Environ Monit Assess (2018) 190:446
https://doi.org/10.1007/s10661-018-6817-x

Coarse taxonomy (tolerance-value averaging) biases Hilsenhoff's family-level biotic index

Erin McGauley · Brett Tregunno · F. Chris Jones

Received: 27 September 2017 / Accepted: 18 June 2018

Abstract Hilsenhoff's family-level index (FBI) combines information about the relative abundances of taxa and their tolerances to pollution. Versions of this index are used extensively in North America to assess water quality. When faced with constraints on time, money, or expertise, bioassessment practitioners have been tempted to calculate a version of the FBI with very coarse (e.g., order-level) taxonomy. Such an approach requires a degree of within-taxon averaging of tolerance values and raises questions about the degree to which accuracy is compromised and bias is introduced. Data from thousands of streams in Ontario (Canada) demonstrated that such tolerance-value averaging produces index scores and associated water-quality classifications that are not equivalent to those calculated with the standard family-level taxonomic precision. Two methods were used in an attempt to correct the order-level FBI scores to equivalence with the family-level index: (1) tolerance scores for the orders included in the calculation were calculated as abundance-weighted means of the scores of their component families, and (2) order-level FBI scores were estimated as predicted values from a polynomial regression of the two versions of the index. The use of abundance-weighted mean tolerance scores greatly improved the accuracy of the order-level index, and the regression-based correction reduced bias by equalizing the distribution of errors across the range of observed FBI values. Nonetheless, equivalence of scores was not demonstrated, and water quality was misclassified in 12 to 80% of cases. Practitioners are discouraged from the practice of tolerance-value averaging and are advised to adhere to the standard family-level FBI.

Keywords Biotic index · Taxonomic precision · Equivalence test · Classification success · Ontario · Streams

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10661-018-6817-x) contains supplementary material, which is available to authorized users.

E. McGauley
Otonabee Region Conservation Authority, 250 Milroy Drive, Peterborough K9H 7M9, Canada

B. Tregunno
Kawartha Region Conservation Authority, 277 Kenrei Road, Lindsay K9V 4R1, Canada

F. C. Jones (✉)
Ontario Ministry of Environment and Climate Change, Dorset Environmental Science Centre, 1026 Bellwood Acres Road, Dorset P1H 2J6, Canada
e-mail: f.chris.jones@ontario.ca

Introduction

Biotic indices allow stream condition to be interpreted (Chang et al. 2014) because they summarize the ecological information that is encoded in the occurrence frequencies, relative abundances, and sensitivities of the taxa that comprise a community (Norris and Georges 1993). William Hilsenhoff formulated genus/species-level (Hilsenhoff 1977, 1987) and family-level (Hilsenhoff 1988) versions of a biotic index, and

Published online: 04 July 2018

tabulated interpretive biocriteria based on known sensitivities of arthropod taxa to organic enrichment (i.e., sewage pollution). His indices are scaled from 0 to 10, and quantitative scores can be interpreted as water-quality classifications, ranging from *excellent* to *very poor* (see leftmost three columns in Table 1). Hilsenhoff's family-level biotic index (FBI; Hilsenhoff 1988) has been widely used in North America to characterize the health of freshwater streams (Reynoldson and Metcalfe-Smith 1992)—indeed, it helped to popularize rapid bioassessment, which is characterized by the use of semi-quantitative sampling methods, coarse taxonomic precision, and fixed-count subsamples (e.g., Carter and Resh 2001).

Conservation authorities in Ontario, Canada, have been employing bioassessment methods to monitor stream condition since the mid-1990s. Following discussions with many conservation authority staff, Jones and Wilcox (2003) proposed methods by which FBI scores could be interpreted as water-quality letter grades (Table 1) and synthesized as public summaries of water quality at the watershed scale. The FBI indicator was ultimately adopted by Conservation Ontario (a non-profit organization that represents Ontario's 36 conservation authorities) and has been used to produce "Watershed Report Cards," which have been published by most conservation authorities every fifth year since 2007.

Many authors have explored the trade-off that exists between ecological information content and monitoring program costs, which both increase as taxonomic precision increases (Carter and Resh 2001; King and Richardson 2002; Jones 2008). Faced with constrained resources, citizen scientists (e.g., Stanfield 2003; EcoSpark 2013) and some conservation authorities involved in the Watershed Report Card program have conducted rapid bioassessments using relatively coarse (e.g., order-level) taxonomic precision. Reporting on stream condition using a version of the FBI calculated with order-level taxonomy requires a degree of tolerance-value averaging and raises questions about bias, accuracy, and agreement with family-level scores (Bailey et al. 2001; Bouchard et al. 2005).

