± 0.13 | 0.77 | 0.23 | 0.85 | 64 |
| | $O/E_{f,0.5}$ | 1.01 ± 0.10 | 1.00 ± 0.08 | 0.73 | 0.28 | 0.87 | 72 |
| Mid-Atlantic Highlands, FW | | (72) | (14) | (456) | | | |
| | raw MIBI | 77.3 ± 14.2 | 75.3 ± 13.0 | 52.7 | 24.6 | 55.1 | 50 |
| | SMIBI | 1.00 ± 0.18 | 0.98 ± 0.17 | 0.68 | 0.32 | 0.71 | 50 |
| | $O/E_0$ | 1.00 ± 0.19 | 0.96 ± 0.20 | 0.78 | 0.22 | 0.74 | 38 |
| | $O/E_{0.5}$ | 1.01 ± 0.17 | 0.98 ± 0.16 | 0.64 | 0.37 | 0.77 | 67 |
| Maine, AS | | (64) | (20) | (452) | | | |
| | $O/E_0$ | 1.01 ± 0.26 | 0.98 ± 0.24 | 0.78 | 0.23 | 0.64 | 33 |
| | $O/E_{0.5}$ | 1.00 ± 0.30 | 1.08 ± 0.23 | 0.72 | 0.28 | 0.60 | 38 |
| Ohio, AS and MH | | (58) | (34) | (322) | | | |
| | raw ICI | 42.8 ± 8.57 | 42.1 ± 7.69 | 33.3 | 9.5 | 30.0 | 35 |
| | SICI | 1.00 ± 0.20 | 0.98 ± 0.18 | 0.78 | 0.22 | 0.70 | 35 |
| | $O/E_0$ | 1.03 ± 0.25 | 1.04 ± 0.20 | 0.90 | 0.13 | 0.71 | 25 |
| | $O/E_{0.5}$ | 1.04 ± 0.16 | 1.01 ± 0.16 | 0.80 | 0.24 | 0.79 | 44 |
| Fish | | | | | | | |
| Ohio, MH | | (114) | (0) | (1438) | | | |
| | raw IBI | 46.6 | | 37.0 | | 33.5 | 40 |
| | SIBI | 1.00 | | 0.82 | 0.18 | 0.75 | 40 |
| | $O/E_0$ | 0.99 | | 0.82 | 0.17 | 0.75 | 39 |
| | $O/E_{0.5}$ | 1.02 | | 0.80 | 0.22 | 0.77 | 44 |

*Notes:* The numbers in parentheses indicate sample sizes. The proximity of the mean value for calibration ($C$) and validation ($V$) data sets to 1 is a measure of global accuracy. Precision is reported as the standard deviation of values obtained from reference sites (calibration and validation data sets). Sensitivity is reported as the difference between mean values obtained from test ($T$) and calibration samples.

† Habitats sampled: MH, multiple habitats; FW, fast-water habitats; AS, artificial substrates.

‡ Where possible, the different measures were standardized (SNCBI, standardized North Carolina biotic index; SMIBI, standardized macroinvertebrate index of biotic integrity; SIBI, standardized [Ohio] index of biotic integrity; SICI, standardized [Ohio] invertebrate-community index) so the mean of calibration samples = 1 to allow direct comparison with $O/E$ values. In the case of the NCBI, values were inverted prior to standardization so that decreasing values implied increasing biological degradation. SNCBI values were further adjusted (ASNCBI) for the effects of four factors (latitude, longitude, distance from source, and calendar day). ASNCBI values are the residuals from the regression of SNCBI values on latitude, longitude, log distance from source, and calendar day. Original residual values had a mean of zero but were incremented by 1 to allow direct comparison with SNCBI and $O/E$ values. The $O/E$ models based on zero and 0.5 probabilities of capture are denoted as $O/E_0$ and $O/E_{0.5}$. Models based on species, genus, and family levels of taxonomic resolution are identified with sp, g, and f subscripts.

§ The percentage of test sites whose assessment values were below the 10th percentile of calibration sample values (%T < 10th% C) was used to show how model precision and sensitivity jointly influence the power of detection of biological impairment.

## MATERIALS AND METHODS

### *Data sets*

I based analyses on five data sets (Table 1). These data included samples of stream benthic invertebrates and associated habitat information from four regions: North Carolina (NCDENR 2003), Maine (Davies and Tsomides 2002), the Mid-Atlantic Highlands, a region that spans several states (Klemm et al. 2002), and Ohio (Ohio EPA 1989). I also included one fish data set from Ohio (Ohio EPA 1989). Invertebrates were identified to the lowest taxon possible in all data sets, including chironomid midges, which were identified to genus or species level. Fish were identified to species. Each data set contained samples collected at reference sites that were used to derive expected conditions at other sites (Stoddard et al. 2006) and samples from a series of test sites that varied in the degree to which they were exposed to stressors and thus in their potential amount of biological impairment. I provide only a brief description of these data sets here. Full descriptions are available in the original reports cited above.

For each data set, I built RIVPACS-type predictive models (Moss et al. 1987, Wright 2000) following the procedures described in Hawkins et al. (2000), Hawkins and Carlisle (2001), and Van Sickle et al. (2005). $O/E$ values were calculated based on two probability-of-capture thresholds: PC > 0 and PC > 0.5. These two

thresholds have been used elsewhere and essentially represent assessments based on either all taxa including those that are expected to be extremely rare (i.e., PC $>$ 0) or only those taxa that are expected to be moderately common at a site (PC $>$ 0.5). Ostermiller and Hawkins (2004) discuss the statistical and biological reasons why use of an intermediate PC threshold such as 0.5 may have advantages over the inclusion of all taxa. I then compared the performance of these $O/E$ assessments with those based on the indicators originally used for each data set.

*North Carolina*.—These data were collected by the North Carolina Department of Environment and Natural Resources and consisted of 208 samples used to calibrate models, 202 validation samples, and 984 "test" samples from potentially impaired sites. Samples were based on multi-habitat, qualitative collections of invertebrates. North Carolina bases biological assessments on the North Carolina biotic index (NCBI), which is calculated from tolerance values assigned to each of the taxa as described above. Tolerance values range from 0 to 10 with low values implying less tolerance to stress.

Because the taxonomic resolution in this data set was exceptionally good, I used this data set to examine if taxonomic resolution affected the performance of $O/E$ assessments by constructing predictive models based on species-, genus-, and family-levels of taxonomic resolution. Six to eight predictive variables were used in these three models, which included elevation, stream width, stream depth, percentage boulder substrate, percentage rubble substrate, calendar day, latitude, longitude, and catchment area. Van Sickle et al. (2005) describe general aspects of the species-level model. Because of the long period of record covered by this data set, I also used this data set to determine if estimates of reference condition were affected by the year in which data were collected. Such an effect could bias assessments if $O/E$ values, or other indicators, were developed from data collected over a restricted period and then applied to data collected in other years. I also examined if $O/E$ values and the NCBI were differentially sensitive to year effects.

*Mid-Atlantic Highlands (MAH)*.—These data were collected in conjunction with the U.S. Environmental Protection Agency (EPA)'s EMAP program (Herlihy et al. 2000). For this study, I used invertebrate data collected from 542 fast-water (riffle) habitats. The U.S. EPA has constructed a multimetric index (the macroinvertebrate index of biological integrity, MIBI) for both riffle and pool habitats (Klemm et al. 2002). The MIBI, which I consider here, included seven individual metrics: mayfly, stonefly, caddis fly, and collector-filterer richness; a biotic index; percentage non-insect individuals; and percentage individuals in the top five dominant taxa. Richness values were adjusted for catchment area, and the overall index was standardized by the authors to scale from 0 to 100, where 100 = the best biological condition.

I constructed a predictive model from the same set of data used to construct the MIBI. Data from 86 reference sites (72 calibration, 14 validation) were used to build the model, and it was applied to 456 test sites. Six variables were used in the predictive model: North Central Appalachian Mountains ecoregion (1 or 0 [present or not]), Central Appalachian Ridge and Valley ecoregion (1 or 0), calendar day, elevation, carbonate concentration, and catchment area. Van Sickle et al. (2005) describe general statistical aspects of the model. Stressor data were also available for many of these sites, which allowed me to compare sensitivities of both $O/E$ and the MIBI to variation in those factors likely causing biological impairment.

*Maine*.—Maine recognizes four aquatic-life use categories (AA, A, B, and C), of which classes AA and A are the highest quality waters defined as having "aquatic life as naturally occurs" (Davies et al. 1995: Table 1), class B includes waters that receive discharges but experience no "detrimental" biological change, and class C includes waters in which discharges may alter assemblage composition but assemblage structure and function are maintained. Waters that do not meet the minimal standards for Class C are grouped in a non-attainment (NA) class. Predictions are derived from a set of hierarchical discriminant-function models in which biological metrics are the predictors of class membership. The Maine Department of Environmental Protection biological data are based on samples collected from artificial substrates (rock-filled baskets, bags, or cones), which are allowed to colonize for about 28 days before they are collected.

I analyzed data from 84 reference-quality samples (64 calibration, 20 validation) and 452 test sites that were collected between 1974 and 1997. Model building resulted in selection of five predictor variables: elevation, distance from stream source (DFS), latitude, the number of freeze-free days, and calendar day. Because Maine uses artificial substrates to collect invertebrate samples, I was also able to compare model performance derived from this type of sampling with the performance of models based on samples collected from natural habitats.