In this paper, we use benthic invertebrate community data from thousands of southern Ontario streams to investigate the equivalence of FBI scores calculated with coarse taxonomy (and tolerance-value averaging), and FBI scores calculated as per the standard family-level version of the index (i.e., Hilsenhoff 1988; Conservation Ontario 2011). Specifically, we pose the following questions: Does the use of family-level and coarse taxonomy result in equivalent FBI scores and water-quality classifications? If not, how large are the discrepancies, at what frequencies do they occur, and can index values calculated with the coarsest taxonomic detail be corrected to equivalence with the FBI?

Methods

Survey design

Different samples are generally made up of different types and numbers of benthic invertebrates, across which tolerance scores must be averaged to calculate a coarse taxonomy version of the FBI. For these reasons, we hypothesized that the amount of bias introduced by the use of coarse taxonomy should be different for

Table 1 Categorical classifications of the Hilsenhoff family biotic index and corresponding conservation authority letter grades, as specified by Conservation Ontario (2011)

FBI score	Narrative interpretation	Indicated degree of organic pollution	Conservation authority letter grade
0.00–3.75	Excellent	Organic pollution unlikely	A
3.76–4.25	Very good	Possible slight organic pollution	A
4.26–5.00	Good	Some organic pollution probable	B
5.01–5.75	Fair	Fairly substantial pollution likely	C
5.76–6.50	Fairly poor	Substantial pollution likely	D
6.51–7.25	Poor	Very substantial pollution likely	F
7.26–10.00	Very poor	Severe organic pollution likely	F

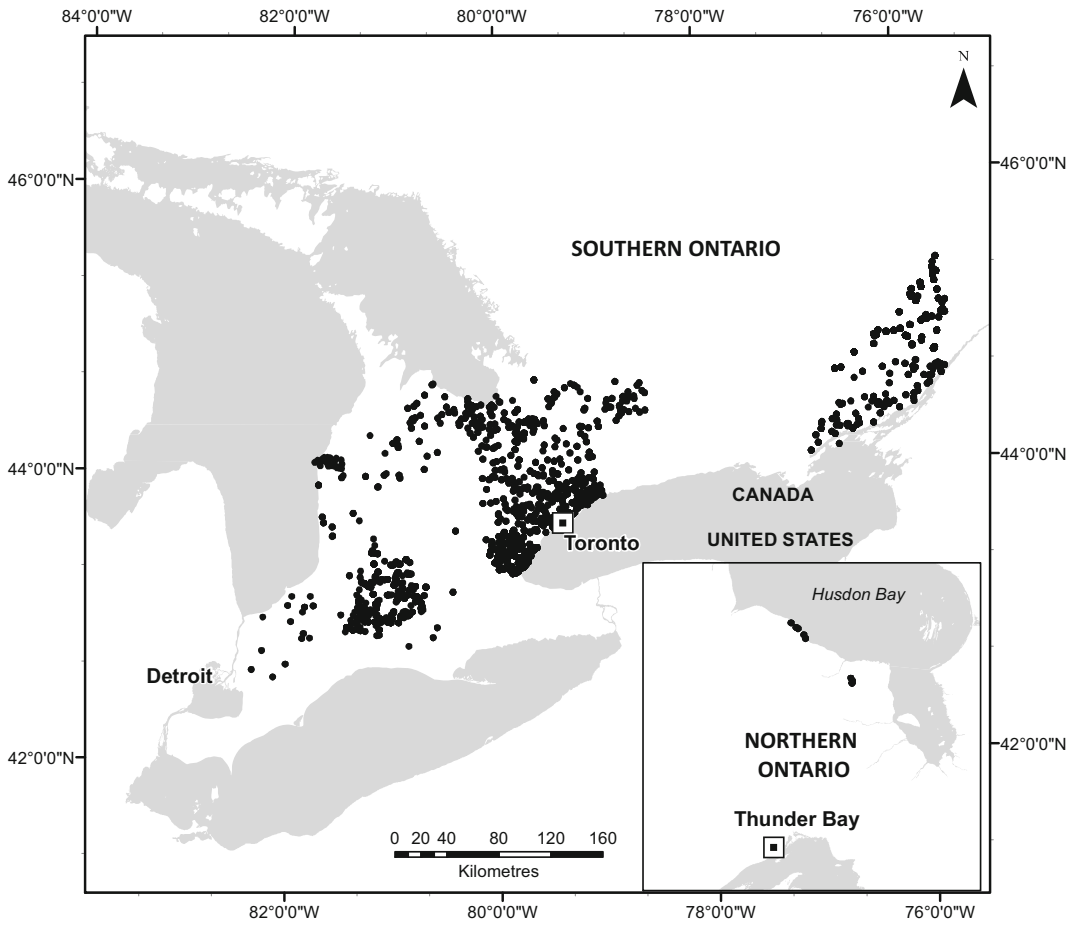

Fig. 1 Sampled locations, regional context

different samples. In order to make generalizations about the consequences of using coarse taxonomy, we compiled a dataset that represented the broad range of stream conditions that exist in Ontario, Canada (Fig. 1). This province contains a variety of ecoregions with different mixtures of natural habitat features and land uses and correspondingly variable benthic macroinvertebrate communities.