*Ohio*.—The Ohio Environmental Protection Agency assesses their rivers and streams with both an invertebrate-community index (ICI) and an index of biological integrity (IBI) based on fish samples (Ohio EPA 1989). For this paper I used data from only those samples for which I could build and apply predictive models. The number of samples that I could use in model building was also restricted by the number of sites for which predictor variables were available. For comparisons based on invertebrates, I used data from 58 reference calibration sites, 34 reference validation samples, and 322 test-site samples. Ohio uses two sampling methods for collecting invertebrates: Hester-Dendy, multiplate

artificial substrates and qualitative kick-net samples from multiple natural habitats. Because Ohio combines these data in their macroinvertebrate ICI, I also used the combined data. However, I also developed preliminary $O/E$ models based only on data derived from Hester-Dendy samplers to assess the performance of models derived solely from artificial substrates. For comparisons based on fish, I used data from 114 reference sites and 1438 test-site samples. No validation samples were used when this model was built. Stressor data were available for a subset of these samples. Six predictor variables were used in the invertebrate model: river basin (Maumee River Basin, 0/1), calendar day (calendar day), ecoregion (Western Allegheny Plateau, 0/1), average relative humidity, log slope of the sampled reach, and log drainage area above the sampling location. Four variables were used in the fish predictive model: latitude, longitude, log catchment area, and log slope of the sampled reach (see also de Zwart et al. 2006).

### *Measures of performance*

The performance of any bioassessment method can be characterized in three ways: precision, bias, and sensitivity to stressors. Evaluation of such criteria is a straightforward process when known standards can be applied under controlled conditions. However, evaluation of the performance of biological indicators is complicated by the fact that the real degree of biological degradation (changes in community structure and function) at a site can never be fully known, i.e., we cannot know how impaired a site is and in all the ways it is impaired prior to sampling it (Cao and Hawkins 2005). We therefore have to compare methods against surrogate measures of biological impairment (e.g., presence of stressors) or against one another and then use indirect means of judging the performance of different methods relative to one another.

I quantified *precision* as the standard deviation (SD) or coefficient of variation (CV) of indicator values derived from the population of reference sites used to establish expected conditions at assessed sites (see Stoddard et al. 2006). Ideally, the only variation in reference-site values would be associated with sampling error, which, if minimized by adequate sampling, would allow detection at test sites of small deviations from expected condition. Because comparisons of precision can be confounded by use of different units of measurement, I standardized all reference site NCBI (North Carolina biotic index) and IBI values to have a mean of 1, the expected mean reference site $O/E$ value derived from predictive models. Raw index values for test sites were then divided by the mean of reference- site values to put NCBI and IBI assessments in the same units of measure as $O/E$. Standard deviations based on such standardized values are equivalent to the CV calculated from raw values. I could not conduct a similar standardization with the Maine assessments because their assessment endpoints are categorical.

For the NCBI, I also adjusted the standardized NCBI values (SNCBI) for environmental setting by calculating the residual values obtained after applying a multiple-regression equation to all sample data. This equation described the effect of naturally occurring environmental variables on SNCBI values and was derived from the calibration samples. Because the residuals for the calibration samples had a mean of zero, I added 1 to all residuals to make these adjusted SNCBI values (ASNCBI) directly comparable with $O/E$ values.

The effect of precision on inferences regarding biological impairment was assessed by determining how many test-site samples fell outside the distribution of reference-site indicator values. For these tests, I used the lower 10th percentile of reference sample indicator values as a standard threshold below which values would be considered biologically degraded. The 10th-percentile threshold value was used solely to standardize comparisons among methods and data sets and should not necessarily be considered a standard for regulatory purposes. Although use of the 10th percentile here might represent an arbitrary choice for regulatory purposes, it should represent a reasonable threshold for statistical comparisons among methods in that indicator values less than this threshold have only a 10% probability of occurring by chance. Hence values this low should usually represent a biologically real response to stress. Use of a more stringent threshold such as the 1st percentile, although leading to greater confidence that a sample is degraded, could confound comparisons of detection frequencies among indicators because such small percentile values can be easily influenced by outliers in the different distributions of reference-site values. Because the percentage of sites declared as degraded by different methods will not necessarily change in parallel with differences in the percentile threshold used, the comparisons based on the 10th percentile cannot be simply extrapolated to other thresholds.

Unrecognized or uncontrolled variation associated with naturally occurring factors can affect the accuracy of assessments in addition to inflating estimates of error above that associated with sampling error. An important area of current bioassessment research focuses on how to best classify reference sites so as to minimize such errors. I examined accuracy of assessments by determining the extent to which indicator values varied with naturally occurring environmental gradients. This analysis can show if methods are locally biased even though they may be globally accurate, i.e., accurate on average across all sites that are assessed. For this analysis, I regressed indicator values derived from calibration samples against the suite of available variables describing the natural setting for each sample location. Those variables typically included measures of stream size, geographic location, elevation, climate, calendar day, and channel habitat condition. I also conducted complementary analyses based on ANOVA

TABLE 2. Regression statistics derived from calibration data sets describing bias in different assessment measures associated with site-specific differences in environmental setting.

| Data set | Assessment measure† | $R^2$ | Factor‡ | Coefficient | Std. coefficient§ | Tolerance‖ | $t$ | $P$¶ |
|---|---|---|---|---|---|---|---|---|
| North Carolina | SNCBI | 0.71 | Constant | −5.332 | 0.000 | | −9.888 | 0.000 |
| | | | longitude | −0.068 | −0.713 | 0.875 | −17.910 | 0.000 |
| | | | log DFS | −0.096 | −0.268 | 0.876 | −6.727 | 0.000 |
| | | | latitude | 0.027 | 0.089 | 0.988 | 2.371 | 0.019 |
| | | | date | −0.0002 | −0.090 | 0.991 | −2.392 | 0.018 |
| | $O/E_{sp0}$ | 0.05 | constant | 0.933 | 0.000 | | 23.205 | 0.000 |
| | | | log DFS | 0.083 | 0.226 | 0.995 | 3.326 | 0.001 |
| | $O/E_{sp0.5}$ | 0.01 | constant | 0.958 | 0.000 | | 35.186 | 0.000 |
| | | | log DFS | 0.042 | 0.139 | 1.000 | 2.011 | 0.046 |
| | $O/E_{f0}$ | 0.02 | constant | 0.951 | 0.000 | | 32.349 | 0.000 |
| | | | log DFS | 0.040 | 0.140 | 1.000 | 2.030 | 0.044 |
| Mid-Atlantic Highlands | SMIBI | 0.08 | constant | 1.200 | 0.000 | | 10.776 | 0.000 |
| | | | log carbonate | −0.118 | −0.286 | 0.947 | −2.451 | 0.017 |
| | | | log WSA | 0.063 | 0.248 | 0.947 | 2.127 | 0.037 |
| Ohio: invertebrates | SICI | 0.07 | constant | 0.685 | 0.000 | | 4.860 | 0.000 |
| | | | habitat index | 0.0043 | 0.296 | 1.000 | 2.243 | 0.029 |
| Ohio: fish | SIBI | 0.46 | constant | −0.310 | 0.000 | | −1.454 | 0.149 |
| | | | annual precipitation | 0.00091 | 0.289 | 0.853 | 3.873 | 0.000 |
| | | | habitat index | 0.0060 | 0.524 | 0.853 | 7.027 | 0.000 |
| | $O/E_0$ | 0.10 | constant | 0.67 | 0.000 | | 7.531 | 0.000 |
| | | | habitat index | 0.0045 | 0.327 | 1.000 | 3.660 | 0.000 |
| | $O/E_{0.5}$ | 0.11 | constant | 0.72 | 0.000 | | 8.899 | 0.000 |
| | | | habitat index | 0.0043 | 0.341 | 1.000 | 3.844 | 0.000 |

† Assessment measure: SNCBI, standardized North Carolina biotic index; $O/E$, observed taxonomic composition as fraction of expected taxonomic composition (sp = species, f = family; 0 and 0.5 = mean probability of capturing a taxon); SMIBI, standardized macroinvertebrate index of biotic interity; SICI, standardized (Ohio) invertebrate-community index; SIBI, standardized (Ohio) index of biotic integrity.

‡ DFS = distance to stream source, WSA = watershed area, date = day of year (i.e., 1–365).

§ Standardized (Std.) coefficients measure the relative strength of associations between indicator values and the different independent variables.

‖ Tolerance is a measure of independence between predictor variables. High values (near 1) imply little colinearity with other predictors.

¶ Two-tailed.

to test for effects of overall regional setting as indicated by the ecoregion from which samples were collected. I also compared the standardized indicator values derived from different methods with one another to determine to what extent one method was a biased estimator of the other. This latter analysis cannot evaluate accuracy per se, but can provide insight regarding the degree to which two methods lead to similar inferences.

Different indicators may be differentially sensitive or responsive to stressors in general or to individual stressors. I measured *sensitivity* in two ways: (1) as the magnitude of difference between the mean standardized indicator values for the populations of reference and test sites examined, and (2) as the magnitude of standardized regression coefficients derived from regressions of indicator values on different measures of stress. Data on stressors were available only for the Mid-Atlantic Highlands and Ohio data sets. Most stressor data were measured as concentrations of specific chemical constituents (e.g., pH, sulfate, nitrogen), but habitat condition was reported as aggregate indices of habitat-quality measures. EPA measured habitat condition in terms of the "Index of in-stream habitat," which includes aspects of channel sinuosity, amount of various types of substrates, water depth, and velocity characteristics (Kaufman et al. 1999). Ohio used the "qualitative habitat evaluation index" (QHEI) to measure habitat condition, an index that is based on similar metrics as used by EPA: substrate, in-stream cover, channel morphology, riparian and bank cover, and stream gradient (Rankin 1989).

For my analyses, I first used multiple regression to test the hypothesis that indicators were sensitive to all measured stressors. Following that test, I conducted another regression analysis on just those variables that were statistically significant to determine how much of the observed variability in indicator values was associated with measured stressors.