Taxa counts from 5047 stream benthic invertebrate samples were obtained from collaborators representing nine conservation authorities and three regional-scale bioassessment projects. These samples were collected in pool and riffle habitats, on one or multiple occasions between 2006 and 2013, using the traveling-kick-and-sweep method of Jones et al. (2007). Samples were collected at different times throughout the year, a majority (two thirds) being collected in either May or October.

Taxa collected in each sample were diagnosed to their "families,"[1] enumerated, and assigned tolerance values (Electronic Supplement 1) so that FBI scores could be calculated as per Conservation Ontario's Guide to Developing Watershed Report Cards (Smith et al. 2009; Conservation Ontario 2011).[2]

[1] Our family-level taxonomy is more precisely referred to as a "mixed-level taxonomic aggregation" (Jones 2008) in which insects, crustaceans, molluscs, and leeches were diagnosed as families, and the Coelenterata, Platyhelminthes, Nemata, Hydrachnidia, and oligochaetous Clitellata were assigned only to these coarse taxonomic ranks.

[2] FBI values were assigned according to the list of taxa-specific sensitivities reported by Smith et al. (2009), meaning that the conservation authorities' formulation of the FBI includes several taxa that were excluded from Hilsenhoff's (1988) index (e.g., Hydrachnidia, Hemiptera, Decapoda, Oligochaeta, Bivalvia, Hirudinea).

Of the 5047 records shared with us by our collaborators, mites were excluded from enumerations in 837 samples. In order to include the corresponding samples in our dataset, we approximated mite abundances in each sample as the global mean relative abundance of Hydrachnidia that was observed among the 4210 samples in the dataset for which these animals were enumerated.

A coarse taxonomy version of the dataset (FBI_{27}) was then assembled by lumping groups in a way that reflected the 27 taxa (a mix of phyla, classes, orders, and families) that constitute the minimum acceptable taxonomic precision for the Ontario Benthos Biomonitoring Network (Jones et al. 2007; Electronic Supplement 1). Tolerance values were assigned to these taxa as the average of the scores associated with their component families.

All samples in the resulting FBI and FBI_{27} datasets (Electronic Supplement 2) satisfied two criteria: (1) after removing taxa for which FBI tolerance values are not assigned, total sample abundance was at least 90 individuals (i.e., sample size was sufficiently large to allow the relative abundances of the community's component taxa to be quantified); and (2) after removing taxa for which FBI tolerances are not assigned, a minimum of 80% of the sample's total abundance was retained in the FBI calculation (i.e., the abundance of taxa that contributed to the FBI calculation was a reasonable approximation of the sample's total abundance).

Statistical analyses

We treated the FBI as a reference standard because it was calculated according to the method specified in Conservation Ontario's Guide to Developing Watershed Report Cards (Conservation Ontario 2011). Against this standard, FBI_{27} values constitute calculated values that are subject to error arising from tolerance-value averaging.[3]

In order to achieve the best possible agreement between these two datasets, we recalculated FBI_{27} scores by applying two different corrections: (1) To improve accuracy, we recalculated tolerance scores for each of the 27 taxa as the weighted average of the scores assigned to their component families (weights reflected the relative abundances of the component taxa in the entire 5047-record dataset; Electronic Supplement 1); (2) to reduce bias by equalizing inaccuracies across the range of FBI values, we then recalculated FBI_{27} as fitted values from the second-order polynomial regression of FBI against FBI_{27}. Polynomial regression was selected because the relationship between FBI and FBI_{27} was best characterized as curvilinear, and addition of a squared predictor in the regression model allowed the sign and magnitude of the correction to vary across the observed range of FBI values. The effect of the regression-based correction was to optimally align $FBI_{27corrected}$ values to the FBI/FBI_{27} 1:1 line of equivalence. Water quality was then classified using two different interpretations: Hilsenhoff's narrative scheme (Hilsenhoff 1988), and the letter-grade system used by Conservation Ontario (Conservation Ontario 2011, Table 1).[4]

We assessed equivalence of FBI and $FBI_{27corrected}$ according to their raw index scores and their associated water-quality classifications. The equivalence of the raw scores was examined using the regression-based equivalence test proposed by Robinson et al. (2005). Test statistics included both the slope (m) and intercept (b) of the $FBI_{27corrected} = m(FBI) + b + error$ regression model. The Robinson et al. (2005) test was useful because it reversed the traditional null hypothesis of no difference by postulating that the two populations being compared were different, and using the data to prove otherwise. The test had two possible outcomes: (1) rejection of the null hypothesis (i.e., $FBI_{27corrected}$ values are significantly equivalent to FBI values) or (2) acceptance of the null hypothesis (i.e., there is insufficient evidence to conclude that FBI and $FBI_{27corrected}$ values are equivalent). Testing equivalence in a regression context, rather than an ANOVA context, further allowed the goodness of fit between *individual* HBI_{fam} and their corresponding $HBI_{27corrected}$ scores to be evaluated directly (whereas ANOVA only allows the equivalence of *means* to be assessed).

The major difficulty with the Robinson et al. (2005) approach is that setting a critical effect size or

[3] We acknowledge that sampling error in FBI estimates also exists, but this source of variance is not pertinent to the questions addressed in the present study.

[4] We recognize that classifying FBI values introduces an assessment bias by potentially splitting similar index values into different interpretive classes or lumping quite different index values into the same interpretive class. Nonetheless, most practitioners interpret the FBI by splitting its range into ordinal classes. Conservation authorities use this approach in their Watershed Report Card process, and we also use it, so that the methods of these water management agencies can be appropriately evaluated.

"equivalence region" is a subjective exercise. The critical effect size can be defined in real units of the predicted variable, and we reasoned that an effect size equal to half the width of the narrowest interpretive water-quality class (i.e., Hilsenhoff's (1988) "Very Good" class, which includes FBI values between 3.76 and 4.25, a width of 0.5 units) would be an acceptable amount of model bias. Given a one-tailed equivalence region, this suggested a critical effect size for the regression intercept of 0.125 FBI units; but it was not clear how to translate this intercept-based criterion into a slope-based criterion. For this reason, we defaulted to a proportional critical effect size (i.e., 5% of the intercept, and 5% of the slope), which matches the 5% significance level that is standard in science. Although FBI and $FBI_{27corrected}$ appeared to be approximately normally distributed (Fig. 2), we sidestepped any potential violation of the probabilistic equivalence test, by calculating confidence limits for regression parameters non-parametrically using bootstrapping.

We evaluated equivalence of FBI and FBI_{27} narrative and letter-grade classifications according to the frequency of misclassifications—misclassifications being defined as cases in which a sample's FBI and $FBI_{27corrected}$ scores were not identically classified—which we report in contingency tables and as overall mean misclassification rates. For both the letter-grade and narrative interpretive schemes, the overall misclassification rate was calculated as the weighted mean of the percentage of incorrect classifications occurring in each interpretive category (each percentage weighted by the number of observed values in its corresponding FBI category). Complementary evidence was provided by Cohen's Kappa (K, a measure of inter-rater agreement [i.e., correlation] that accounts for the probability of chance agreements), and the p value from chi-square tests (estimated by the Monte Carlo process with 1000 simulated replicates), which represented the likelihood that narrative or letter-grade $FBI_{27corrected}$ classifications were drawn from their expected distributions (i.e., from the observed distribution of FBI results).

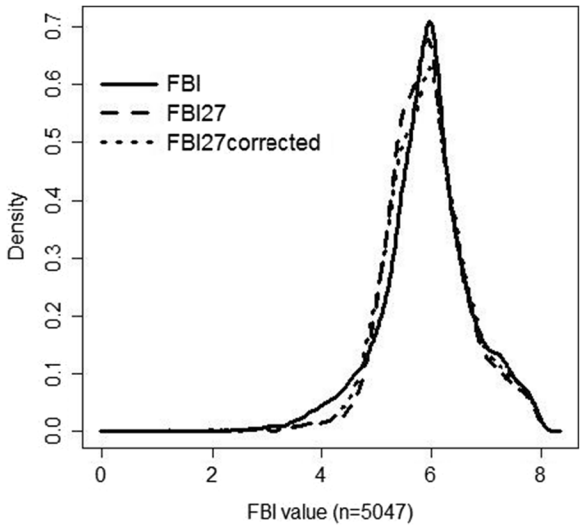

Fig. 2 Probability densities of FBI scores

The FBI and FBI_{27} datasets were compiled in Microsoft Excel. All analyses were completed for the full ($n = 5047$) dataset, with calculations performed using scripts written in the R language and environment for statistical computing (R Core Team 2016; Robinson 2016; Electronic Supplement 3). The α-level for all hypothesis tests and confidence intervals was 0.05.

Results

Simple graphical summaries of FBI and FBI_{27} indicated non-equivalence of these two indices. For this reason, the following results focus solely on comparisons between the FBI and $FBI_{27corrected}$ datasets.