## RESULTS

### *Comparisons of $O/E$ with other indicators*

*North Carolina stream invertebrates.*—The standardized North Carolina biotic index (SNCBI) was both slightly less precise (reference-sample SD) and less sensitive in detecting departure from reference conditions than was the observed taxonomic composition as a fraction of the expected taxonomic composition, $O/E$, based on species-level data, the level of resolution used in the NCBI (Table 1). $O/E$ based on the probability of capturing a taxon, PC, at PC $> 0.5$ ($O/E_{sp,0.5}$) was slightly more precise than $O/E$ based on

TABLE 3. Associations ($r^2$) between indicator values and ecoregion setting, together with mean indicator values for each ecoregion.

| Data source, ecoregion | $N$ | $O/E_{05}$ | NCBI | SNCBI | ASNCBI | MIBI | SMIBI | ICI | SICI | IBI | SIBI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| North Carolina† | | | | | | | | | | | |
| Blue Ridge Mountains | 141 | 1.02 | 3.35 | 1.09 | 1.01 | ... | ... | ... | ... | ... | ... |
| Piedmont | 38 | 0.99 | 4.83 | 0.84 | 0.96 | ... | ... | ... | ... | ... | ... |
| MACP/SP‡ | 29 | 1.00 | 5.09 | 0.91 | 1.02 | ... | ... | ... | ... | ... | ... |
| $r^2$ | | 0.01$^{NS}$ | *0.60* | *0.60* | *0.06* | | | | | | |
| Mid-Atlantic Highlands | | | | | | | | | | | |
| North-central Appalachians | 18 | 1.02 | ... | ... | ... | 77.7 | 1.01 | ... | ... | ... | ... |
| Blue Ridge Mountains | 6 | 1.05 | ... | ... | ... | 78.2 | 1.01 | ... | ... | ... | ... |
| Central App. Ridge and Valleys | 39 | 1.02 | ... | ... | ... | 79.4 | 1.03 | ... | ... | ... | ... |
| Central App. Mountains | 9 | 0.93 | ... | ... | ... | 66.9 | 0.87 | ... | ... | ... | ... |
| $r^2$ | | 0.03$^{NS}$ | | | | 0.08$^{NS}$ | 0.08$^{NS}$ | | | | |
| Ohio: invertebrates | | | | | | | | | | | |
| Eastern Corn Belt Plains | 26 | 1.04 | ... | ... | ... | ... | ... | 43.9 | 1.03 | ... | ... |
| Erie Drift Plain | 13 | 1.07 | ... | ... | ... | ... | ... | 44.2 | 1.03 | ... | ... |
| Huron/Erie Lake Plain | 3 | 1.04 | ... | ... | ... | ... | ... | 41.3 | 0.97 | ... | ... |
| Interior Plateau | 3 | 0.92 | ... | ... | ... | ... | ... | 39.0 | 0.91 | ... | ... |
| Western Allegheny Plateau | 13 | 1.03 | ... | ... | ... | ... | ... | 40.5 | 0.95 | ... | ... |
| $r^2$ | | 0.04$^{NS}$ | | | | | | 0.04$^{NS}$ | 0.04$^{NS}$ | | |
| Ohio: fish | | | | | | | | | | | |
| Eastern Corn Belt Plains | 55 | 1.03 | ... | ... | ... | ... | ... | ... | ... | 46.6 | 1.04 |
| Erie Drift Plain | 16 | 0.97 | ... | ... | ... | ... | ... | ... | ... | 42.5 | 0.95 |
| Huron/Erie Lake Plain | 11 | 0.97 | ... | ... | ... | ... | ... | ... | ... | 33.1 | 0.74 |
| Interior Plateau | 6 | 1.02 | ... | ... | ... | ... | ... | ... | ... | 45.9 | 1.02 |
| Western Allegheny Plateau | 26 | 1.05 | ... | ... | ... | ... | ... | ... | ... | 47.3 | 1.05 |
| $r^2$ | | 0.03$^{NS}$ | | | | | | | | *0.30* | *0.30* |

*Notes:* All $r^2$ values are statistically significant ($P < 0.05$) unless noted as nonsignificant (NS). For assessment-measure codes, see Table 2.

† For North Carolina, $O/E_{05}$ is $O/E_{sp,05}$.

‡ Middle Atlantic Coastal Plain and Southeastern Plateau.

PC > 0 ($O/E_{sp,0}$). Both the lower precision and sensitivity of the SNCBI were associated with the strong dependency (71% of variation) of SNCBI values on naturally occurring environmental conditions (Table 2). Ecoregion accounted for less (60%) variation in SNCBI values than the regression did (Table 3). The adjustment of SNCBI values(ASNCBI) for variation in longitude, distance from source, latitude, and calendar day, resulted in ASNCBI being more precise than that of $O/E_{sp,0.5}$, and, consequently, the number of test sites that were detected as being different from reference increased (Tables 1 and 4). However, the average difference between test and reference sites decreased markedly after adjusting for natural setting (Table 1), a consequence of removing apparent differences between observed and expected values that were caused by systematic variation among test sites in natural features. Adjusting SNCBI values for longitude, distance from source, latitude, and calendar day also resulted in the removal of most, although not all, of the association of SNCBI values with ecoregion (Table 3).

In contrast to the SNCBI, none of the $O/E$ models exhibited substantial site-specific bias with respect to geographic location (latitude, longitude), stream size (log DFS [distance from stream source]), and calendar day, although most models slightly underpredicted

TABLE 4. Concurrence between $O/E_{0.5}$ assessments and the four other biotic indices (ASNCBI, SMIBI, SICI, and SIBI [fish]) in inferring if test sites are in reference or nonreference condition.

| Case | Percentage of samples† | | | |
|---|---|---|---|---|
| | ASNCBI | SMIBI | SICI | SIBI |
| Both $O/E$ and assessment method concur in reference condition | 9 | 30 | 46 | 45 |
| Both $O/E$ and assessment method concur in not reference condition | 67 | 46 | 27 | 29 |
| Only $O/E$ implies reference condition | 13 | 4 | 8 | 11 |
| Only $O/E$ implies not reference condition | 11 | 20 | 19 | 15 |

*Notes:* Key to abbreviations: ASNCBI, adjusted standardized North Carolina biotic index; SMIBI, standardized macroinvertebrate index of biological integrity; SICI, standardized Ohio invertebrate-community index; SIBI, standardized Ohio index of biotic integrity. $O/E$ for the ASNCBI and SIBI are based on species-level data; other $O/E$ models are based on variable taxonomic resolution.

† Values are the percentages of samples that were in each of four possible categories of agreement.

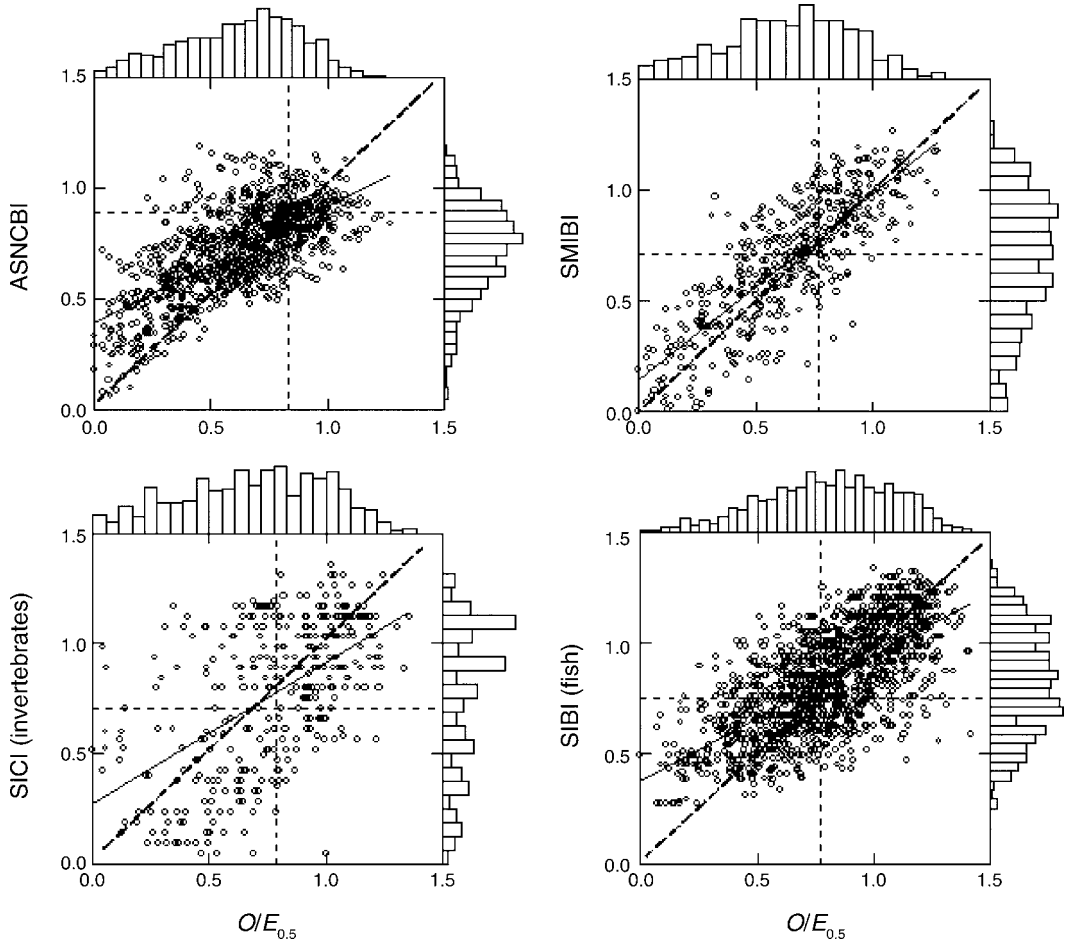


FIG. 1. Relationships between the adjusted standardized NCBI (ASNCBI), the standardized MIBI (SMIBI), the standardized ICI (SICI), and the standardized IBI (SIBI) with $O/E_{0.5}$ values derived from test site samples from each data set. Models for the North Carolina invertebrates and Ohio fish are based on species or near-species taxonomic resolution. The other models are based on highest possible resolution, generally genus. Differences between the solid regression lines and the dashed 1:1 lines show the extent to which the other indicators and $O/E$ are biased predictors of one another. The histograms show the distribution of sample values as measured by each method. Vertical and horizontal dashed lines indicate the lower 10th percentile values derived from the reference calibration sample values for each indicator and are equivalent to the 10th% C values in Table 1, which were used to estimate the percentage of samples in nonreference condition (%T $<$ 10th% C).

richness with increasing stream size (Table 2). None of the variation in $O/E$ values was associated with ecoregion setting (Table 3) showing that the models accounted for the effects of those stream-habitat factors that vary with ecoregion in North Carolina.