FBI and $FBI_{27corrected}$ values had similar means of approximately 6 (suggesting *fairly poor* water-quality conditions, or a *D* letter grade; Table 1 or Fig. 2). The two versions of the index were strongly correlated ($R^2 = 0.89$, Fig. 3), but FBI was more variable (variance = 0.68, coefficient of variation = 14%) than $FBI_{27corrected}$ (variance = 0.61, coefficient of variation = 13%). Bootstrapped 95% confidence intervals for the regression model's intercept (0.00–0.01) fell within the defined equivalence region; however, those of the slope coefficient (0.80–0.83) did not. Non-equivalence of FBI and $FBI_{27corrected}$ was, therefore, demonstrated. The pattern of agreement was inconsistent across the observed range of FBI values: $FBI_{27corrected}$ values in the range of approximately 3–5 (*good* to *excellent*) were generally higher and made water quality appear worse than what was indicated by their corresponding FBI values. To the contrary, $FBI_{27corrected}$ values between approximately 5 and 7 (*fair* to *poor*) were generally lower and made water quality appear better than what was indicated by their corresponding FBI values (Figs. 2 and 3).

Lack of equivalence of index values resulted in a large number of water-quality misclassifications (Tables 2 and 3). The percentage of $FBI_{27corrected}$

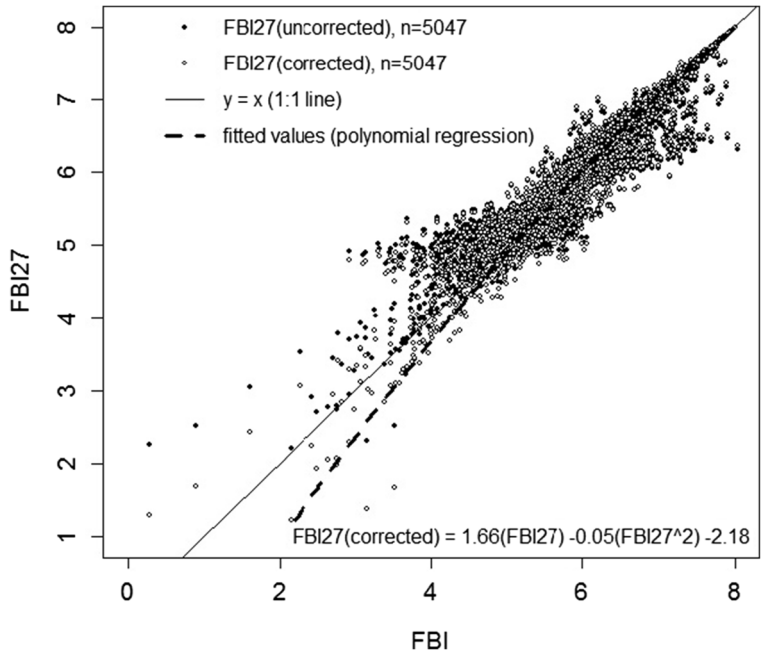

Fig. 3 Goodness of fit of FBI_{27} and $FBI_{27corrected}$ scores, relative to those of FBI. The dashed line represents the best fit second-order polynomial regression of FBI_{27} against FBI (coefficients from this equation were used to calculate $FBI_{27corrected}$). The solid line represents the 1:1 line of equivalence (scatter about this line demonstrates non-equivalence, and uneven distribution of residuals about this line demonstrates bias)

narrative results that were correctly classified ranged from 20% (*very good*) to 86% (*fair*), and the overall mean misclassification rate was 21% (Table 2). Designations of *poor* water quality arose with approximately equal frequency in the FBI and $FBI_{27corrected}$ datasets, whereas *excellent*, *very good*, *good*, *fairly poor*, and *very poor* designations occurred less frequently, and *fair* designations occurred more frequently in the $FBI_{27corrected}$ dataset ($K = 0.90$; chi-square p value < 0.001; Fig. 4).

Similarly, when FBI scores were expressed as Conservation Ontario's letter grades, the frequency of correct assignments ranged from 44% (*A*) to 87% (*F*), and the overall mean misclassification rate was 20% (Table 3). *A*, *B*, *D*, and *F* letter grades arose less frequently, and *C* grades arose more frequently in the $FBI_{27corrected}$ dataset than in the FBI dataset ($K = 0.89$; chi-square p value < 0.001; Fig. 5).