Although both $O/E$ and the ASNCBI led to generally similar inferences regarding the percentage of sites that were not in reference condition (Table 1), $O/E$ and the ASNCBI often resulted in markedly different site-specific assessments (Fig. 1). These differences occurred because the ASNCBI was a biased predictor of $O/E$ as revealed by the $>0$ intercept and $<1$ slope of this relationship ($P < 0.05$). This relationship showed that the relative degree of impairment estimated by these two indicators was a function of the overall degree of impairment at a site. In particular, the ASNCBI implied that biological condition was better than that implied by $O/E$ at the most degraded sites, and this difference declined as sites approached reference condition. The overall outcome of these differences was that use of the two assessment methods led to dissimilar inferences regarding the biological status (reference or not) of a test-site sample in 24% of the cases examined based on use of a 10th–percentile-of-reference-values threshold (Table 4).

*Mid-Atlantic Highlands stream invertebrates.*—The standardized macroinvertebrate index of biological integrity for the Mid-Atlantic Highlands (MAH), SMIBI, was marginally less precise than $O/E_{0.5}$ assessments, and $O/E_0$ was less precise than either $O/E_{0.5}$ or the SMIBI (Table 1). In all cases, precision was less than observed for the North Carolina data set. Regressions of reference-sample indicator values on naturally occurring factors showed that the MIBI produced slightly ($R^2 = 0.08$) biased assessments depending on setting (Table 2). Reference-sample SMIBI values decreased with increas-

TABLE 5. Regression statistics describing the response of $O/E$ and standardized multimetric indices to variation among test sites in potential stressors measured at sites in the Mid-Atlantic Highlands (MAH) and Ohio, USA.

| Data set† | Assessment measure‡ | $R^2$ | Factor | Coefficient | Standardized coefficient | Tolerance§ | $t$ | $P$\|\| |
|---|---|---|---|---|---|---|---|---|
| MAH | SMIBI | 0.35 | Constant | 0.763 | 0.000 | | 9.667 | 0.000 |
| | | | $SO_4$ | −0.220 | −0.361 | 0.987 | −10.303 | 0.000 |
| | | | TSS¶ | −0.084 | −0.145 | 0.917 | −4.001 | 0.000 |
| | | | habitat index | 0.039 | 0.376 | 0.918 | 10.350 | 0.000 |
| | $O/E_0$ | 0.25 | constant | 0.449 | 0.000 | | 3.465 | 0.001 |
| | | | pH | 0.081 | 0.245 | 0.949 | 6.387 | 0.000 |
| | | | $SO_4$ | −0.219 | −0.387 | 0.989 | −10.308 | 0.000 |
| | | | habitat index | 0.021 | 0.226 | 0.942 | 5.861 | 0.020 |
| | $O/E_{0.5}$ | 0.35 | constant | 0.061 | 0.000 | | 0.471 | 0.000 |
| | | | pH | 0.085 | 0.238 | 0.949 | 6.673 | 0.000 |
| | | | $SO_4$ | −0.234 | −0.384 | 0.989 | −10.973 | 0.000 |
| | | | habitat index | 0.041 | 0.401 | 0.942 | 11.177 | 0.000 |
| Ohio invertebrates | SICI | 0.19 | constant | 0.159 | 0.000 | | 0.067 | 0.067 |
| | | | habitat index | 0.0093 | 0.390 | 0.996 | 7.100 | 0.000 |
| | | | log $NH_3$ | −0.475 | −0.191 | 0.996 | −3.472 | 0.001 |
| | $O/E_0$ | 0.06 | constant | 0.884 | 0.000 | | 43.911 | 0.000 |
| | | | Pb | −0.027 | −0.257 | 1.000 | −4.306 | 0.000 |
| | $O/E_{0.5}$ | 0.14 | constant | 0.509 | 0.000 | | 6.833 | 0.000 |
| | | | Pb | −0.029 | −0.296 | 0.998 | 3.456 | 0.000 |
| | | | habitat index | 0.0045 | 0.227 | 0.998 | −5.158 | 0.000 |
| Ohio fish | SIBI | 0.25 | constant | 0.341 | 0.000 | | 9.700 | 0.000 |
| | | | habitat index | 0.0069 | 0.444 | 0.990 | 16.284 | 0.000 |
| | | | $NH_3$ | −0.064 | −0.128 | 0.988 | −4.685 | 0.000 |
| | | | Zn | −0.0011 | −0.088 | 0.889 | −3.047 | 0.002 |
| | | | Pb | −0.0100 | −0.091 | 0.882 | −3.145 | 0.002 |
| | | | hardness | 0.00017 | 0.069 | 0.986 | 2.538 | 0.011 |
| | $O/E_0$ | 0.09 | constant | 0.542 | 0.000 | | 14.682 | 0.000 |
| | | | habitat index | 0.0044 | 0.244 | 0.977 | 8.138 | 0.000 |
| | | | $NH_3$ | −0.0485 | −0.081 | 0.996 | −2.724 | 0.007 |
| | | | Zn | −0.00064 | −0.071 | 0.911 | −2.295 | 0.022 |
| | | | Pb | −0.0081 | −0.065 | 0.913 | −2.082 | 0.038 |
| | | | Cd | −0.061 | −0.067 | 0.983 | −2.232 | 0.026 |
| | $O/E_{0.5}$ | 0.15 | constant | 0.390 | 0.000 | | 8.590 | 0.000 |
| | | | habitat index | 0.0059 | 0.317 | 0.990 | 10.858 | 0.000 |
| | | | $NH_3$ | −0.049 | −0.081 | 0.988 | −2.788 | 0.005 |
| | | | Zn | −0.0019 | −0.126 | 0.889 | −4.098 | 0.000 |
| | | | Pb | −0.0100 | −0.075 | 0.882 | −2.431 | 0.015 |
| | | | hardness | 0.00018 | 0.061 | 0.986 | 2.081 | 0.038 |

*Notes:* No stressor data were available for North Carolina or Maine. Habitat condition was measured with indices in which increasing values imply better quality habitat.

† Number of samples for which stressor values were available: MAH, 456; Ohio invertebrates, 264; Ohio fish, 1013.

‡ Key to abbreviations: SMIBI, standardized macroinvertebrate index of biological integrity; $O/E$, observed taxonomic composition as a fraction of expected taxonomic composition (subscripts 0 and 0.5 denote mean probability of capturing a taxon); SICI, standardized Ohio invertebrate-community index; SIBI, standardized Ohio index of biotic integrity.

§ A measure of independence between predictor variables; high values near 1 imply little collinearity with other predictors.

|| Two-tailed.

¶ Total suspended solids.

ing concentrations of carbonate in stream water and increased with watershed area. None of the variation in reference-site SMIBI values was significantly associated with ecoregion setting (Table 3). $O/E_0$ and $O/E_{0.5}$ values were not related to any naturally occurring individual factor that I was able to examine, nor were these measures associated with ecoregion (Table 3).

Both indicators showed that test sites were substantially degraded relative to reference conditions (Table 1). $O/E_{0.5}$ and the SMIBI had nearly identical mean values for test sites. Although mean test-site values were similar, the slightly greater precision of the $O/E_{0.5}$ model resulted in ~30% more test sites being inferred as in nonreference condition than the MIBI. In contrast, the lower precision of the $O/E_0$ model resulted in it assessing ~35% fewer sites than the MIBI as being in nonreference condition. As observed in the comparison between the ASNCBI and $O/E$, the intercept was >0 and the slope <1 ($P < 0.05$), which caused degraded sites to appear more degraded by $O/E$ assessments than by SMIBI assessments. Although the association between $O/E_{0.5}$ and SMIBI values for test sites was stronger than for any other comparison (Fig. 1), the combination of differences in precision and bias resulted in the MIBI and $O/E_{0.5}$ leading to different conclusions regarding impairment in 24% of samples (Table 4). These disagreements were not symmetric with respect to the percentage of samples inferred to be in reference or nonreference condition. Of the 109 samples for which the two assessments disagreed, $O/E_{0.5}$ was five

times more likely than the MIBI to imply a sample was degraded.

Both the MIBI and $O/E$ varied in response to stressor gradients, but the two measures differed in their responsiveness to the suite of stressors present in the MAH region (Table 5). Variation in both $O/E$ indicators was most strongly associated with the same three stressors (pH, $SO_4$, and habitat modification), but more of the variation in $O/E_{0.5}$ was associated with these stressors than that for $O/E_0$, primarily because $O/E_{0.5}$ was more strongly associated with the measure of habitat quality than was $O/E_0$. The SMIBI was similar to $O/E_{0.5}$ in how it declined with decreasing measures of habitat quality and increasing values of $SO_4$, but in contrast to $O/E_{0.5}$, it was not sensitive to pH but was sensitive to total suspended solids (TSS). Variation among sites in levels of stressors accounted for similar amounts of variation in the two types of indicators. Classifying sites by their dominant types of stressors provided somewhat different insights regarding the relative sensitivities of the different indicators (acid mine drainage = metals and pH, pH = acid deposition, nutrients = phosphorus and nitrogen, mixed = general habitat degradation plus other stressors). Although values of both SMIBI and $O/E_{0.5}$ varied similarly among stressor categories (Fig. 2: top panel), ANOVA based on the sample-wise differences between $O/E_{0.5}$ and the S-MIBI showed that the two measures were differentially sensitive to these broad categories of stress (Fig. 2: bottom panel, $F = 5.96$, df = 4537, $P < 0.0005$). In general, $O/E_{0.5}$ appeared to be more sensitive to pH and acid mine drainage than was the MIBI, but there was little difference in sensitivity between the two measures at sites dominated by nutrients or mixed stressors.