FBI narrative water-quality categories were distributed among 3–4 of the corresponding $FBI_{27corrected}$ categories, meaning that misclassification errors were severe in some cases. For example, *very good* FBI results were classified as *excellent* in 11% of cases, as *good* in 60% of cases, and as *fair* in 10% of cases (Table 2). Similarly,

Table 2 Distribution of FBI narrative results among corresponding $FBI_{27corrected}$ categories. The percentages of family-level results that were correctly classified with 27-group taxonomy are shown on the diagonal in italics

		$FBI_{27corrected}$						
		Excellent ($n = 50$) (%)	Very good ($n = 41$) (%)	Good ($n = 367$) (%)	Fair ($n = 1620$) (%)	Fairly poor ($n = 1985$) (%)	Poor ($n = 703$) (%)	Very poor ($n = 281$) (%)
FBI	Excellent ($n = 67$)	*57*	9	30	4	0	0	0
	Very good ($n = 114$)	11	*20*	60	10	0	0	0
	Good ($n = 395$)	0	3	*53*	44	0	0	0
	Fair ($n = 1348$)	0	0	5	*86*	9	0	0
	Fairly poor ($n = 2119$)	0	0	0	13	*82*	5	0
	Poor ($n = 694$)	0	0	0	0	16	*80*	4
	Very poor ($n = 310$)	0	0	0	0	5	12	*82*

Table 3 Distribution of FBI letter-grade results among corresponding $FBI_{27corrected}$ categories. The percentages of family-level results that were correctly classified with 27-group taxonomy are shown on the diagonal in italics

		$FBI_{27corrected}$				
		A ($n = 50$) (%)	B ($n = 41$) (%)	C ($n = 367$) (%)	D ($n = 1620$) (%)	F ($n = 1985$) (%)
FBI	A ($n = 181$)	*44*	49	8	0	0
	B ($n = 395$)	3	*53*	44	0	0
	C ($n = 1348$)	0	5	*86*	9	0
	D ($n = 2119$)	0	0	13	*82*	5
	F ($n = 1004$)	0	0	0	13	*87*

FBI letter grades were distributed among 2–3 of the corresponding $FBI_{27corrected}$ classes. For example, FBI *A*-grades were classified as *B*s in 49% of cases, and as *C*s in 8% of cases (Table 3).

The coefficients used to create the $FBI_{27corrected}$ dataset are shown in Eq. 1.

$$FBI_{27corrected} = 1.66(FBI_{27}) - 0.05(FBI_{27}^2) - 2.18 \ (n = 5047, R^2 = 0.89). \quad (1)$$

Applying this polynomial correction did not reduce scatter about the 1:1 line of $FBI/FBI_{27corrected}$ equivalence, but served to equalize error rates across the interpretive classes (Tables 4 and 5, Fig. 3).

Conclusion and discussion

We conclude that FBIs calculated with coarse (i.e., 27-group) taxonomy are not equivalent to the standard family-level index. Calculating tolerance values as abundance-weighted means of component taxa greatly improves accuracy relative to using unweighted means, and regression-based correction reduces bias by equalizing error rates across the range of FBI values. Nonetheless, we caution biomonitoring practitioners that the tolerance-value weighting factors and regression coefficients used to create our $FBI_{27corrected}$ dataset may not be equally appropriate in all watersheds (i.e., the corrective performance of these measures will depend on the richness and relative abundances of the different invertebrate families, which likely differ markedly from the provincial averages in some localities). In general, even when both of these corrections are used, $FBI_{27corrected}$

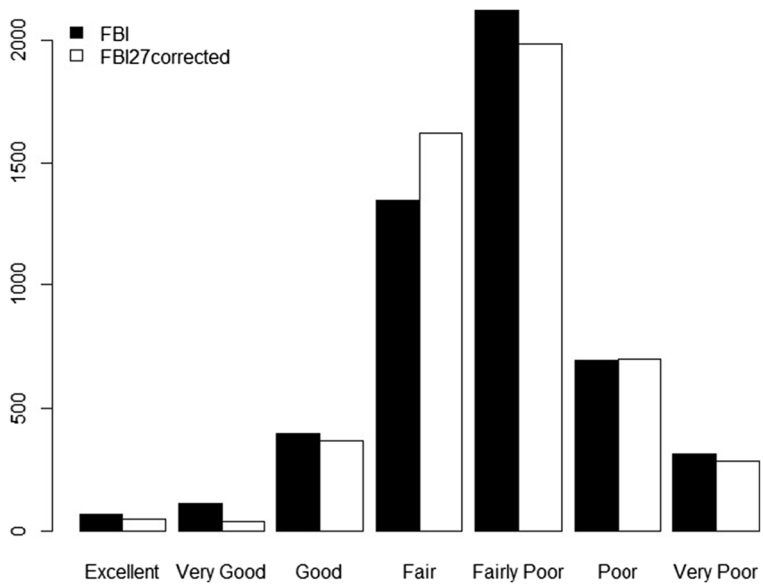

Fig. 4 Occurrence frequencies of FBI and $FBI_{27corrected}$ narrative interpretive results

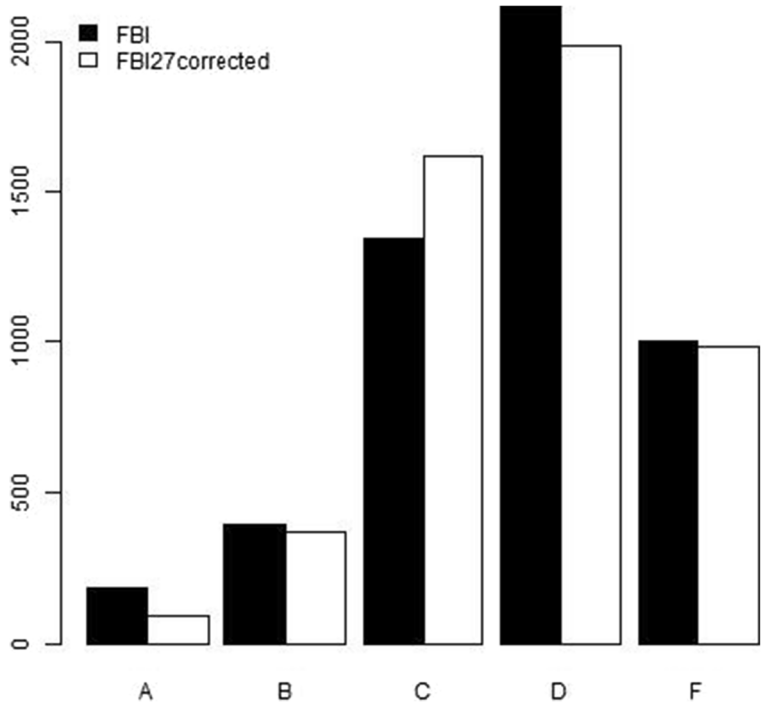

Fig. 5 Occurrence frequencies of FBI and $FBI_{27corrected}$ letter grades, as assigned in the Conservation Ontario (2011) Watershed Report Card process

values do not capture all of the variation in the FBI dataset, and practitioners can expect narrative or letter-grade interpretations to be off by at least one ordinal class in 12–80% of cases.

Results from our case study dataset suggest that differences in the nature of the coarse taxonomy bias depend on whether one is interested in the FBI value itself or in its corresponding narrative or letter-grade water-quality classification. It is unsurprising that the degree and type of bias depends on the interpretive scheme, given that the narrative classes and letter grades partition the range of FBI values differently, and because, in our study, the different water-quality classes were represented by different numbers of sampled locations.

This paper highlights the use of coarse taxonomy and tolerance-value averaging as a source of potential bias in bioassessments. Specific to Ontario, we advise conservation authorities involved in watershed reporting to avoid this practice. Our review of Conservation Ontario's watershed reporting guidance document (Conservation Ontario

Table 4 Effect of regression-based correction on classification success—narrative interpretations (i.e., the change in the percent of cases for which FBI and $FBI_{27corrected}$ narrative interpretations match, relative to the percent of cases for which FBI and FBI_{27} interpretations match; *NC*, no change)

		Change in classification success: $FBI_{27corrected}$ vs FBI_{27}						
		Excellent	Very good	Good	Fair	Fairly poor	Poor	Very poor
FBI	Excellent	+ 12%	− 10%	+ 3%	− 4%	NC	NC	NC
	Very good	+ 11%	NC	− 2%	− 9%	NC	NC	NC
	Good	NC	+ 2%	+ 12%	− 14%	NC	NC	NC
	Fair	NC	NC	+ 2%	− 3%	+ 1%	NC	NC
	Fairly poor	NC	NC	NC	NC	− 2%	+ 2%	NC
	Poor	NC	NC	NC	NC	− 5%	+ 3%	+ 3%
	Very poor	NC	NC	NC	NC	− 2%	− 1%	+ 3%

Table 5 Effect of regression-based correction on classification success—letter grades (i.e., the change in the percent of cases for which FBI and $FBI_{27corrected}$ letter grades match, relative to the percent of cases for which FBI and FBI_{27} grades match; *NC*, no change)

		Change in classification success: $FBI_{27corrected}$ vs FBI_{27}				
		A	B	C	D	F
FBI	A	+7%	NC	−7%	NC	NC
	B	+2%	+12%	−14%	NC	NC
	C	NC	+2%	−3%	−1%	NC
	D	NC	NC	NC	−2%	+2%
	F	NC	NC	NC	−4%	+4%

2011) suggests a need to investigate other potential sources of bias associated with survey design and indicator selection. A key challenge of regional monitoring programs is to make inferences about the population of streams from site-scale data. As noted by several authors (e.g., Herlihy et al. 2000; Stevens 1994; Thompson 2002), accuracy may be enhanced by randomizing site selection and proportionally representing stream types. Finally, given that the FBI was developed in Wisconsin, with tolerance scores assigned based on taxa-specific responses to organic enrichment, questions remain about its suitability as the singular indicator for reporting on water quality in streams draining southern Ontario's mixed-use watersheds. Modeling studies that investigate the FBI's response to multiple stressors of interest in the Ontario region are warranted. Such studies could facilitate the development of locally relevant interpretive criteria and might suggest the need to consider other indicator metrics.