*Maine stream invertebrates*.—The Maine $O/E$ indicators were among the least precise of the models examined (Table 1). Although the models showed no systematic bias with respect to any natural environmental gradient examined (ecoregion, Maine biophysical region, latitude, basin size, elevation, channel gradient, temperature, calendar day), the models were less precise than the other $O/E$ indicators examined. The large reference-site $O/E$ standard deviation for these models implies that little of the variation in assemblage composition across reference sites was associated with variation in the predictor variables used (elevation, distance from stream source, latitude, number of freeze-free days, and calendar day). This low precision resulted in a small percentage of test-site samples falling below the 10th percentile of reference-sample values even though mean $O/E$ values estimated for test-site samples were not substantially different from that observed in other data sets.

Mean $O/E$ values declined with decreasing water-quality class as generally expected given how classes were defined by the Maine Department of Environmental Protection. Classes AA and A were combined because they both imply excellent biological integrity

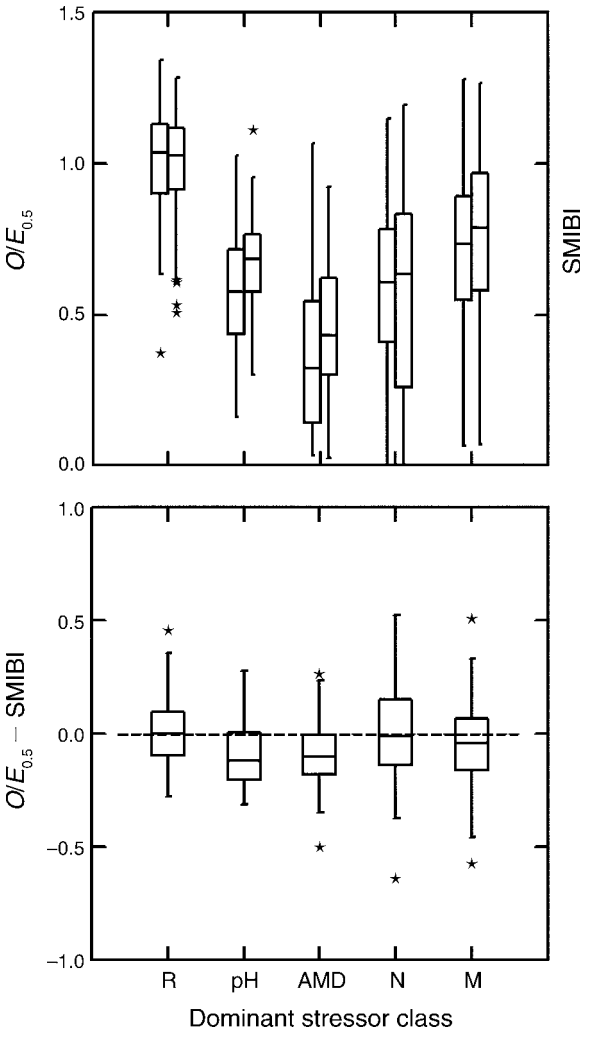


FIG. 2. Box plots of the $O/E_{0.5}$ (left member of pair) and SMIBI (right member of pair) values (top panel) and sample-wise differences between the two indicators showing the differential sensitivity of the two indicators with respect to four classes of dominant stress occurring at each site: acid deposition (pH), acid mine drainage (AMD), nutrients (N), and mixed stressors (M). Reference sites (R) are included for comparison. Box plots show the medians, first and third quartiles (top and bottom of boxes), and lower and upper inner fence values ($\pm 1.5 \times$ inner quartile range). Outliers are shown by stars.

(Davies et al. 1995, Davies and Jackson 2006). However, there was substantial variation in $O/E$ values among samples assigned to any of the water quality classes (Fig. 3). Only 31% ($O/E_0$) and 38% ($O/E_{0.5}$) of the variation in $O/E$ values for test-site samples was associated with the water-quality class to which samples were assigned by the Maine method.

*Ohio stream invertebrates*.—Assessments based on the invertebrate-community index (ICI) and $O/E_{0.5}$ model resulted in similar estimates regarding the average biological condition of test site samples, but the $O/E_0$ model resulted in substantially higher estimates of mean condition for test site samples than either the ICI or

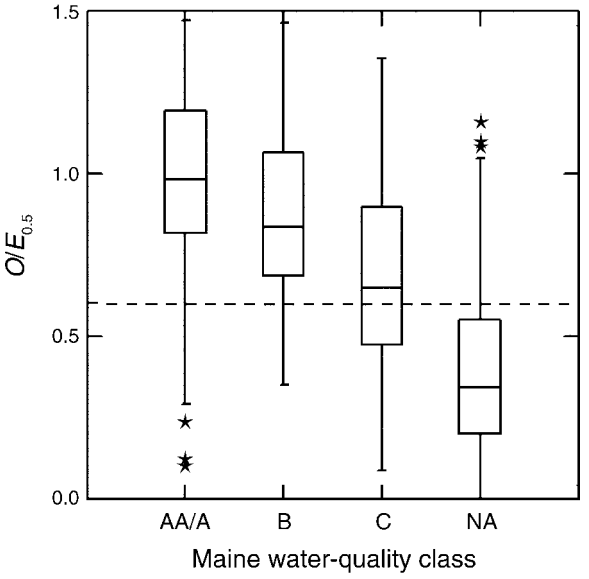


FIG. 3. Variation in test-site $O/E_{0.5}$ values within and among the four different water-quality classes to which samples were assigned by the Maine water-quality class predictive model. Classes AA and A have been combined (since both imply excellent biological integrity). NA represents non-attainment according to Maine water-quality criteria. The horizontal dashed line represents the 10th percentile of reference-site $O/E_{0.5}$ values. Box plots show the medians, first and third quartiles (top and bottom of boxes), and lower and upper inner fence values ($\pm 1.5 \times$ inner quartile range). Outliers are shown by stars.

$O/E_{0.5}$ (Table 1). The $O/E_{0.5}$ model produced the most precise assessments followed by the standardized ICI (SICI) and then the $O/E_0$ model. These differences in precision resulted in corresponding differences in the number of test-site samples that would be inferred as being in nonreference condition (Table 1). The performance of $O/E$ assessments based on Ohio invertebrates was similar to that of the MAH predictive models in terms of bias and precision, but the average condition of test samples was higher and the number of test samples that would be considered to be in non-reference condition was lower for the Ohio data than the MAH data (Tables 1 and 4).

Neither type of assessment method was strongly biased by environmental setting. None of the variation in reference-sample indicator values was associated with ecoregion (Table 3), but ICI values did vary slightly with differences among reference sites in qualitative habitat-evaluation index (QHEI) values (Table 5). Neither of the $O/E$ measures derived from reference-site samples varied with either ecoregion (Table 3) or QHEI (Table 5). ICI and $O/E_{0.5}$ assessment values for individual test site samples were only weakly associated with one another (Fig. 1), and as in the other comparisons, the intercept was $>0$ and the slope $<1$ ($P < 0.05$).

For these data the ICI and $O/E$ values were only weakly associated with estimates of stressors (Table 5). For the 264 samples for which stressor values were available, SICI values were most strongly associated with the index of habitat quality (increased as habitat scores increased) and were less strongly associated with $NH_3$ concentrations (decreased with increasing concentrations). $O/E_0$ values were associated with only Pb concentrations (decreased with increasing concentrations), but $O/E_{0.5}$ values were associated with both habitat quality and Pb. The association of $O/E_{0.5}$ with Pb was stronger than that with habitat. Because Pb and log $NH_3$ concentrations were correlated ($r = 0.51$, $P < 0.001$), both stressor variables may be indicators of the same overall suite of stressors affecting biota at these sites.

*Ohio stream fish.*—The general performance of the index of biological integrity (IBI) and both $O/E$ measures were very similar (Table 1). The precision of the standardized IBI (SIBI) was slightly better than that for $O/E_{0.5}$, which was more precise than $O/E_0$ assessments. The mean condition of test-site samples was also very similar among indicators (Table 1) as was the percentage of test-site samples that were assessed as being in nonreference condition (Table 4). Even though the $O/E_{0.5}$ model was slightly less precise than the SIBI, it assessed a slightly higher percentage of test site samples as being in nonreference condition than the SIBI did because $O/E_{0.5}$ assessed test sites as slightly more degraded on average than did the SIBI. As in other data sets, values of the SIBI and $O/E_{0.5}$ were correlated, and the SIBI was a biased predictor of $O/E$ (intercept $>$ 0, slope $< 1$, $P < 0.05$, Fig. 1). This bias, together with differences in precision, resulted in 26% of test-site samples being assessed differently in terms of whether they were in reference condition or not. $O/E_{0.5}$ had a slightly higher tendency to imply samples were in nonreference condition than the IBI did (Table 4).

Both methods were subject to bias associated with differences between reference sites in environmental setting, the IBI substantially so (Table 2). Thirty percent of the variation in SIBI values was associated with ecoregion (Table 3). Regression analysis showed that even more of the variation among reference sites (46%) in the SIBI was associated with differences among sites in habitat quality and annual precipitation (Table 2). This result implies that an ecoregion classification was only partly successful in accounting for natural variation in fish assemblages among reference, sites and that variation in aspects of climate and channel features affect assemblage structure within ecoregions. In contrast, relatively little of the variation in reference-site $O/E$ values was associated with environmental setting. About 10% of the variation in both $O/E_0$ and $O/E_{0.5}$ was related to variation among reference sites in habitat-quality scores (Table 2), but no variation in either $O/E$ measure was associated with any other channel or regional (e.g., climate, ecoregion) variable (Tables 2 and 3). The $O/E$ models were therefore successful in

accounting for natural variation in biotic structure that occurred both among and within ecoregions.

Both types of indicators were generally similar in their response to stressors (Table 5). The statistical tests of response to seven stressors showed that the IBI and both $O/E$ measures varied with differences among test-site samples in measures of habitat condition, $NH_3$, Zn, and Pb. The IBI and $O/E_{0.5}$ were sensitive to the same five stressors, including hardness, to which $O/E_0$ was not sensitive. $O/E_0$ showed sensitivity to one stressor (Cd) to which the IBI and $O/E_{0.5}$ did not. Variation in stressors accounted for between 9% and 25% of the variation in indicator values. More of the variation in the IBI was associated with stressors than were the $O/E$ measures, although most of the variability in the IBI was associated with habitat condition ($r^2 = 0.20$), a factor that also varied substantially among reference sites. Both $O/E$ measures were also more strongly associated with variation in habitat condition than variation in other potential stressors.