Acknowledgements We thank the members of the Stream Monitoring and Research Teams (southern Ontario and eastern Ontario groups) for providing useful feedback at various points in the study process. We acknowledge the efforts of our collaborators, Erin Carrol, Holly Evans, Adrienne Lewis, Martha Nicol, Adrienne Ockenden, Ian Ockenden, Kim Ootjers, John Schwindt, Angela Wallace, and Rob Wilson, who shared their data with us. Thanks also to Jordan Ahee and Angela Wallace for their assistance with data compilation, and Nancy Aspden for Geographical Information Systems support.

Funding information This research was conducted jointly via operational program funding provided by Kawartha Region Conservation Authority, Ontario Ministry of Environment and Climate Change, and Otonabee Region Conservation Authority.

References

Bailey, R. C., Norris, R. H., & Reynoldson, T. B. (2001). Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. *Journal of the North American Benthological Society, 20*(2), 208–286.

Bouchard, R. W., Huggins, D., & Kriz, J. (2005). *A review of the issues related to taxonomic resolution in biological monitoring of aquatic ecosystems with an emphasis on macroinvertebrates*. Lawrence: Central Plains Center for BioAssessment Prepared in fulfillment of USEPA Grant X7-99790401.

Carter, J. L., & Resh, V. H. (2001). After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society, 20*(4), 658–682.

Chang, F. H., Lawrence, J. E., Rios-Touma, B., & Resh, V. H. (2014). Tolerance values of benthic macroinvertebrates for stream biomonitoring: assessment of assumptions underlying scoring systems worldwide. *Environmental Monitoring and Assessment, 186*(4), 2135–2149.

Conservation Ontario. (2011). Guide to developing Conservation Authority Watershed Report Cards. Conservation Ontario, Newmarket, Ontario.

EcoSpark. (2013). Water quality monitoring with benthic macroinvertebrates—field manual. Toronto, ON.

Herlihy, A. T., Larsen, D. P., Paulsen, S. G., Urquhart, N. S., & Rosenbaum, B. J. (2000). Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP mid-Atlantic pilot study. *Environmental Monitoring and Assessment, 63*(1), 95–113.

Hilsenhoff, W. L. (1977). *Use of arthropods to evaluate water quality of streams. Technical Bulletin No. 100*. Madison: Department of Natural Resources.

Hilsenhoff, W. L. (1987). An improved biotic index of organic stream pollution. *Great Lakes Entomologist, 20*(1), 31–40.

Hilsenhoff, W. L. (1988). Rapid field assessment of organic pollution with a family-level biotic index. *Journal of the North American Benthological Society, 7*(1), 65–68.

Jones, F. C. (2008). Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Review, 16*(NA), 45–69.

Jones, C., & Wilcox, I. (2003). *Conservation Ontario Discussion Paper: Recommendations for monitoring Ontario's water quality*. Newmarket: Conservation Ontario Resource document. http://conservationontario.ca/projects/pdf/CO_Water_Quality.pdf. Accessed 24 April 2017.

Jones, C., Somers, K. M., Craig, B., & Reynoldson, T. B. (2007). *Ontario benthos biomonitoring network: Protocol manual*. Toronto: Ministry of the Environment.

King, R. S., & Richardson, C. J. (2002). Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. *Journal of the North American Benthological Society, 21*(1), 150–171.

Norris, R. H., & Georges, A. (1993). *Analysis and interpretation of benthic macroinvertebrate surveys*, *Freshwater biomonitoring and benthic macroinvertebrates*. New York: Chapman and Hall.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing URL: https://www.R-project.org/.

Reynoldson, T. B., & Metcalfe-Smith, J. L. (1992). An overview of the assessment of aquatic ecosystem health using benthic invertebrates. *Journal of Aquatic Ecosystem Stress and Recovery (Formerly Journal of Aquatic Ecosystem Health), 1*(4), 295–308.

Robinson, A. (2016). Package 'equivalence'. Technical Report. https://cran.rproject.org/web/packages/equivalence/equivalence.pdf.

Robinson, A. P., Duursma, R. A., & Marshall, J. D. (2005). A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiology, 25*(7), 903–913.

Smith, A., Heitzman, D., & Duffy, B. (2009). Standard operating procedure: biological monitoring of surface waters in New York State. New York State Department of Environmental Conservation, Division of Water.

Stanfield, L. W. (Ed.). (2003). *Guidelines for designing and interpreting stream surveys: A compendium manual to the Ontario Stream Assessment Protocol*. Picton: Ontario Ministry of Natural Resources.

Stevens, D. L. (1994). Implementation of a national monitoring program. *Journal of Environmental Management, 42*(1), 1–29.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.