*Summary of comparisons of O/E with other indicators*.—For each comparison, Fig. 1 shows graphically the lower 10th-percentile value for each indicator. Points that fall within the upper right and lower left quadrants as defined by these lines would be assessed similarly as either in reference condition or not in reference condition by the two methods as summarized in Table 4. Points that fall within the other two quadrants would be assessed differently by the two methods, which is also summarized in Table 4. Relationships between the different indicators and $O/E$ were: ASNCBI $= 0.395 + 0.527 \times O/E_{sp,0.5}$, $r^2 = 0.46$; S-MIBI $= 0.145 + 0.839 O/E_{0.5}$, $r^2 = 0.66$; SICI $= 0.226 + 0.686 \times O/E_{0.5}$, $r^2 = 0.34$; SIBI $= 0.368 + 0.569 \times O/E_{0.5}$, $r^2 = 0.46$. In all cases, intercepts and slopes were statistically different ($P < 0.05$) from 0 and 1, respectively.

### *Factors potentially affecting comparability of different O/E assessments*

*Effects of taxonomic resolution on O/E assessments*.—Taxonomic resolution influenced both the precision and sensitivity of $O/E$ indicators (Table 1). In general, model precision improved with decreasing taxonomic resolution (i.e., species to family), and as a consequence the magnitude of biological change that could be detected decreased. Furthermore, models based on PC thresholds of >0.5 were more precise than those based on PC > 0. However, differences in taxonomic resolution affected sensitivity as well as precision, both of which in combination affected assessments. For example, the difference in mean $O/E$ values between test and reference sites increased with increasing taxonomic resolution, and as a consequence so did the percentage of test-site samples with $O/E$ values below the 10th percentile of reference site values even though precision decreased. Exclusion of locally rare taxa (PC > 0.5 models) also resulted in both increasing differences between mean reference- and test-site samples and the number of test sites with $O/E$ values below the 10th percentile of reference-sample values. In general, the species-level model based on PC > 0.5 most strongly discriminated between reference and test sites.

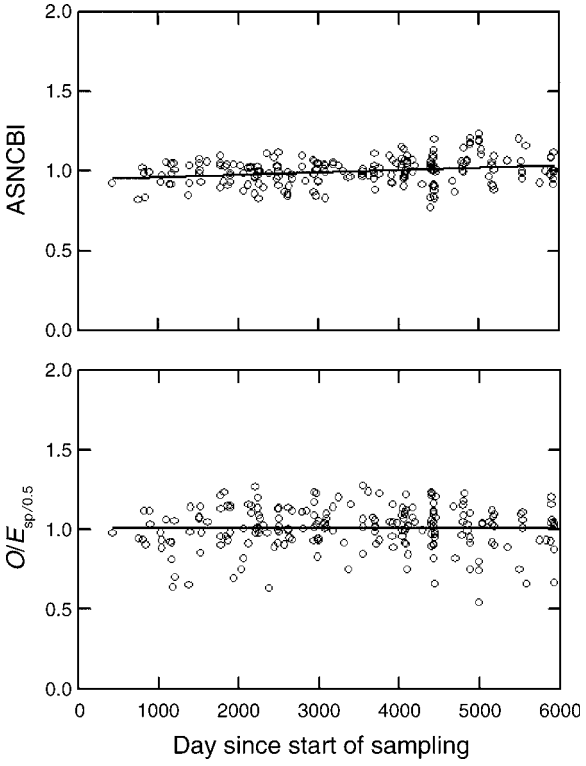


FIG. 4. Variation in adjusted standardized North Carolina biotic index (ASNCBI) and $O/E$ values for the 208 North Carolina reference sites used for model calibration over a 15-yr period of record. No two of these samples were collected from the same site. The linear best-fit line is shown for each indicator.

*Effect of sampling method on O/E assessments*.—The two $O/E$ models that were developed with invertebrate data collected from artificial substrates were distinctly less precise than the $O/E$ models developed from samples collected from natural substrates. The precision of the Ohio model that I developed from invertebrates collected from Hester-Dendy samples was the lowest observed for any model ($O/E_{0.5}$: SD = 0.35, not reported in Table 1). The Maine $O/E$ model based on invertebrates collected from artificial rock baskets was also imprecise (SD = 0.26) relative to the models derived from invertebrate samples collected from natural habitats (Table 1). The precision of both of these models is similar to that observed for null models (Van Sickle et al. 2005), which implies taxonomic composition in these samples varied in a nearly random way between sites.

*Variation in reference-site indicator values with year of sampling*.—In general, indicator values for both the ASNCBI and $O/E$ showed little obvious systematic variation across the ~5500-day period of record in the North Carolina data (Fig. 4) even though this period of time included both droughts as well as regional flooding

associated with hurricanes. None of the variation in $O/E_{sp,0.5}$ was associated with time since sampling started, and only 6% of the variation in ASNCBI values was associated with time. These results imply that, on a regional scale, invertebrate assemblages at reference sites were generally stable from year to year.

## Discussion

In this paper, I compared the performance of $O/E$, the observed taxonomic condition of a site as a fraction of the site's expected (i.e., natural) composition, with the other main types of biological indicators used in the United States and elsewhere. My general aims were to determine (1) how $O/E$ performed relative to other indicators and (2) if $O/E$ could serve as a standard means of quantifying biological condition and thus comparing biological conditions across either politically or ecologically defined regions. This evaluation required that I examine performance measures for both $O/E$ and other indicators as well as factors that potentially affect their performance.

### *Comparability among indicators*

For different biological indicators to be comparable, they must, in general, measure the same properties and respond to stress in parallel. Despite the fact that $O/E$ and the other biological indicators are based on somewhat different biological attributes, on average, $O/E$ assessments were generally similar to assessments based on other indicators. This result might imply that all of these indicators measure to a large extent the same fundamental underlying property of biological assemblages from which different measures of assemblage structure are derived. Because tolerance values, biotic indices, and other types of metrics are derived from the same raw information on composition that $O/E$ measures, these indicators should be correlated with $O/E$.

Differences in indicator performance that were apparent were associated with (1) differences between indicators in the precision with which we estimate expected values (Table 1) and hence the effect size (departure from reference condition) that could be detected; (2) differences in sensitivity among indicators to natural environmental variability among sites (Table 2) that affected bias of assessments; and (3) differences in sensitivity among indicators to different stressors (Table 5, Fig. 2). These factors in combination resulted in both the North Carolina biotic ineex (NCBI) and all three multimetric indices (MMIs) being biased predictors of $O/E$ in such a way that the agreement with $O/E$ for a given sample decreased with increasing biological degradation. This bias in turn resulted in disagreement between indicators in the estimated proportion of sites that we would infer to be in nonreference condition (Table 4).

The tendency for different indicators to frequently (~25% of comparisons, Table 4) disagree in whether an individual sample was in reference condition or not must arise from differences between indicators in either their biological or statistical properties. Both types of effects are likely contributing to imperfect agreement between indicators in their sample-specific assessments. Indicators that combine raw information on taxonomic composition into aggregate metrics, such as stonefly richness or percentage tolerant individuals, could conceivably be either more or less responsive to stressors affecting a site than changes in raw taxonomic composition. Whether such an aggregate metric is more or less sensitive to stress than $O/E$ could easily be dependent on the philosophy used when selecting metrics, i.e., a priori selection based on ecological principles or a posteriori selection based on empirical discrimination between reference and stressed sites. Many of these differences in biological properties among indicators are not transparent to typical users, and to my knowledge we have seldom delved very deeply into how we should interpret these metrics. In the future, we may want to scrutinize how well the indicators we use both measure overall biological integrity and are reflective of the values society actually places on freshwater ecosystems.

The consequences of differences between indicators in their statistical properties are more easily understood. In this analysis, $O/E_{0.5}$ assessments were more likely to detect departures from reference condition than MMI assessments because of their greater precision as well as their slightly greater sensitivity to whatever stressors existed at test sites (Table 1, Figs. 1 and 2). In many cases differences in precision appeared small (e.g., 0.01–0.02 SD units) and therefore potentially not meaningful. Although such differences may not be significant in some specific instances, the general tendency for $O/E_{0.5}$ assessments to consistently have lower SD than other indicators is evidence that such differences are likely real. In general, the precision of $O/E_{0.5}$ models vary from ~0.10 to >0.20 SD, thus differences of 0.01 SD units may therefore represent a 10% difference in precision. Models with SD $< 0.15$ are relatively good, account for a substantial portion of the variation in assemblage structure between sites, and can approach pure sampling error among replicate samples within a site (e.g., Ostermiller and Hawkins 2004, Van Sickle et al. 2005). Models with SD $> 0.20$ are relatively imprecise, are similar in precision to null models, and often account for little of the biotic variation among sites.

In practice, the greater precision gained from modeling may at least partly disappear if expected ranges of indicator values are adjusted by geographic or other strata, as is done for both the NCBI and the Ohio ICI (invertebrate-community index) and IBI (index of biological integrity) (Lenat 1993, Ohio EPA 1989). I demonstrated such an effect when the precision of the NCBI was greatly improved by adjusting for environmental setting by modeling (71% of variation in SNCBI) and to a lessor extent by adjusting for ecoregion (60%). In this case, however, the adjustments had mixed effects

on assessment outcomes. The increase in precision only had a marginal effect on the number of samples that were detected as being in nonreference condition, mainly because adjusting for local environmental setting resulted in estimates of the average condition of reference- and test-site samples becoming more similar to one another (Table 1).

Although apparent at the site level, the effects of biological and statistical differences between indicators appeared to largely disappear when individual site assessments were aggregated across samples, a process that would be applied when conducting regional-scale comparisons of biological condition. Mean standardized IBI values for test-site samples from Ohio were remarkably similar to mean $O/E_{0.5}$ values (within 0.04 units, Table 1). These differences were similar to previously observed differences between $O/E$ assessments derived from models that were based on samples taken either from different habitat types (fixed-area riffle vs. timed multi-habitat) or with different organism counts (50–450) (Ostermiller and Hawkins 2004). Variation in mean $O/E_{0.5}$ values across the Maine water quality classes (Fig. 3) also points to some basic consistency between how $O/E$ and the Maine method assess biological condition. From a statistical perspective, these results imply that regional and national syntheses may be achievable by either harmonizing indicators via standardizing indicator values to common nondimensional units or by reanalysis of raw data to estimate $O/E$ values. The only substantial inconsistency between assessments of mean condition were for the NCBI and $O/E$. The mean of the adjusted, standardized NCBI, which is more comparable to how North Carolina applies the NCBI in practice by adjusting for ecoregion, differed from $O/E_{0.5}$ by 0.10 units for the species model and 0.07 units for the genus model. In this type of situation, it will be more difficult to harmonize indicators by a simple standardization or re-scaling of indicator values.

In general, because all $O/E$-based assessments are designed to measure the same biological property, $O/E$ assessments conducted in different regions should be more comparable to one another than comparisons based on either another type of indicator or on a mix of standardized indicators. Even if we can show empirically that assessments based on different indicators are statistically equivalent, biological inferences could remain problematic. For example, comparing across MMIs will require either that we assume their component metrics are ecologically equivalent or that we develop ways to map different MMIs to a common biological condition scale, i.e., the harmonization approach described by Davies and Jackson (2006). This is an especially problematic issue when comparing assessments across landscapes the size of the entire United States, where assemblage composition and structure, and hence the ecological relevance of any individual metric, will vary markedly across sites and regions. In theory, $O/E$ avoids these problems by basing assessments on the degree to which the observed taxa list matches the expected one. Unfortunately, even comparisons between different $O/E$ assessments are not without problems.

### *Three factors affecting direct comparability between O/E indicators*

$O/E$ assessments are potentially sensitive to at least three of the factors that can also influence other assessment methods: equivalence of reference sites, the taxonomic resolution used, and the sampling method used to collect biota. Stoddard et al. (2006) treat the first issue in detail, but the data examined here also illustrate the problem well. The magnitude of an $O/E$ value is dependent not only on what biota are observed but on the estimate of what biota should occur. If reference sites are of high quality, as many were in the Maine data set, estimates of $E$ may represent something close to the historical potential of a water body. On the other hand, if a region has experienced severe landscape alteration, the least-disturbed sites will likely represent something considerably below historical condition. Such a situation is certainly the case for Ohio and probably North Carolina and the Mid-Atlantic Highlands as well. For example, direct comparison of the mean $O/E_{0.5}$ value for test-site samples from Maine (0.72) with that for Ohio (0.80) implies that biological conditions at Ohio test sites are less impaired than those from Maine. It is unlikely that Ohio streams and rivers are less impaired than those in Maine given the history of landscape alteration in each state. Direct comparison of $O/E$ requires either "equivalent" reference-site quality, something that may be difficult to determine, or societal acceptance that $E$ represents not the historical biological potential of aquatic ecosystems but regionally specific desired or best attainable condition. Although the latter case may greatly complicate comparisons of biodiversity loss or the degree to which systems meet the biointegrity objective of the Clean Water Act (see Stoddard et al. 2006), use of least-impaired reference sites can at least establish a fixed benchmark to which future assessments within a region can be compared. To some extent then, "equivalence" is in the eye of the beholder and dependent on the criteria used to define "expected." Defining expected condition will often have both scientific and social components.

Differences between data sets in taxonomic resolution and sampling method can also confound comparisons between $O/E$ assessments, but can potentially be controlled by standardization of methods (e.g., Ostermiller and Hawkins 2004). The trends observed among the North Carolina models based on different levels of taxonomic resolution (Table 1) are largely consistent with patterns emerging from the literature (Lenat and Resh 2001, Waite et al. 2004). In general, there appears to be a trade-off between precision and sensitivity to stress that varies with the taxonomic resolution used.

Use of coarse taxonomy results in more precise expectations, probably because there are fewer rare taxa that are difficult to model and hence add error to predictions. However, lumping of more highly resolved taxa obscures the signal shown by sensitive taxa within a given group. This trade-off was clear in the North Carolina data, where the mean test site $O/E_{0.5}$ value changed from 0.62 for species to 0.65 for genera and 0.73 for families (Table 1). In contrast, the SD of reference site $O/E$ values changed from 0.14 (species) to 0.13 (genera) to 0.10 (families). In this case, detecting departure from reference condition was more sensitive to responsiveness to stress than precision given that the percentage of sites that were assessed as impaired relative to the 10th reference-sample percentile decreased as taxonomic resolution decreased (Table 1). In practice, decisions regarding the level of taxonomic resolution to use are made by balancing the sensitivity needed against the costs of identifications. These issues are not as problematic for assessments based on fish, for which species-level identifications can often be conducted in the field and are the norm.

Finally, this analysis provided evidence that the performance of $O/E$ models is affected by use of artificial substrates to sample biota (Table 1). The most precise models were those based on the North Carolina data in which all natural habitats at a site were exhaustively sampled (SD for genus-based $O/E_0$ and $O/E_{0.5}$ models = 0.15 and 0.13, respectively). The least precise models were those based on the Maine (SD = 0.26 and 0.30) and Ohio (SD = 0.30) invertebrate data that were collected from artificial substrates and which had been allowed to colonize for 28 to 42 days. Models for the Mid-Atlantic Highlands (MAH) and for Ohio invertebrates were intermediate in precision and were either based on less exhaustive sampling (MAH) or were based on a combination of data collected from artificial and natural substrates (Ohio). These results do not appear to be consistent with analyses that show variance among replicate artificial substrates to be lower than that among samples taken from natural habitats (Rosenberg and Resh 1982, Morin 1985). However they are interpretable in terms of how well (or poorly) the fauna that initially colonize a new, standard habitat patch characterizes the fauna occurring either in the variety of natural habitats that occur within reaches or among reaches that can vary substantially in the types of habitats present. Furthermore, it is not clear how between-sample variance within an individual site would be related to predictive-model precision because the error in these models is based on entire sites as sampling units, not individual subsamples within a site. Although this issue needs to be addressed more rigorously, it seems likely that, at a minimum, detectability of impairment will be affected by sampling technique, and, as a consequence, so will the percentage of test samples that are inferred as impaired (Table 1). It is less clear that sampling method will affect estimates of the average condition of test-site samples (Table 1; also Ostermiller and Hawkins 2004). In one sense, it is unlikely that artificial substrates adequately characterize the natural biotic structure at a site and hence their use would likely be unsuitable for assessments of conservation status or potential. On the other hand, the biota that colonize artificial substrates may provide sufficient signal to detect biological changes relevant to Clean Water Act mandates.

### *A fundamental assumption affecting interpretation of all indicators*

Most bioassessment methods in use today compare the observed biota at a site to that estimated from samples collected at several appropriate reference sites (Bailey et al. 1998, Reynoldson and Wright 2000, Stoddard et al. 2006). Most of the information for these reference sites is usually collected over a short time period, and application of these data to future time periods requires an assumption that the distribution of conditions across reference sites does not change significantly over time (Stoddard et al. 2006). In general, we often assume that the spatial variance in conditions observed across environmentally similar reference sites sampled over a short period of time is equivalent to the temporal (year-to-year) variance we would observe at a single site. Such a space-for-time substitution can greatly reduce the cost of assessments by alleviating the need to constantly monitor individual reference sites and adjust yearly assessments accordingly. However, this assumption has not been well tested.

The stability of the distributions of both $O/E$ and ASNCBI values observed across a 15-year period (Fig. 4) implies that the space-for-time assumption may often be reasonable and that conditions estimated at one time apply to other times. However, other studies have noted significant variation in assemblage structure over time periods as long as 20 years, some of which were associated with climatic conditions (Bradley and Ormerod 2001, Metzeling et al. 2002, Daufresne et al. 2004). Although the results reported here were encouraging, more documentation is clearly needed describing long-term variation in reference-site conditions, especially within different climatic settings, and how well the overall distribution of reference-site indicator values mimics the long-term variation at individual sites.

### *Concluding comments*

In recent years there has been vigorous, and sometimes passionate, debate among researchers and practitioners regarding how to measure the biological integrity of freshwater ecosystems (Gerritsen 1995, Norris 1995, Karr and Chu 1999, 2000, Downes 2000, Norris and Hawkins 2000). The debate has both conceptual and empirical origins (NRC 1994, Boulton 1999, Karr and Chu 2000, Norris and Hawkins 2000). Some view the general concept of biological integrity as heuristically useful but essentially unmeasureable (e.g., NRC 1994).

Some have expressed different views regarding the specific biological attributes that should be measured (Karr and Chu 2000, Norris and Hawkins 2000). Others have questioned the adequacy of different indicators or analytical methods when measuring biological condition (Gerritsen 1995, Norris 1995, Fore et al. 1996). These differences in opinions have usually been fueled by both a scarcity of data and a failure by participants to clearly distinguish between the technical merits of different indicators and the ecological and social values that users attach to those indicators. The analyses presented here show that choice of an indicator, or indicators, for use in regional-scale assessments should depend, in part, on both the biological properties society wishes to measure and the statistical properties of each indicator. There are clearly no perfect indicators that will satisfy all users or uses; however, the numerical simplicity of $O/E$, its ease of biological interpretation, and its inherent standardization to site-specific conditions make it an excellent candidate as a general measure of biological integrity for both local and regional/global assessments.

## Acknowledgments

The research on which this paper was based was supported by a cooperative agreement with the Office of Science and Technology of the United States Environmental Protection Agency (CX-826814-01). I thank Susan Davies and David Courtemanch of the Maine Department of Environmental Protection, Trish McPherson and David Lenat of the North Carolina Department of Environment and Natural Resources, and Chris Yoder and Ed Rankin formerly of the Ohio Environmental Protection Agency for access to their state's databases and for providing comments that improved the manuscript. I also thank EPA EMAP for access to the Mid-Atlantic Highlands data and Alan Herlihy for his help in both defining reference sites and providing stressor data for the Mid-Atlantic Highlands data set. Comments by Mark Vinson, Yong Cao, Daren Carlisle, Lester Yuan, Susan Davies, and two anonymous reviewers substantially improved earlier drafts of the manuscript.

## Literature Cited

Bailey, R. C., M. G. Kennedy, M. Z. Dervish, and R. M. Taylor. 1998. Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual benthic invertebrate communities in Yukon streams. Freshwater Biology **39**:765–774.

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. Second edition. EPA 841-B-99-002. U.S. Environmental Protection Agency, Office of Water, Washington, D.C., USA.

Boulton, A. J. 1999. An overview of river health assessment: philosophies, practice, problems and prognosis. Freshwater Biology **41**:469–479.

Bradley, D. C., and S. J. Ormerod. 2001. Community persistence among upland stream invertebrates tracks the North Atlantic Oscillation. Journal of Animal Ecology **70**: 987–996.

Cao, Y., and C. P. Hawkins. 2005. Simulating biological impairment for evaluating ecological indicators. Journal of Applied Ecology **42**:954–965.

Chutter, F. M. 1972. An empirical biotic index of the quality of water in South African streams and rivers. Water Resources **6**:19–30.

Clarke, R. T., J. F. Wright, and M. T. Furse. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modeling **160**:219–233.

Daufresne, M., M. C. Roger, H. Capra, and N. Lamouroux. 2004. Long-term changes within the invertebrate and fish communities of the Upper Rhône River: effects of climatic factors. Global Change Biology **10**:124–140.

Davies, S. P., and S. K. Jackson. 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecological Applications **16**:1251–1266.

Davies, S. P., and L. Tsomides. 2002. Methods for sampling and analysis of Maine rivers and streams. Maine Department of Environmental Protection, Bureau of Land and Water Quality, Division of Environmental Assessment, Augusta, Maine, USA.

Davies, S. P., L. Tsomides, D. Courtemanch, and F. Drummond. 1995. Maine biological monitoring and biocriteria development program. DEP-LW108. Maine Department of Environmental Protection, Augusta, Maine, USA.

de Zwart, D., S. D. Dyer, L. Posthuma, and C. P. Hawkins. 2006. Predictive models attribute effects on fish assemblages to toxicity and habitat alteration. Ecological Applications **16**: 1295–1310.

Downes, B. 2000. Book review (Restoring life in running waters: better biological monitoring). Freshwater Biology **43**: 663–665.

Fore, L. S., J. R. Karr, and R. W. Wisseman. 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. Journal of the North American Benthological Society **15**:212–231.

Frey, D. G. 1977. Biological integrity of waters: an historical approach. Pages 127–140 *in* R. K. Ballentine and L. J. Guarraia, editors. The integrity of water: a symposium. U.S. Environmental Protection Agency, Washington, D.C., USA.

Gerritsen, J. 1995. Additive biological indices for resource management. Journal of the North American Benthological Society **14**:451–457.

Hawkins, C. P., and D. M. Carlisle. 2001. Use of predictive models for assessing the biological integrity of wetlands and other aquatic habitats. Pages 59–83 *in* R. Rader, D. Batzer, and S. Wissinger, editors. Bioassessment and management of North American freshwater wetlands. John Wiley & Sons, New York, New York, USA.

Hawkins, C. P., R. H. Norris, J. N. Hogue, and J. W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications **10**:1456–1477.

Heinz Center [The H. John Heinz III Center for Science and the Environment]. 2002. The state of the nation's ecosystems: measuring the lands, waters, and living resources of the United States. Cambridge University Press, New York, New York, USA.

Herlihy, A. T., D. P. Larsen, S. G. Paulsen, N. S. Urquhart, and B. J. Rosenbaum. 2000. Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP mid-Atlantic pilot study. Environmental Monitoring and Assessment **63**:95–113.

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

Houston, L., M. T. Barbour, D. Lenat, and D. Penrose. 2002. A multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. Ecological Indicators **1**:279–292.

Johnson, R. K., T. Wiederholm, and D. M. Rosenberg. 1993. Freshwater biomonitoring using individual organisms, populations, and species assemblages of benthic macroinvertebrates. Pages 40–158 *in* D. M. Rosenberg and V. H. Resh, editors. Freshwater biomonitoring and benthic macroinvertebrates. Chapman & Hall, New York, New York, USA.

Karr, J. R. 1981. Assessment of biotic integrity using fish communities. Fisheries 6 (6):21–27.

Karr, J. R., and E. W. Chu. 1999. Restoring life in running waters: better biological monitoring. Island Press, Washington, D.C., USA.

Karr, J. R., and E. W. Chu. 2000. Sustaining living rivers. Hydrobiologia **422–423**:1–14.

Karr, J. R., and D. R. Dudley. 1981. Ecological perspective on water quality goals. Environmental Management **5**:55–68.

Kaufmann, P. R., P. Levine, E. G. Robison, C. Seeliger, and D. D. Peck. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. U.S. Environmental Protection Agency, Washington, D.C., USA.

Klemm, D. J., K. A. Blocksom, W. T. Thoeny, F. A. Fulk, A. T. Herlihy, P. R. Kaufmann, and S. M. Cormier. 2002. Methods development and use of macroinvertebrates as indicators of ecological conditions for streams in the Mid-Atlantic Highlands region. Environmental Monitoring and Assessment **78**:169–212.

Lenat, D. R. 1993. A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assigning water-quality ratings. 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 North American Benthological Society **20**:287–298.

Metzeling, L., D. Robinson, S. Perriss, and R. Marchant. 2002. Temporal persistence of benthic invertebrate communities in south-eastern Australian streams: taxonomic resolution and implications for the use of predictive models. Marine and Freshwater Research **53**:1223–1234.

Morin, A. 1985. Variability of density estimates and the optimization of sampling programs for stream benthos. Canadian Journal of Fisheries and Aquatic Sciences **42**: 1530–1534.

Moss, D., M. T. Furse, J. F. Wright, and P. D. Armitage. 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology **17**:41–52.

National Research Council. 1994. Review of EPA's environmental monitoring and assessment program. Committee to Evaluate Indicators for Monitoring Aquatic and Terrestrial Environments, Board on Environmental Studies and Toxicology, and the Water Science and Technology Board, Commission on Geosciences, Environment and Resources. National Academies Press, Washington, D.C., USA.

National Research Council. 2000. Ecological indicators for the nation. National Academy of Sciences Press, Washington D.C., USA.

Norris, R. H. 1995. Biological monitoring: the dilemma of data analysis. Journal of the North American Benthological Society **14**:440–450.

Norris, R. H., and C. P. Hawkins. 2000. Monitoring river health. Hydrobiologia **435**:5–17.

North Carolina Department of Environment, Health and Natural Resources. 2003. Standard operating procedures: biological monitoring. Division of Water Quality, Raleigh, North Carolina, USA.

Ohio Environmental Protection Agency. 1989. Biological criteria for the protection of aquatic life. Volume III. Standardized biological field sampling and laboratory methods for assessing fish and macroinvertebrate communities. State of Ohio Environmental Protection Agency, Ecological Assessment Section, Division of Water Quality, Columbus, Ohio, USA.

Ostermiller, J. D., and C. P. Hawkins. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. Journal of the North American Benthological Society **23**:363–382.

Rankin, E. T. 1989. . The qualitative habitat evaluation index (QHEI): rationale, methods, and application. State of Ohio Environmental Protection Agency, Ecological Assessment Section, Division of Water Quality, Columbus, Ohio, USA.

Revenga, C., and Y. Kura. 2003. Status and trends of biodiversity of inland water ecosystems. Technical series number 11. Secretariat of the Convention on Biological Diversity, Montreal, Quebec, Canada.

Reynoldson, T. B., and J. F. Wright. 2000. The reference condition: problems and solutions. Pages 293–304 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse, editors. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

Rosenberg, D. M., and V. H. Resh. 1982. The use of artificial substrates in the study of freshwater benthic macroinvertebrates. Pages 175–235 *in* J. Cairns, Jr., editor. Artificial substrates. Ann Arbor Science, Ann Arbor, Michigan, USA.

Simpson, J. C., and R. H. Norris. 2000. Biological assessment of river quality: development of AusRivAS models and outputs. Pages 125–142 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse, editors. Assessing the Biological Quality of Freshwaters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications **16**:1267–1276.

USEPA [U.S. Environmental Protection Agency]. 2002*a*. National water quality inventory—2000 report. EPA-841-R-02-00. U.S. Environmental Protection Agency, Office of Water, Washington, D.C., USA.

USEPA [U.S. Environmental Protection Agency]. 2002*b*. Summary of biological assessment programs and biocriteria development for states, tribes, territories, and interstate commissions: streams and wadeable rivers. EPA-822-R-02-048. U.S. Environmental Protection Agency, Office of Water, Washington, D.C., USA.

USEPA [U.S. Environmental Protection Agency]. 2003. Draft report on the environment.Technical document. EPA-600-R-03-050. United States Environmental Protection Agency, Washington, D.C., USA.

U.S. General Accounting Office. 2000. Water quality: key EPA and state decisions limited by inconsistent and incomplete data. Report to the Chairman, Subcommittee on Water Resources and the Environment, Committee on Transportation and Infrastructure, House of Representatives. GAO/RCED-00-54. U.S. General Accounting Office, Washington, D.C., USA.

Van Sickle, J., C. P. Hawkins, D. P. Larsen, and A. T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. Journal of the North American Benthological Society **24**:178–191.

Waite, I. R., A. T. Herlihy, D. P. Larsen, N. S. Urquhart, and D. M. 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.

Wright, J. F. 2000. An introduction to RIVPACS. Pages 1–24 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse, editors. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.